Prediction of Parallel In-Register Amyloidogenic Beta-Structures In Highly Beta-Rich Protein Sequences By Pairwise Propensity Analysis

By

Allen Wayne Bryan, Jr.

B.S. Physics
Mississippi State University, 2000

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2009

Signature of Author: ______________________________
Division of Health Sciences and Technology
July 13, 2009

Certified by: ______________________________ and ______________________________
Bonnie A. Berger, PhD
Professor of Applied Mathematics
Thesis Co-Supervisor

Susan L. Lindquist, PhD
Professor of Biology
Thesis Co-Supervisor

Accepted by: ______________________________
Ram Sasisekharan, PhD
Director, Harvard-MIT Division of Health Sciences and Technology
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering.

Prediction of Parallel In-Register Amyloidogenic Beta-Structures In Highly Beta-Rich Protein Sequences By Pairwise Propensity Analysis

By

Allen Wayne Bryan, Jr.

Submitted to the Harvard-MIT Division of Health Sciences and Technology on July 13, 2009 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Medical Engineering and Medical Physics

Abstract

Amyloids and prion proteins are clinically and biologically important beta-structures, whose supersecondary structures are difficult to determine by standard experimental or computational means. In addition, significant conformational heterogeneity is known or suspected to exist in many amyloid fibrils. Recent work has indicated the utility of templates and pairwise probabilistic statistics in beta-structure prediction. A new suite of programs, BETASCAN, STITCHER, and HELIXCAP, are presented to address the problem of amyloid structure prediction. BETASCAN calculates likelihood scores for potential beta-strands and strand-pairs based on correlations observed in parallel beta-sheets. The program then determines the strands and pairs with the greatest local likelihood for all of the sequence's potential beta-structures. BETASCAN suggests multiple alternate folding patterns and assigns relative *ab initio* probabilities based solely on amino acid sequence, probability tables, and pre-chosen parameters. STITCHER processes the output of BETASCAN and uses dynamic programming to 'stitch' structures from flexible abstract templates defined by constraints for amyloid-like all-beta structures. The 'stitched' structures are evaluated by a free-energy-based scoring algorithm incorporating BETASCAN scores, bonuses for favorable side-chain stacking, and penalties for linker entropy. The analyses of STITCHER structures emphasize the importance of side-chain stacking ladders in amyloid formation. HELIXCAP detects a class of end-caps, called beta-helix caps, which stabilize known beta-helix structures. These structures are known to stabilize globular beta-helix proteins and prevent their amyloidogenesis; their presence in a sequence is a powerful negative predictor of amyloid potential. Together, these algorithms permit detection and structural analysis of protein amyloidogenicity from sequence data, enhancing the experimental investigation of amyloids and prion proteins.

Thesis Co-Supervisor: Bonnie Berger, PhD

Title: Professor of Applied Mathematics; HST Affiliated Faculty

Thesis Co-Supervisor: Susan L. Lindquist, PhD

Title: Professor of Biology; Member, Whitehead Institute for Biomedical Research; Howard Hughes Medical Institute Investigator

Biographical Note: Allen W. Bryan, Jr.

Educational History:

From	To	School	Degree	Department *(concentration(s))*
2000	*2011*	Harvard U. /MIT	M.D.	Health Sciences and Technology *(medicine)*
2000	2009	Harvard U. /MIT	Ph.D.	Health Sciences and Technology *(bioinformatics)*
1996	2000	Miss. State Univ.	B.S.	Physics *(pre-medicine, math), magna cum laude*

Research Experience:

June 2003 – present, Drs. Bonnie Berger (MIT Math.) and Susan Lindquist (Whitehead Inst./ MIT Biology)

Computational biology: Prion and amyloid structure prediction and analysis

Programs developed: BETASCAN (pairwise probabilistic prediction of protein β-structure); HELIXCAP (HMM-based predictor of caps for β-helix protein structure); STITCHER (dynamic discrete flexible-template prediction of amyloid β-structure)

Summers 1998 – 2000, Dr. Peter Kim (Whitehead Inst. / MIT Biology)

Computational biology: Sequence analysis of viral-entry and similar proteins

Experimental biology: Protein purification and crystallization

Achievements: Prepared Five-Helix, a highly agglutinative protein construct based on HIV gp41 viral-entry surface protein, for crystallization and X-ray structure determination.

Programs developed: SKIPCOIL (coiled-coil protein structure detector).

January 1996 - June 1997, Dr. John T. Foley, (PI, The Optics Project / Miss. State Univ. Physics)

Physics simulation: Educational software design

Achievements: Designed format and properties of the Geometrical Optics module of the TOP program. Also wrote a tutorial designed for use with the Geometrical Optics module.

Programs: TOP (Geometrical Optics Module), web-accessible: http://www.webtop.org

June 1995 - July 1995, Dr. Johnathan Harris (MIT Chemical Engineering)

Chemical engineering simulation: Molecular dynamics of supercritical fluid

Achievements: Received scholarship to 1995 Research Science Institute at MIT. Performed FORTRAN-based analysis of chemical and physical properties under varying conditions. *(see presentation below)*

August 1993 – May 1994, Dr. John Boyle (Miss. State Univ. Biochemistry)

Applied experimental biology: Development of DNA-based veterinary diagnostic test

Achievements: Performed PCR-based DNA fingerprinting tests on field samples. *(see presentation below)*

Academic Honors:

2004 Member, *Informatics for Integrating Biology and the Bedside* (a National Institutes of Health National Center for Biomedical Computing)

2000 *Medical Scientist Training Program* grant, National Institutes of Health

1998 *Society of Scholars*, College of Arts and Sciences, Mississippi State University

1997 *Golden Key* Honorary Fraternity; *Gamma Beta Phi* Society; USAA *All-American Scholar*

1997 *Phi Kappa Phi* Honorary Society; *Alpha Epsilon Delta* Society

1996 *John M. Stalnaker Scholar for Science and Mathematics*, National Merit Scholarship Corporation

1996 *Stan Rundel Scholar in Physics*, Mississippi State University

1996 Scholarship to *Princeton Lectures on Biophysics* (sponsor: NEC Institute, theme: protein folding)

Teaching Experience:

2005-2007 Undergraduate and high school student mentor/supervisor, Bonnie Berger lab, MIT

2004 Teaching assistant, MIT Department of Biology 7.20 (Human Physiology)

1998 Volunteer for primary school reading and science, Columbus Municipal School District

1997-2000 Tutor in physics, Mississippi School for Mathematics and Science

1996-1997 Tutor in physics, Mississippi School for Mathematics and Science (minority students)

Clinical Experience:

2002 (July -August) 3rd year rotation, Internal Medicine, Beth Israel Deaconess Hospital, Boston, MA

2002 (February - May) Introduction to Clinical Medicine, Brigham & Women's Hospital, Boston, MA

1998 - 1999 Physical Therapy Volunteer, Baptist Mem. Hosp.-Golden Triangle, Columbus, MS

Publications:

Bryan AW Jr, Menke M, Weiland S, Lindquist SL, Berger BA. BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comp. Biol. 2009 Mar;5(3):e1000333. Epub 2009 Mar 27.

Conferences and presentations:

2005 Research in Computational Molecular Biology, Boston, MA

2000-2003 Harvard School of Medicine MD/PhD Symposium, Waterville Valley, NH

Presentation, 2001: Bryan AW Jr, Kim PS., Singh MP. SKIPCOIL: a method for detecting register shifts in coiled coil proteins.

2000 Mississippi Academy of Sciences, Tupelo, MS

Presentation, 2000: Bryan AW Jr, Kim PS. "Identification and characterization of register shifts in coiled coil proteins."

1995 Research Science Institute Symposium, Cambridge, MA

Presentation, 1995: Bryan AW Jr, Harris JP. Computer modeling of water: simulation and analysis.

1994 Mississippi State University Research Opportunities Symposium

Presentation, 1994: Bryan AW Jr, Boyle J. The use of polymerase chain reaction in the detection of channel catfish virus.

Acknowledgements

A full accounting of the people who made my journey to this point possible would become a thesis document in and of itself. A cast of hundreds smoothed my journey from northeast Mississippi to the Infinite Corridors, and a full and fair accounting thereof is probably impossible and certainly impractical. The man who is more responsible for my arrival at MIT than any other is Dr. Peter Kim, who took a chance on a boy willing to ask penetrating questions, and mentored my research and career during critical years of my academic development. His inventiveness, cheerful approach to life and work, and instant willingness to "talk science" are an enduring inspiration. Special mention must also be given to outstanding teachers at all levels, particularly Ms. Rosemary Lamar, Ms. Alma Turner, Dr. Mary Davidson, Mrs. Helen Perry, Mrs. Debbie Fancher, Dr. John Foley, and Dr. Don Downer.

The rapid prototyping of BETASCAN and STITCHER would not have been possible without the wizard coding skills of Matt Menke, who never failed to come through with code in the clutch. Many other members of the Berger and Lindquist labs were unhesitatingly helpful and invaluable. Charles O'Donnell and Jerome Waldispuehl were excellent coworkers who skillfully finessed a particularly close research topic from a potential conflict into a beneficial collaboration. Andrew McDonnell, Nathan Palmer, Rohit Singh, and Shannon Weiland helped convert the probability tables from BETAWRAP for use by BETASCAN, and were wonderful for everything from sounding boards to comradely stress relief. Michael Baym deserves special mention for being "Moriarty" to my "Holmes" during the search for thesis topics. Kenny Lu made the BETASCAN website functional at a critical juncture. Irene Kaplow was a terrific UROP even before her admission to MIT, and was a ray of sunshine I was pleased to help advise. On the "wet lab" side, Peter Tessier, Ram Krishnan, Sven Heinrich, Joshua Kritzer, Randall Halfmann, and Oliver King were generous with their time and data, and patient with the visitor from the strange world of computation.

I have been blessed with not one, but two outstanding supervising professors. Dr. Susan Lindquist was open and generous with aid for a project outside the direct scope of her investigations, patient with my stumbles in biological terminology, and always skillful in negotiating the sometimes rocky shoals of interdisciplinary collaboration. Dr. Bonnie Berger has been a shepherdess and "den mother" to me and to all her students, mentoring on all topics from grant development and manuscript editing to hair care. While her research was the ultimate basis for my own, this was only a small part of her role in this work and my life. Her care for not only my research but my career and personal development has been a joy during the stressful graduate student experience. I am proud to have had six years to call these extraordinary women "boss". Dr. Lenore Cowen was incredibly supportive and patient in the long process of shepherding the BETASCAN paper to publication, and taught me much about the art of article preparation. My committee chair, Dr. Pete Szolovits, went the extra mile – even so far as to telecommute to the defense – not only willingly, but cheerfully.

Finally, my heartfelt thanks and deepest love go out to my family, Wayne, Betty, and Dan Bryan. After dozens of visits and hundreds of calls ranging from panic abatement, meal reminder, and sanity checks to long political and philosophical discussions, I am more profoundly aware than ever before of the value and importance of close family. Though thousands of miles may separate us, I could not be who I am – and certainly could not still be sane – without your deep support and deeper love.

Table of Contents

Symbols used

c: frequency of a residue, derived from an amino acid frequency table C.

C: an amino acid frequency table.

$C_{allproteins}$: the amino acid frequency table for all proteins, calculated from SWISS-PROT 50.4.

$C_{allyeasts}$: the amino acid frequency table for GC-sparse yeasts, calculated from the genome of Saccharomyces cerevisiae.

ΔE_c: free energy change due to contact enthalpy.

ΔE_{el}: free energy change due to electrostatic enthalpy.

f, g: the two strand-pairs between a STITCHER link.

$f(s, p, l, o)$: propensity of beta-strand formation; a subsidiary calculation in BETASCAN.

ΔG: free energy change due to folding.

$g(s, p, l)$: propensity of strand sequence occurrence; a subsidiary calculation in BETASCAN.

$h(s, p, l, o)$: odds ratio of beta-strand formation, the singleton score output of BETASCAN.

i: first residue of the upper strand of a strand-pair considered for a STITCHER structure, derived from p in BETASCAN.

j: first residue of the lower strand of a strand-pair considered for a STITCHER structure, derived from q in BETASCAN.

$j(s, p, q, l, o)$: pairwise propensity of strand-pair formation; a subsidiary calculation in BETASCAN.

k: upper bound of strand and strand-pair length calculated by BETASCAN.

$k(s, p, q, l, o)$: odds ratio of strand-pair formation, the pairwise score output of BETASCAN.

l: length of strand or strand-pair in BETASCAN, counted in residues. Becomes L in STITCHER.

l_{aa}: the distance between alpha carbons in a peptide chain.

L: length of strand-pairs in STITCHER, derived from l in BETASCAN, counted in residues.

L_{link}: number of residues in linker.

m: number of rungs per polypeptide chain in STITCHER.

n: number of strands per rung in STITCHER.

n_{XX}: number of occurrences of residue-pairs $\{X, X\}$ in a structure.

o: orientation of a strand-pair in BETASCAN.

p: position of first residue of strand or upper strand-pair in BETASCAN. Becomes i in STITCHER.

P: probability, in the general sense.

q: position of first residue of lower strand-pair in BETASCAN. Becomes j in STITCHER.

r: a rung-pair in STITCHER, $1 \leq r \leq m$.

$r_1, r_2, \ldots r_n$: the strand-pairs in rung-pair r.

r_{link}^{0-1}: the linker from rung-pair r to the previous rung-pair, if any. Not defined for $r = 1$.

$r_{link}^{1-2}, \ldots r_{link}^{(n-1)-n}$: the linkers in rung-pair r.

R: the gas constant.

R_1, R_2: residues in upper and lower strands, respectively.

s: polypeptide sequence(s) of a strand or strand-pair in BETSACAN.

ΔS: change in entropy.

t: summation variable, measuring length along strands or strand-pairs.

T: temperature, set to 300K for all calculations in this work.

v: occurrence of a residue, derived from residue occurrence table V.

v_{ends}: the volume that may be occupied by the linked ends of a circular peptide chain.

$V(X_1, \theta)$: the 1x2x20 table of residue occurrences in amphipathic parallel beta-strands.

w: occurrence of a residue-pair, derived from residue-pair occurrence table W.

$W(X_1, X_2, \theta)$: the 2x2x20 table of residue-pair occurrences in amphipathic parallel beta-sheets.

X: a residue, selected from the set of twenty possible amino acid sidechains.

X_1: upper residue in residue occurrence table V and residue-pair occurrence table W.

X_2: lower residue in residue-pair occurrence table W.

Z: the number of structures calculated by STITCHER; a prescribed value.

β_1, β_2: the strands of a strand-pair.

θ: orientation in residue occurrence table V and residue-pair occurrence table W.

λ: parameter of the distribution of amphipathic parallel strand-pair lengths.

Introduction

Protein structure and folding

For sixty years, understanding the formation of three-dimensional protein structures has been one of the greatest challenges in the life sciences. The identification of secondary structures, such as the β-pleated sheet [1], was only the first step in classifying and analyzing the incredible variety of folds found in nature. Two experimental methods, X-ray diffraction crystallography and nuclear magnetic resonance determination, have been used over decades to collect thousands of structures [2]. Both convergent evolution and common genetic inheritance have been found to contribute to the existence of fold families and superfamilies, which have now been extensively catalogued [3]. However, the difficulty, uncertainty, and expense of these methods has so far precluded a large-scale experimental canvass of protein structure [4] analogous to the Genome Project [5]. Investigations of the necessary conditions of protein folding led to the postulation of Anfinsen's dogma of a unique and strongly attractive thermodynamically stable native state [6, 7]. However, this notion was tempered and challenged by Levinthal's recognition of the mismatch between the size of polypeptide conformational space and the time required for protein folding [8].

The heritage of protein structure databases

The proteins produced by living organisms have been long recognized to adopt many and varied three-dimensional conformations, usually referred to as protein folds. Depending on the function of the protein, folds can form nearly any construction required for cell function, from pores to structural struts and from small free-floating enzymes to gigantic complexes designed for metabolism, replication, or signal transduction. Understanding the origin and production of this incredible diversity from a limited set of building blocks – the common peptide backbone and the standard twenty amino acid side-chains – has been an enduring and profound challenge to the field of molecular biology.

The first major breakthrough in the understanding of protein structure came with the seminal work of Linus Pauling and colleagues in 1951 [9]. These studies built on Pauling's theoretical work on resonance theory, which predicted the planar nature of the peptide bond, and X-ray diffraction studies of individual amino acids and short peptides, which provided critical data on bond lengths and angles. Combining these aspects, a short series of papers laid out the fundamental components of secondary structure. First to be predicted were the α-helix and its rare cousin the γ-helix [10], with their localized hydrogen-bonding, non-integer residue/turn ratios, and repeating spiral conformations. Then, Pauling and Corey turned their attention to the β-pleated sheet [1]. In the β conformation, the backbone of the polypeptide chain forms multiple extended regions forming near-linear, laterally oriented "strands" linked by hydrogen-bonds alternating in orientation with every other residue. Two possible configurations, presented in their first figure (and reproduced here as Figure 1-1), were presented: one with each peptide strand's bonds oriented opposite those of neighboring strands (Figure 1-1, left), and the other with every peptide bond oriented identically (Figure 1-1, right). These conformations, now referred to as the "antiparallel" and "parallel" β-sheets respectively, were borne out as fundamentally accurate representations of protein structure with the first direct observations of β-structures in the X-ray diffraction crystallography of whole proteins in 1965 [11].

With the advent of X-ray crystallographic analysis of entire proteins, which had begun by the late 1950's [12, 13], the importance of structural complexities beyond the α and β conformations became evident. In the terms of the classification system devised by Kaj Ulrik Linderstrøm-Lang in 1952

[14] and illustrated by Figure 1-2, the interest shifted to the *tertiary* structure of proteins, i.e., the arrangement of various helices, sheets, and connecting loops (*secondary* structure) within a single polypeptide chain (whose sequence makes up the *primary* structure of the protein). With the increasing worldwide experience and skill in crystallizing proteins in native conformations and collecting sufficient data for analysis, the number of solved crystal protein structures began an exponentiation which has continued to this day [15]. The foundation of the Protein Data Bank (PDB) [2, 16] in 1971 began the process of assembling the coordinates of structures derived from X-ray crystallography, and later, from nuclear magnetic resonance data utilizing the nuclear Overhauser effect[17] to devise constraints on structural geometry [18].

By the 1980's, the recurrence of motifs and whole folds in the PDB led to the description of families, and later superfamilies, of folds. These could be classified by secondary structure content (all-α, all-β, or α-β), by shape (e.g., the Greek-key motif), and sometimes by function (as in the G-protein signaling family). The 1981 introduction of the now-standard 'cartoon' visualization method pioneered by Jane Richardson [19] made visualization and understanding of these folds practical. As the number of folds expanded, databases of related folds such as SCOP [3] and CATH [20] were created to classify and organize knowledge of protein architectures. Today (mid-2009) there are over 59,000 total protein structures in the Protein Data Bank, of which over 34,000 are non-redundant. These structures have been classified by their fold architecture [21] into around 4,000 families of related conformations.

Anfinsen's dogma and Levinthal's paradox

The discovery of *in vitro* folding of proteins [6] demonstrated that the conformation of proteins could be intrinsically determined by the amino acid sequence of the polypeptide chain, and could be adopted outside the cell under the correct environmental conditions [7]. Several key conclusions were drawn from these observations that influenced future protein folding experiments. First, the 'native' or fully folded state was believed to lie at a strong and unique free-energy minimum, making the native conformation a strong thermodynamic attractor state. This minimum was held to be surrounded by steep energy barriers, which would prevent instability of the native state once formed. Despite these energy barriers, there must be kinetically accessible path(s) from the large number of unfolded states to the unique native state. There was therefore strong interest in trying to trace the kinetic folding pathway from the unfolded to the native state. Finally, the context of other molecules *in vivo* (particularly chaperone proteins) and other "*trans*-acting factors" was de-emphasized in favor of study of polypeptides in isolation.

In 1968, Cyrus Levinthal pointed out [8] that the number of potential conformations available to any polypeptide chain is astronomical. Indeed, according to Levinthal's calculations, if every atom in the universe were used to assemble a conformational variation of a single 100-residue protein every second since the formation of the universe, only a tiny fraction of the available conformational space of the polypeptide chain would have been explored. However, biologically active polypeptide chains typically proceed from these vast deserts of random coil conformations to folded states in less than a second. Any assumption that this rapid assertion of structural order occurs through random searches of the available conformational space is therefore precluded by the so-called "Levinthal's paradox", namely, that the needle of folding is impossibly small compared to the haystack of conformational space. As Levinthal noted, it must therefore be possible to reach the native state quickly from nearly any conformation. Despite this remarkable intuition, the random-search or 'golf course' [22] model of protein folding persisted for many years, spurring searches for unique pathways of folding for globular proteins.

Amyloid

In parallel with the investigation of the folding of globular proteins, investigations began on a distinct and highly organized type of protein aggregation that belied several of the assumptions made by Anfinsen. *Amyloids* were originally named when a particular type of extracellular inclusions previously described as 'waxy' were suspected by Virchow to be composed of starch because of their agglutinative properties [23]. The term 'amyloid' had already become standard by the 1870's, when component molecules were discovered to be proteinaceous [24]. Detection of amyloid was originally accomplished by observing the binding of various organic dyes [25]. However, Congo red dye [26] became dominant due to its superior definition, reliability, and its peculiar property of apple-green birefringence upon amyloid binding [24]. The last major dye to be introduced for amyloid detection was thioflavin T, favored for its yellow-green fluorescence upon binding [27]. Despite the historical association with extracellular assembly, intracellular inclusions with amyloid properties have also been observed [28].

The modern understanding of amyloids began with the first electron microscope images of amyloid [29], which revealed non-branching fibrils composed of aggregated proteins as a common form of diverse types of amyloid. The nature of these fibrils was clarified by X-ray diffraction studies of amyloid [30]. A characteristic sharp, intense signal at 4.7 Å was observed, indicating a structure termed "cross-β". Cross-β is a not a specific fold, but a general descriptor indicating that the β-strands are oriented perpendicular to the long axis of the amyloid fibrils. The β-strands were originally believed to be anti-parallel in all cases due to a weak reflection at 9.3 Å, but subsequent studies [31] showed this signal to be an artifact of sample processing for at least some fibrils. A variety of other experimental methods [32], including circular dichroism spectroscopy [33] and electron microscopy [34], have since confirmed the fibrillar model of amyloid. Thus, amyloid is currently understood as a highly stable family of structures composed of many protein monomers arranged into β-sheet–rich fibrils with cross-β orientation [35] (see Figure 1-3).

Prions

Prions were first postulated by Griffith [36] and Prusiner [37] as the causative agent of scrapie, an infectious neurodegenerative disease of sheep. The then-incredible suggestion of a proteinaceous infectious agent lacking any genetic material was originally a hypothesis of exclusion, after tests revealed the incredibly small size and radiation resistance of the scrapie agent [38]. Fifteen years later, the hypothesis was confirmed when prion protein (PrP) was linked to the presence of scrapie[39]. The PrP protein is now generally regarded as the cause of the entire class of spongiform encephalopathies, including Cruetzfeld-Jakob disease and kuru in humans and 'mad cow disease' in bovines [40].

Since then, the definition of prion has been repeatedly and controversially extended as new properties and functional roles have come to light [41, 42]. The term 'prion' now describes not only self-templating aggregative proteins in mammals and avians, but also fibrillar proteins in yeast that contribute to epigenetic inheritance [43-45], switching elements in sea slug axons [46], and control elements for mating compatibility in fungi [47]. Prion fibrils and infectivity has been transmitted across cell divisions, across cell membranes, and even between individuals of different species [48]. Currently, a protein generally agreed to be a prion would have, in addition to its native and usually soluble conformation, at least one additional conformation (which is typically β-sheet rich and insoluble) that is self-perpetuating and infectious [35].

The structural natures of amyloid and prion proteins have been highly controversial topics [49-51]. The debate has been complicated by the morphological heterogeneity of amyloid structures suggested by EM imagery [52, 53] and the demonstration of prion 'strains' or 'variants' with differing growth and stability phenotypes [54-56]. For the Sup35 prion, strains are known to maintain specificity through serial passage [55] and have been correlated with differences in conformation [57].

Functional associations of prion and amyloid

Amyloids were originally studied due to their association with a host of diseases [24, 58]. The systemic amyloidoses include familial or inherited amyloidosis, primary or AL amyloidosis, and secondary or AA amyloidosis. The clinical distinction between these syndromes lies in their etiology, respectively held to be genetic mutation [24], overproduction of plasma cells manufacturing light-chain immunoglobulins [59], and overconcentration of a serum inflammatory protein [60]. Another major class of amyloid-associated diseases is the neurodegenerative amyloidoses [61], including Alzheimer's [62], Parkinson's [50], and Huntington's [63]. A variety of other amyloidoses have been observed to be associated with neuropathy, angiopathy, nephropathy, and type II diabetes [34, 64].

By analogy to the systemic amyloidoses, the non-systemic amyloid diseases were originally assumed to also be caused by the associated amyloids. However, more recent evidence suggests that the formation of amyloids may more commonly be a protective mechanism. Especially in the case of the neurodegenerative amyloidoses, aggregation may act to sequester misfolded polypeptides that would otherwise dwell in more toxic, and more highly interactive, oligomeric species. Structurally speaking, the amyloid fibrils formed in all these diseases are strikingly similar considering the diversity of their amyloid precursors – twenty-five of which were identified in humans as of the last major classification [65].

More recent searches for amyloid function have revealed that amyloids serve important biological functions across many species. Amyloid fibers composed of the protein Pmel17 are part of the production system for melanin [66]. In bacteria, biofilm stability contributes to the tenacity of infections, a property induced in part by the amyloid curli [67]. The self-templating amyloids in fungi may faithfully pass their conformations on to daughter cells, providing new phenotypes [68] and in some cases beneficial [69] buffered genetic diversity [70, 71]. Self-perpetuating prion-like switches in the state of CPEB also appear to play a role in neuronal learning and memory, by maintaining a translation factor involved in the maintenance of synapses in an active and highly localized state [46]. Most recently, amyloid was discovered to be a primary means of stockpiling a wide variety of hormones in the Golgi-derived secretory granules of various mammalian endrocrine glands [72].

The mystery of detailed structure in amyloid and prion

Amyloid fibrils, 10-30 nm in diameter, resolve into protofilaments measuring typically 2-6 nm in width under transmission electron and atomic force microscopy [32, 34, 50, 73]. However, below this resolution, microscopy must give way to alternate methods for detailed structural determination. The standard methods for such investigation have been frustratingly difficult. The cross-β structure, once formed, is highly stable and resistant to proteinase activity or denaturation by heat or chemical solvent [74]. Crystallization of the cross-β structure for X-ray structural determination has proven impractical except for extremely short segments [51, 75]. Likewise, solid-state NMR studies have been pursued due to the insolubility of amyloid folds [76]. Biochemical studies of amyloid and prion structure have therefore tended to utilize indirect observation of formation using dye-binding assays [24-26], fluorophore activation [27, 77], or functional assays [78]. Recently, direct observations through solid-state nuclear magnetic resonance (ssNMR) studies have permitted direct confirmation of parallel β-structure in a few specific cases [79-82]. Figure 1-3 depicts one example of these structures.

Amyloid and prion protein structures are of great interest to multiple disciplines for multiple reasons. Physicians seek a greater understanding of the origin, assembly, diagnostic and pathological implications of *in vivo* amyloid aggregation and prion infectivity. Biologists desire to understand the mechanisms of amyloid and prion assembly and the evolutionary and phenotypic implications of the presence and absence of amyloid states. Molecular biophysicists and computational biologists seek greater clarifications of the underlying influences of primary structure, environmental effects, and folding cofactors on the folding of polypeptides. Theorists of protein folding are challenged by the bi-stable nature of prion conformational change and the context dependency of amyloid folding, which violate the principles of Anfinsen's dogma, yet must still somehow avoid the pitfalls of Levinthal's paradox. Despite this great concentration of interest, the experimental tools to elucidate detailed structures of amyloid and prion proteins are insufficient. We therefore turn next to computational methods of β-structure prediction to continue searching for the conformational details of these vital proteins.

Figures

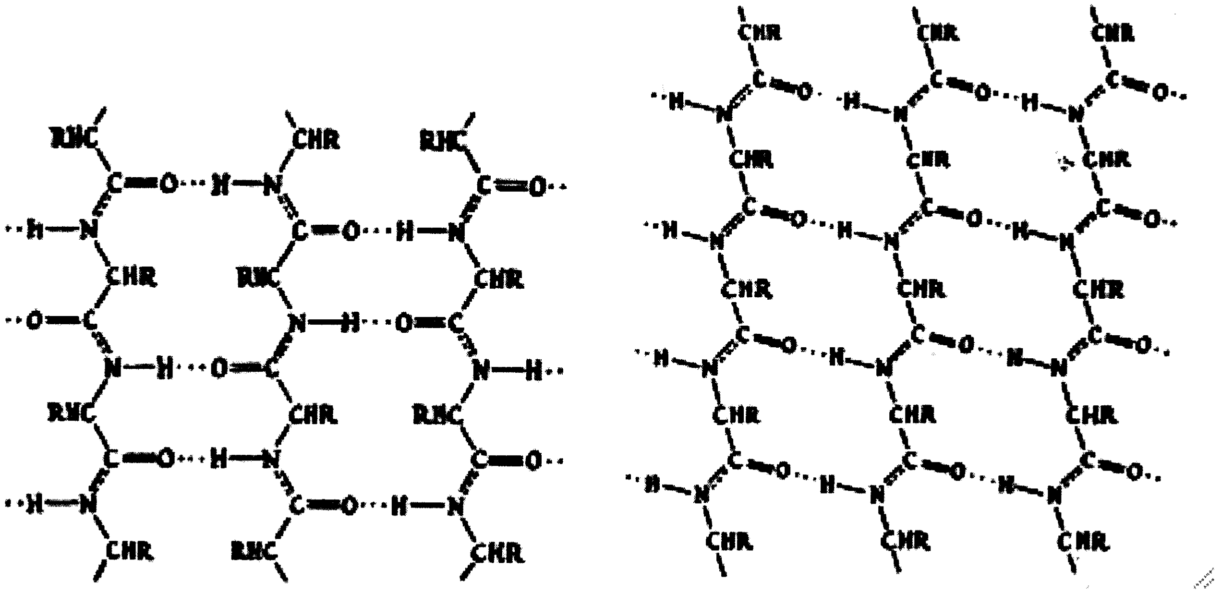


Figure 1-1: the original drawings of the antiparallel, left, and parallel, right, β-sheet conformations. Reproduced from [1].

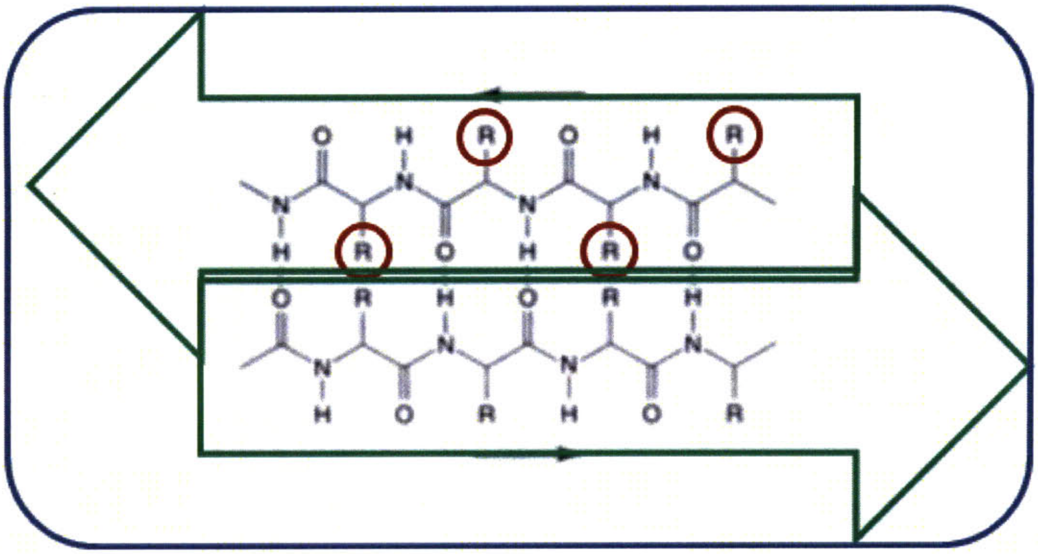


Figure 1-2: An illustration of the levels of protein structure, as originally expounded by Linderstrøm-Lang in 1952 [14]. Red circles indicate the locations of amino acid residues, the variable portion of the protein *primary structure.* Green arrows indicate the individual beta-strands, an element of *secondary structure.* The blue rectangle indicates the whole fold of the protein, described as *tertiary structure.*

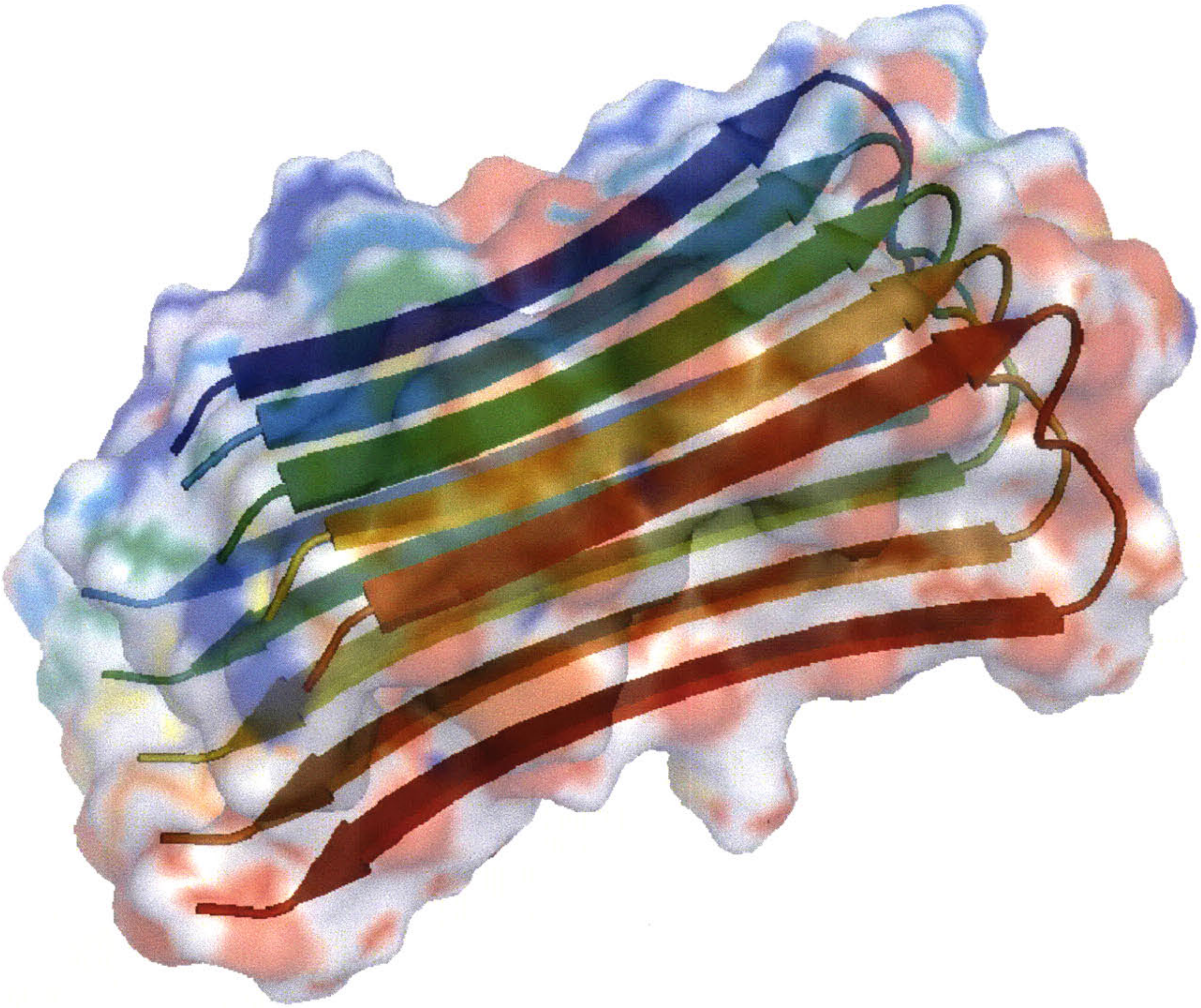


Figure 1-3: An example of amyloid structure, depicting the cross-β nature of the conformation and the stacking of multiple copies of a polypeptide sequence into a aggregated fibril. Structure after Lührs et al. [80], PDB structure 2BEG, rendered by PyMol [83].

Computational Strategies

Despite the fundamentally spurious nature of its most naïve formulation, the Levinthal paradox has nonetheless posed a challenge to computational efforts to solve the protein folding problem. In the absence of means to define the dynamic free-energy state or the kinetic folding pathway(s) for a particular protein, any conformational state appears *a priori* to be of equal probability for the purposes of computer algorithms. The problem of protein structure prediction may be attacked in two ways: relieving the Levinthal paradox by elimination of impossible or highly improbable conformations, and preventing the effect of the paradox by finding the free-energy 'landscapes' that encourage folding [64]. Despite extensive efforts [84], a generalized understanding of the mechanisms of protein folding have proven insufficient for reliable, high-fidelity *ab initio* prediction.

Ab initio prediction

Various studies have described the general protein folding problems as NP-hard for conditions implying the naïve equal-probability assumption. Preliminary attempts simplified by solving for straight-line subsequences and for graphs reflecting simple folds [85, 86]. Eventually, even the simplest realistic description of protein folding, modeling amino acids simply as hydrophobic or polar and the available conformation space as a cubic 3D lattice, was shown to be incomplete [87]. These studies provided confirming evidence for the Anfinsen postulate that folding pathway(s) are encoded into the amino acid sequence of proteins [88].

Interest has thus alternatively turned to utilizing free-energy calculations in the context of statistical mechanics [89]. Whole-protein *ab initio* theoretical methods were developed for globally solving protein structures. Monte Carlo minimization [90] introduced the concept of pushing past local minima to search for global optimization by perturbing protein structures in the process of optimization. Molecular dynamics simulation [91] attempted to simulate both the kinetics and dynamics of protein folding, in line with Levinthal's suggestion that kinetics would play a role in the selection of the pathway(s) to the native state. These efforts were aided by the recognition that side-chain optimization was as important to high-resolution structure determination as backbone packing [92, 93] and that the optimization of interactions between secondary structures by manipulating them as unitary objects is an important step in global alignment [94, 95]. Despite these improvements, constant testing and optimization, and the sporadic success, a generally applicable, *ab initio*, whole-protein method for predicting tertiary structure from primary sequence remains elusive [4, 84], with only incremental improvements in the latest global assessment of performance, the critical assessment of methods of prediction (CASP).

The primary difficulty with general *ab initio* protein structure determination is the importance of interactions between residues that are distant from each other as measured by backbone position, but highly proximal in the folded structure. One clue to how nature overcomes this difficulty was provided by the recognition that one β-strand can induce the formation of the next β-strand in a β-sheet [96]. Since β-strands, unlike α-helices, can be far distant in backbone position yet specific in their assembly, β-structure can provide a scaffold for the folding of some proteins. This realization is particularly vital to the understanding of amyloid and prion proteins, which are not only all-β in structure, but are known, i.e. in the mammalian PrP protein, to be formed of sequences that form α-structure in other contexts [97].

Secondary structure prediction efforts have met with somewhat more success than tertiary structure prediction. Starting with the initial propensity statistics of Chou and Fasman [98, 99], increasingly complex and varied methods have been developed to incorporate the accumulated knowledge of protein homology and similarity into modeling of secondary structure [100], including Bayesian inference [101], support vector machines [102] and neural networks [103]. However, because of the increased distance between β-strands as opposed to the highly local 3-4 residues/turn connections of α-helices, β-strand prediction remains significantly poorer even for globular proteins [104]. The poor sequence conservation and lack of homology between amyloid and prion proteins effectively prevent utilization of any of these methods, which instead predict a lack of structure for sequences that produce some of the most stable protein fibrils known.

Previous strategies

In the absence of a general *ab initio* solution to the protein folding problem, strategies with more restricted scope have been developed. One of the most successful is known as threading, after the THREADER algorithm that initially demonstrated the principle [105]. Threading can extrapolate from a known protein structure to another, if the two are related in both sequence and structure. To do so, the amino acid sequences of the two proteins are typically first aligned [106]. This alignment guides the placement of the second protein in a conformation directly analogous to the structure of the first. Localized methods are then employed to adjust this rough approximation into a high-quality structure. The previously known, or template, structure is required to sufficiently restrict the conformational space to make the prediction of high-quality structure tractable. In addition, the sequences must be sufficiently similar to permit an accurate alignment [107]. In the “twilight zone” of protein structure

[108], proteins have similar structures yet insufficient sequence similarity for accurate alignment, despite better-conserved structural similarity [109-111].

Another approach to protein structure prediction attempts to use thermodynamic principles as its guide [112]. Such methods seek to describe structures in terms of their free energy, accounting for the enthalpy of hydrogen bonds and ionic interactions and the entropy caused by restricting the freedom of the peptide backbone and sidechains, as well as the so-called "hydrophobic interactions". This method is used to great effect as part of the localized optimization methods in threading. However, attempts to apply free-energy calculations globally stumble on several limitations: the uncertainty of measurement of free-energy parameters, the simplifying assumptions of many approaches, and the lack of a computationally tractable method for describing the hydrophobic collapse of most globular proteins. Without these features, the free-energy method cannot provide the reduction of conformational space provided by the known structure of a template.

Yet a third method of protein structural prediction uses the thousands of determined protein structures to determine statistics relating sequence patterns to structural features [113]. According to this philosophy, the dependence of structure on sequence implies that similar interactions will occur repeatedly throughout the database of known protein structures. By cataloguing these patterns, predictions for proteins of unknown structure may be evaluated in terms of probability and likelihood. Some local patterns, such as coiled coils [114], can be extracted by this method without recourse to templates or previous knowledge of protein structures.

A new approach for prediction of amyloid-like β-structures

Recently, novel efforts to push the boundaries of protein structure predictions have begun to combine these three methods in different ways. The challenge of identifying the parallel β-helix fold (Figure 2-1), a fold postulated as similar to amyloid [115], from its sequence was addressed by BETAWRAP [116, 117]. BETAWRAP was the first program to incorporate the important long-range pairwise interactions into a computational method to predict β-structure. In doing so, BETAWRAP was able to predict *strand-pairs*, defined as any two β-strands connected by the hydrogen bonds of a β-sheet. The predictive power, and the greatest limitation, of the BETAWRAP approach originated from its *flexible* template of relative β-strand positions. One turn between two β-strands (designated T2 by the authors of BETAWRAP, indicated by red in Figure 2-1) was discovered to consistently be of a particular length, connecting two β-strands of also consistent length. This pattern was repeated many times within each β-helix protein, and also replicated across proteins of widely divergent amino acid sequences. This structural observation permitted the use of probabilistic prediction to both search for and align the β-strands of the protein, facilitating the near-complete assembly of tertiary structure from primary sequence – but only for a restricted sub-family of proteins.

The mechanism of an abstract structural template is more broadly applicable than the three-dimensional fitting procedure of threading. The exact configurations of backbones and amino acids are subsumed in this method by a succinct description of backbone contacts. Furthermore, abstract structural templates can provide accurate predictions across a wider variety of primary sequences than traditional threading. However, as currently implemented, this strategy depends on the manual identification of a structural motif across multiple proteins. While such a strategy has been successful in other limited applications [118, 119], the general approach can only be extended slowly – and cannot address at all the vast majority of proteins without clear family similarities of structure.

The divide and conquer plan for amyloid structure prediction

The approach outlined in subsequent chapters attempts to utilize the most successful aspects of many of the previously mentioned approaches in a piecewise fashion. This "divide and conquer" plan reflects the insight that secondary structure in general, and β-structure in particular, can be used as building blocks to assemble and optimize tertiary structures. Because this plan applies only to β-structure, it must be applied to sequences without significant α-helix content.

The first step is to identify individual parallel β-strands, which is accomplished with singleton probabilities derived from a pairwise probability database. Single strands are next combined into *strand-pairs*, units of two adjacent parallel β-strands, utilizing the same pairwise probability database. A maxima-finding strategy prunes the strand-pairs to reveal the most likely connections between distant residues (Chapter 3, BETASCAN). Next, these strand-pairs are 'stitched' together into sheets and complete structures by chaining strand-pairs together (Chapter 4, STITCHER). A key component of the stitching algorithm is the verification of the relative positional alignments of the β-sheets. The effects of side-chain stacking interactions, which create ladders of stabilizing residues along the sheets and turns, are the most important influence on the selection of properly aligned strand-pairs. Free-energy methods are used to choose the optimum combinations of stacking, linker loops, and pairwise interactions to identify a set of near-optimum structures. Insights from experimental data may be used to enhance and select the best structural predictions of amyloid structures. Lastly, to verify that these structures are amyloidogenic and not the related globular β-helix structure, a hidden Markov model identifies the β-helix caps peculiar to those proteins (Chapter 5, HELIXCAP). The β-helix cap prevents the exposure of the β-strands and hydrophobic core of a folded peptide both to solvent and to fibril ends that promote amyloid assembly. Therefore, in contrast to the positive detection methods of BETASCAN and STITCHER, HELIXCAP represents a negative detection method that identifies sequences not prone to amyloid or prion formation.

Figures

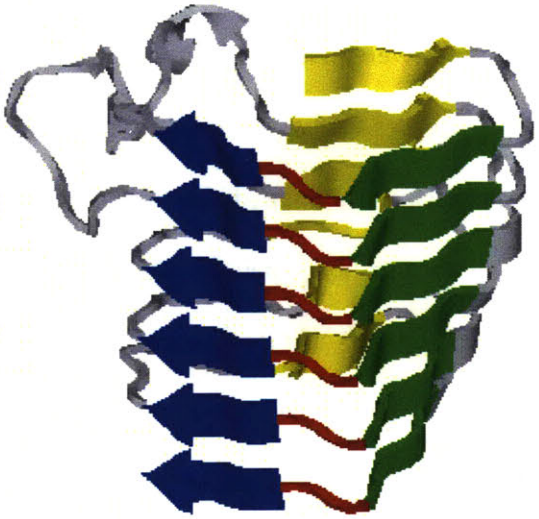


Figure 2-1: Pectate lyase C from *Erwinia chrysanthemi*, one of the first β-helices to be crystallized [120]. The T2 turn, the vital element of the BETAWRAP abstract template, is highlighted in red.

BETASCAN: Prediction of β-strands and strand-pairs[1]

Introduction

As described in chapter 1, "amyloid" is a term used to describe a particular type of protein structure that can be adopted by a very wide variety of proteins with completely unrelated primary amino acid sequences [58, 64]. It is a form of protein aggregation, but of a distinct and highly ordered type. It has recently been realized that, given the right conditions, a great many, perhaps most, proteins have the potential to form amyloids. This appears to be due to intrinsic properties of the peptide backbone, a finding of great importance for understanding the evolution of protein folds. A much smaller fraction of proteins, and protein fragments, assemble into amyloid under normal physiological conditions, and these are of great interest in diverse aspects of biology and medicine [121].

Many amyloids first came to our attention because they were associated with a wide variety of diseases, from systemic amyloidoses to neurodegenerative diseases such as Alzheimer's [97]. It had initially been assumed, therefore, that amyloids were toxic species. This indeed may be the case in peripheral amyloidoses, where the massive accumulation of amyloid fibers may physically disrupt normal tissue function [122]. Increasingly, however, evidence suggests that the formation of amyloids may more commonly be a protective mechanism which, especially in the case of the neurodegenerative amyloidoses, acts as to sequester misfolded polypeptides that would otherwise dwell in more toxic, and more highly interactive, oligomeric species. It has also recently been realized that amyloids serve important biological functions in a number of different situations. For example, in melanocytes, amyloid fibers formed by Pmel17 play a role in the production of melanin [66], and in bacteria extracellular

[1] The material in this chapter previously appeared in *PLoS Computational Biology:* Bryan, A.W., Jr., et al., *BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis.* PLoS Comput Biol, 2009. 5(3): p. e1000333.

amyloids are a key feature of the biofilms that are so difficult to eradicate in various infectious processes [67]. In fungi, a special class of self-templating amyloids serve as elements of inheritance: these bi-stable proteins can persist as soluble or amyloid species and the change in function that occurs when with the switch to the amyloid form, is passed from generation to generation as mother cells faithfully pass amyloid (prion) templates through the cytoplasm to their daughter cells [123, 124]. Such self-perpetuating prion-like switches in state also appear to play a role in neuronal learning and memory, by maintaining a translation factor involved in the maintenance of synapses in an active and highly localized state[46]. There is, therefore, great interest in deciphering the structures that underlie amyloid states.

Several methods have established that amyloids are generally rich in β-strands aligned perpendicular to the long axis of the fibril [32-34, 47, 125]. Beyond this, frustratingly little is known about their structure. Crystallization for X-ray structure determination has proven impossible except for extremely short segments [51, 75]. Notably, the importance of interactions between side-chains in these structures establishes that a detailed understanding of such interactions will be necessary to comprehend the physical and biological properties of other amyloids. The insolubility of amyloids has also precluded NMR-based structural determination until very recently, when solid-state nuclear magnetic resonance (ssNMR) studies have yielded partial, specifically parallel β-structures in a few specific cases [80-82, 126, 127]. Due to the scarcity of direct evidence, the nature of amyloid and prion supersecondary structures and their relation to sequence have been highly contentious topics [49, 51, 128]. The debate has been complicated by the morphological heterogeneity of amyloid structures suggested by EM imagery [52, 53] and the demonstration of prion 'strains' or 'variants' with differing growth and stability phenotypes [54-56]. In the case of the yeast prion protein Sup35, such variants have been demonstrated to maintain specificity through serial passage [55] and have been correlated with differences in conformation [57]. These results underscore the need to consider alternate supersecondary structures for amyloid and prion strands.

Given the difficulty of direct observation of supersecondary structure, computational modeling of amyloid folding has been attempted. Unfortunately, barriers exist to the effective application of sequence-based computational analysis. Several homologous prion-forming domains, while functionally conserved over evolutionary time, have sequence identities of under 25%, with sufficient additional rearrangement as to preclude multiple sequence alignment via standard algorithms such as CLUSTAL [106]. Analysis of amyloidogenic proteins has not revealed overall commonalities of sequence, except in individual residue frequencies [129-131] and a tendency for imperfect repeats to appear [132, 133]. Secondary structure prediction algorithms [100, 134] identify many amyloid- and prion-forming domains as random coil without structure. The amyloid-forming domains of these sequences are removed by the low-complexity filters of local sequence alignment tools such as BLAST [135], rendering another family of methods ineffective.

The strong evidence for β-structure in amyloid suggests that, as an alternate means of secondary structure prediction, computational methods designed to predict globular β-structure should be assessed. BETAWRAP [116, 117] was the first program to incorporate the important long-range pairwise interactions into a computational method to predict β-structure. In doing so, BETAWRAP was the first program to predict *strand-pairs*, defined as any two β-strands connected by the hydrogen bonds of a β-sheet. The program is restricted by a template of strand lengths to predict only one sub-family of the parallel β-helices, a fold widely cited as similar to amyloid [52, 115, 136]. BETAPRO [137] is a general method that incorporates pairwise properties into a neural net to learn globular β-strands and strand-pairs.

A variety of other approaches have been implemented in the search for a reliable detector of protein aggregation. TANGO [138] utilizes a statistical mechanics approach to make secondary structure predictions, including differentiation of β-aggregation from β-sheets. The TANGO algorithm presumes that all residues of an aggregate will be hydrophobically buried. Zyggregator [139] models aggregation propensity per residue as a combination of four factors intrinsic to a sequence: charge, hydrophobicity, secondary structure propensity, and the "pattern" of alternating hydrophobic and hydrophilic residues. Zyggregator derives its statistical basis from a study of effects of mutation on aggregation [140] and calculates its scores based on a sliding window of 21 residues. SALSA [141] uses a sliding window to sum the cumulative Chou-Fasman parameter score, then selects the 400 best scores and sums each residue's contribution. Finally, PASTA [142, 143] calculates singleton and pairwise propensities for individual residues and residue-pairs by calculating a weighted average of the contribution of that residue or pair to β-strand formation. The pairwise scores, in turn, were calculated according to a Boltzmann energy function derived from the adjacencies in a database of 500 annotated structures.

We introduce a program, BETASCAN, to predict prions and amyloids as well as other forms of parallel β-structure. Like PASTA, BETASCAN relies on calculation of strand propensities. However, BETASCAN makes use of a novel hill-climbing algorithm to find the most preferred β-strands and strand-pairs. Our hypothesis is that BETASCAN will be able to determine the location and length of the β-strands present in the amyloid and prion protein sequences. Coupled with a more statistically robust method to estimate pair propensities and the consideration of the amphipathic environment of amyloid β-sheets, the hill-climbing method leads to favorably comparable performance by BETASCAN compared to previous methods, as determined by existing experimental data.

Results

BETASCAN was designed, in principle, to predict parallel β-structure in all cases where the two surfaces of the β-sheet have significant environmental differences. Our strongest subset of interest within this area of competence was the set of prion and amyloid proteins. We therefore tested BETASCAN on five amyloids with known structures and a set of aggregating proteins. In order to verify the accuracy of BETASCAN predictions, we ran BETASCAN on a non-redundant set of crystallized parallel β-helix proteins. This set of structures provided the closest analogue to prion and amyloid proteins with detailed crystal structures available.

Test sets

In addition to testing BETASCAN and competing programs on amyloids, we also test them on their ability to detect β-strands in a superfamily of parallel β-folds with solved crystal structures, namely the parallel β-helices. Here we test BETAWRAPPRO (the improved version on the BETAWRAP algorithm specifically designed for predicting β-helices), BETAPRO (the neural network for predicting β-strands), PASTA, SALSA, TANGO, Zyggregator, and our program BETASCAN for correct detection of β-strands in a β-helix data set. In addition, we compare our program's predictions to those of the other algorithms in light of the known experimental structural evidence for amyloid proteins.

Available verification data

X-ray crystallography results were available from 34 independent, non-redundant structures of β-helix sequences excluded from the pairwise and singleton probability tables (Supplementary Table 1).

In addition, deuterium-exchange solid-state NMR β-sheet detection results were available for amyloid A-β [80, 126], the *Podospora* Het-s prion [81, 82], portions of α-synuclein [144], and the PHF43 fragment of the tau protein [145]. A theoretical model of amylin/islet amyloid polypeptide [146] was also used in the verification of PASTA, and its analysis was included as well. An additional structure of A-β was considered [147], but was too low in resolution to determine lengths of component β-strands.

Output formats

As an example of typical BETASCAN visual outputs, Figure 3-1 presents sample outputs from BETASCAN for the β-helix domain of *Erwinia crysanthemi* pectate lyase C. The heat-map at top (Figure 3-1A) depicts the assignment of a likelihood score to each point on a lattice of possible β-strands. In these graphs, the starting point of a putative β-strand is indicated horizontally, and the length of such a β-strand increases vertically. Organized in this fashion, a likely β-strand appears as a triangular signal of high probability against a low-probability background; the strand location and length may be read at the triangle's apex. The residues on the strands may face in one of two directions (starting inward or outward, relative to fibril core); therefore, two graphs are presented to depict the effects of residue orientation.

The strands with local maximal likelihood were calculated for output as described in Materials and Methods; Figure 3-1B offers a concise version of the results. Here, the potential β-strand lengths and locations are depicted horizontally; the vertical axis indicates the score for each potential strand. An analogous procedure was then executed for all strand-pairs, resulting in the set of local maximum likelihood strand-pairs depicted in Figure 3-1C. As in Figure 3-1B, strand-pairs lengths and locations are depicted, with the horizontal and vertical axes indicating the starting points in the sequence of the first and second strands of the pair.

Predictions of specific β-strands and strand-pairs may now be made directly from the sets of local maximal likelihood structures. For instance, the marked strands and strand-pairs in Figures 3-1B and 3-1C are the set of non-overlapping structures with the highest score. These correspond well to the β-strands and strand-pairs observed in the PDB-deposited crystal structure (purple bars in Figures 3-1B and 3-1C). Confidence in the prediction for any strand, strand-pair, or subsequence thereof may be inferred from the additional predictions for the location.

Verification from crystal structures of β-helices

β-helices have been widely noted as the closest globular protein analogue of amyloid and prion structures [52, 115, 136, 148]. Because of this similarity, the β-helices were removed from consideration during the computation of the probabilistic database. Therefore, these structures formed a useful test set to evaluate the accuracy of the BETASCAN algorithm in β-strand detection. The BETASCAN results for the non-orthologous β-helices (listed in Supplemental Table 3-S1) were compared to the STRIDE analysis of β-strands in crystal structures. Statistics were collected on the accuracy of predictions by strand and by residue, and on the accuracy of left- and right-edge locations.

The accuracy of BETASCAN, counted by strand and by residue, is depicted in Figure 3-2A. Examined strand-by-strand, BETASCAN had an effective sensitivity of 94-96% to correct strands. In addition, as long as the maximum β-strand length is equal or greater to the average β-strand length, BETASCAN achieved 80% or greater sensitivity, measured residue-by-residue, for this data set. As shown in Figure 3-2B, the error in the predicting the left and right edges of each strand was between one and two residues each. The error in edge localization was minimized when the maximum β-strand length was closest to the average β-strand length, and increased considerably when longer strands were considered. The residue-by-residue sensitivity is thus reflective of the error in edge localization.

Our hypothesis that pairwise probabilities reflect the occurrence of β-strands in structures necessarily implied that the scores discriminate between β-forming subsequences and sequences that form loops or other structures. Figures 3-2C, 3-2D, and 3-2E describe the sensitivity and specificity or positive predictive value of BETASCAN scores for residues and for strands. Sensitivity and specificity, measured by residue, were markedly reduced if strands of sufficient length are not considered, while additional strand length had little or no effect. However, strand-by-strand sensitivity, specificity, and PPV were slightly improved. For all lengths and scores examined, negative predictive value (NPV) was 95% or higher.

While portions of a loop may be found in β-conformation without the distinctive hydrogen bonds of a β-sheet, the majority of β-strands are found in β-sheets. Therefore, the scores of the predicted pairwise contacts may indicate whether a given postulated β-strand is present in native or amyloid structures. With this hypothesis in mind, a filter was devised to exclude strands without significant associated pairwise contacts from the BETASCAN single-strand maxima results. A strand was considered to have poor pairwise contacts if the summed scores of pairwise contacts with the strand as the first element was less than some threshold value. Figures 3-2F, 3-2G, 3-2H reveal the effect of increasingly filtering pairwise-poor strands on sensitivity, specificity, and strand-by-strand PPV for the β-helix test set. The best results for this set, as determined by the receiver-operating characteristic method, were at a pairwise filter threshold score of 17. For strands with summed pairwise scores above this value, 91-94% sensitivity and 84-94% specificity was observed, with the longest allowed strand lengths yielding the best statistics. Measured residue by residue, 72-83% sensitivity and 74-81% specificity were achieved, with the best statistics observed at a maximum length of four residues. PPVs achieved were 68-70% for strands and 70-73% for residues.

Comparison to BETAWRAPPRO results.

Comparisons were also made between the BETASCAN β-helix results and the highest-scoring predictions of BETAWRAPPRO, the latest BETAWRAP algorithm [117]. Since the BETAWRAPPRO algorithm incorporates structural information specific to β-helices, comparison of BETASCAN and BETAWRAPPRO indicates the relative utility of structure-specific knowledge.

BETAWRAPPRO predicted 276 strands in its top results for each of the β-helices studied. When compared to the 631 β-strands in the crystal structures, 189 were found to correspond, for a sensitivity of 30% and a positive predictive value of 68.4%. Of these 189 strands, 183 strands were considered matched by BETASCAN under the same conditions used for matching in the BETASCAN analysis above. These results were unchanged by changes in maximum β-strand length. Thus, BETASCAN effectively reproduces the correct results of BETAWRAPPRO without structure-specific knowledge. While markedly increasing sensitivity to β-strands, especially to those outside the canonical β-strand pattern, BETASCAN also maintains the positive predictive value achieved by BETAWRAPPRO.

Verification from solid-state NMR analyses of amyloids, and comparison with PASTA and SALSA

Solid-state NMR analysis was used by Petkova et al. [126], Luehrs et al. [80], Ritter et al. [81], and Wasmer et al. [82] to determine strand-pair contacts in A-β 1-42 and the *Podospora* Het-s prion. Briefly, ^{1}H-NMR signals were taken before and after one week of immersion in D_2O. Deuterium exchange occurred in all locations except where retarded by the energy wells of hydrogen bonds, allowing the identification of residues taking part in β-sheet hydrogen bonding.

The structure of A-β 1-42 (Figure 3-3A) under differing conditions was determined independently by Petkova et al. [126] and Luehrs et al. [80]. As determined by Luehrs, the structure included two β-strands formed at residues 15-24 and 30-42, each forming in-register interchain strand-pairs. The structure as determined by Petkova included a strand from 10-14 and a region from 30-35 that was ambiguously determined as one or two strands. Predictions by PASTA and SALSA suggested β-structure in the regions 10-22 and 29-42 without elaboration. The BETASCAN algorithm, as its top specific prediction, produced β-strands at residues 9-13, 15-22, and 30-42.

The heterokaryon compatibility prion Het-s from *Podospora anserina* [47] was found by Ritter et al. [81] to form four β-strands, with one β-sheet composed of alternating copies of β-strands 1 and 3, and another β-sheet similarly composed by strands 2 and 4 (Figure 3-3B). The results of Wasmer et al. [82] indicated breaks in each of these four β-strands, thus predicting a total of eight closely spaced strands. In addition, the new results indicated a reversal of orientation at the breaks in strands 1 and 3. PASTA predicted two strands and the possibility of a third, corresponding to Ritter's strands 2, 3, and 4. The BETASCAN algorithm strongly predicted Ritter's strands 2, 3, and 4 at their full length. While BETASCAN's prediction matched only the C-terminal half of Ritter's strand 1, it matched both strands 1a and 1b of the Wasmer model at lower probability. Wasmer strands 2a, 2b, 3a, 3b, and 4a were all indicated at high probability, and Wasmer strand 4b at the same probability as strand 1a. Each of the strand-pairs observed by Ritter and by Wasmer was found in the strand-pair set predicted by BETASCAN, although the signal was not clearly distinguishable from other potential pairings.

α-synuclein has been analyzed by Heise et al. [144] to contain a total of seven strands. The two highest-scoring, the third and sixth strands, were detected by both PASTA and BETASCAN with high accuracy (Figure 3-3C). SALSA predicted the first two, with a large and vague prediction of amyloid propensity covering the remaining strands. While some predictions were low-scoring, only BETASCAN indicated the possibility of the seven strands detected by experiment.

Amylin, also known as islet amyloid polypeptide (IAPP), was modeled by Kajava et al. [146] to be an in-register amyloid composed of three strands. PASTA predicted an amyloidogenic region from 15-32. BETASCAN results (Figure 3-3D) suggested two of the three strands predicted by Kajava, part of the third strand (30-33), and an additional strand at residues 3-7. This potential strand may be related to the intrachain cysteine bond between residues 3 and 8.

Aggregation in the tau protein centers on the repeat domain, which takes the conformation of random coil in the native state [145]. Interest has more specifically centered on the protease-resistant PHF43 sequence, though other regions of the protein product have been suggested to play roles [149]. Trovato and colleagues only analyzed the PHF43 domain itself, verifying the importance of the hexapetpide VQIVYK at residues 306-311. Here, the region between residues 205 and 441 is analyzed. A more extensive run of the PASTA algorithm finds strands at 258 and 338. SALSA weakly detects strands at about 235, as well as at 390 and 410. The more expansive BETASCAN analysis presented in Figure 3e underscores the importance of residues 306-311, as it is the most likely β-strand to form in the entire tau protein; it also detects strands at 255, 338, and 390-410.

Comparison across multiple prediction programs

A synopsis comparing the predictions of BETASCAN to those of PASTA (as provided in [143]), SALSA (as provided in [141]), TANGO [138], Zyggegator [139], and BETAPRO[137] is presented in Table 3-1. As a control, the results of JPRED [100] and PSIPRED[150] are included to represent traditional secondary structure prediction. Because of the difficulty in translating one program's scores to another's, predictions were indicated as 'strong' or 'weak' depending on relative internal scoring and extent of prediction. For IAPP, Het-s, and α-synuclein, our program BETASCAN was the only program to detect the correct number of strands. All algorithms successfully predicted the strands of A-β, although some did not detect all of strand 2. All algorithms were also able to detect the strongest strands for the tau protein, except that BETAPRO and TANGO did not detect the first strand near residue 235. BETAPRO tended to miss strands, while PASTA, SALSA, and Zyggegator had difficulty separating strands. TANGO tended to miss strands at the edges of β–rich regions.

Analysis of aggregating sequences

We also used BETASCAN to analyze a larger database of sequences derived from proteins observed to aggregate in experimental settings, and the results compared favorably to those of PASTA (see Figure 3-4). The data set previously used for analysis by [138, 142] was considered as a benchmark. However, the redundant content of the set was found to cause a loss of robustness, as determined by an analysis involving the removal of one or two clusters' redundant sequences (see Supplemental Table 3-S2). Therefore, a non-redundant version was calculated by CD-HIT [151] with 40% sequence similarity cutoff, and the resulting 120 sequences (see Supplemental Table S-S3) were analyzed by BETASCAN. The specificity and sensitivity curves for the top β-strand score of each sequence intersect at 81%, which compares favorably to previously reported PASTA results [142].

Discussion

We have introduced the program BETASCAN and showed its improved performance over previous methods for identifying β-strands in parallel β-structures, most importantly amyloid structures. The BETASCAN approach depends upon the idea that, while all sequences display some tendency towards the β conformation, sequence details determine the relative likelihood of β-strand and strand-pair formation at all scales. Thus, sequence has a broad effect not only on secondary structure, but also on the supersecondary structural assembly of β-strands into a β-sheet. This concept is the driving force of both the scoring and maxima-finding algorithms in BETASCAN. The score is designed to allow unbiased comparisons between β-strands differing not only in sequence, as in BETAWRAP, but also by length and orientation. Correspondingly, the maxima-finding algorithm uses these comparisons to explore strand and pair space for locally optimal β-strands and strand-pairs.

In addition, BETASCAN may owe some of its strong performance, compared to PASTA, to its ability to distinguish strands of different lengths in relation to their rate of occurrence in nature. PASTA scores are generated for residues and residue pairs based upon the weighted-sum scores of every potential β-strand that could be formed using that residue or residue pair. In contrast, the emphasis in BETASCAN is placed upon finding specific high-scoring strands. Any region with a high PASTA score will also contain high-scoring BETASCAN predictions, which supply additional information about where strands are likely to begin, end, and pair. The concentration on the strand as the fundamental unit of β-structure also improves residue-by-residue detection of β-structure.

BETASCAN is highly sensitive for potential strands and excellent at determining when sequences will not contribute to β– and amyloid structure (high negative predictive value). However, the effort to identify all potential β-structure variants can cause significant over-prediction of β-structure, as Figures 3-1, 3-2, and 3-3 all reflect. While the highest-scoring strands consistently reflect real structures, only some of the lower-scoring strands are found in experimental data. The low-scoring strand at residue 146 in the 2PEC structure (Figure 3-1B) is an example of a low-scoring strand extant in crystal structure. By synthesizing the singleton and pairwise maxima results of BETASCAN, a better predictive capacity is achieved. The optional pairwise-based filter demonstrated in Figures 3-2F, 3-2G, and 3-2H can identify structural strands with better performance than exclusion by low score alone, and retains low-scoring strands that readily form strand-pairs. Additional factors, discernable by experimental data or by more astute analysis, may be used as additional specificity filters to distinguish which *potential* β-strands are contributors to either amyloid or native structures. However, the knowledge that amyloid structures include multiple 'strains', may be heterogeneous even within a single fibril, and frequently include β-strands not found in the native fold of the parent protein, argues for the inclusion of hypothetical β-strands in analysis until excluded by evidence.

Most interestingly, BETASCAN is capable of revealing details and variants of protein structure previously inaccessible to computational methods. For instance, two solid-state NMR studies of A-β protein [80, 126] produced conflicting results in the region between residues 30 and 42. However, each result is reflected in the BETASCAN results (Figure 3-3A), where both short strands corresponding to the Petkova results and long strands corresponding to the Luehrs results are high-scoring maxima. Likewise, two solid-state NMR studies of Het-s [81, 82] were differentiated by the presence or absence of interruptions in the β-strands. Both the elongated and truncated versions of these strands were isolated by the maxima-finding subroutine of BETASCAN. Thus, BETASCAN can distinguish the local attractor states that the two pairs of experimental samples occupied, opening the possibility of understanding the influence of environmental conditions and/or folding kinetics on "prion strains" and other amyloid folding variations.

BETSACAN forms part of a synergistic strategy for the evaluation of all-β structure. Additional β-strand specificity may be found using experimental contextual clues, such as discernment of physical attributes, specific links between residues such as cysteine bonds and side-chain ladders, and constraints on the conformational space of the amyloid. The variants indicated by BETSACAN may also be distinguished *in vitro* or *in* vivo by additional exterior factors, such as pH, osmolarity, and the presence of seeding factors or chaperone proteins. By distinguishing folding variants and providing specific location and likelihood data, BETASCAN thus boosts to the efficacy of both experimental and computational efforts to understand the parallel β-sheets of amyloids and prions.

Materials and Methods

Algorithmic strategy

BETASCAN calculates likely β-strands and strand-pairs for an input sequence presumed to contain parallel β-structure. Every contiguous subsequence of length 2 up to *k* is initially considered as a possible parallel β-strand (*k* defaults to 13, the length of the longest parallel β-strand in our source database). For each pair of possible strands, a score is determined corresponding to a prediction of how likely their pairing would be (see **Strand state propensity** and **Pair state propensity** below). This probability is based on the observed preferences for each pair of residues in the strands to be hydrogen-bonded (see **Probability tables** below). Maxima-finding algorithms, also known as "hill-climbing" algorithms, are then used to detect all local maxima of formation propensity across strand-pair space (see **Maxima finding** below). The outputs of BETASCAN are score-ordered lists of all locally optimal strands and strand-pairings.

Note that BETASCAN can return strands and/or strand-pairs that inconsistently overlap in the local-optimum list. These results reflect the potential, under differing conditions, for alternate β-strand folding patterns.

Probability tables

Pairwise probability tables to capture the preference for each pair of amino acids to be hydrogen-bonded in a β-sheet was estimated using a method similar to McDonnell *et al.* [117] Briefly, the non-redundant structures of the Protein Data Bank [2] as of June 8, 2004, were filtered to remove the set of structures in Supplementary Table 3-S1. These structures, including all three-stranded right-handed β-helices, were removed for two reasons. First, their similarity to known and theorized amyloid structures was considered a potential source of bias. Second, their removal allowed their use as a control test set (see **Test set construction** below). The STRIDE algorithm [152] was used to on the remaining structures to find all amphipathic β-sheets, namely β-sheets with solubility differences between its two faces. The frequencies of occurrence of hydrogen-bonded pairs (X_1,X_2,θ) were tabulated, where the orientation θ distinguished β-sheet faces with lesser (zero) or greater (one) solubility. The frequencies were normalized to sum to 1, generating the 20x20x2 pairwise statistical table *W.* (Symbols identify structures as indicated in Figure 3-5A.)

V, the 1x20x2 singleton probability table, represents the propensity of a side-chain X_1 to be present in an amphipathic β-strand. *V* was calculated by summing the pairwise probability tables across rows.

Background probability tables were generated by counting single amino acid frequencies across all protein sequences (not only β-structures). Background probability pairwise tables were formed by squaring the singleton frequencies, corresponding to an independence assumption for the null hypothesis. The default table $C_{allproteins}$ is derived from the release 50.4 (July, 2006) of the SWISS-PROT database. Prion and amyloid sequences derived from genomes of yeast species with amino acid distributions potentially biased by sparse GC content, as determined by whole-genome phylogenetic analysis, were analyzed using a table $C_{allyeasts}$ derived from the genome of *Saccharomyces cervisiae* as of July, 2006.

Strand state propensity

Figure 3-5A serves as a visual reference for the following formulas. For a possible β-strand starting at position *p* with length *l* and orientation *o*, the propensity *h(s, p, l, o)* of formation for a strand state *(p, l, o*) forming from a polypeptide sequence *s* is calculated as the ratio of the propensity *f(s, p, l, o)* of the strand sequence to form a β-sheet and the propensity *g(s, p, l*) of the strand sequence to occur randomly. The background propensity *g* is calculated as the product of the occurrence rates *c* of each residue in the possible strand, derived from the background table *C*. (The table $C_{allproteins}$, as derived above, is the default for *C*.) The strand propensity *f* calculation similarly begins by multiplying each residue's frequency *v* in the singleton probability table *V* (as calculated above) for the orientation *o*. The calculation of *f* also includes dividing by a length correction term to model the effect of length on the formation of a β-strand. The length correction term is included to enable comparison of strands with different lengths on an equal basis, a requirement for the maxima-finding subroutine (see **Maxima finding** below). The form of the correction was chosen to reflect the observed histogram of parallel strand-pair lengths in the PDB [153]. The best-fit curve of this independently derived data was found to be a Poisson distribution with parameters (*l* - 1, 3.15). A potential explanation for the Poisson distribution is the modeling of each residue's addition to the β-strand as a Poisson process.

Including the correction term, the propensity of formation is therefore

$$h(s,p,l,o)=\frac{f(s,p,l,o)}{g(s,p,l)}=\frac{\dfrac{\prod_{ti=0}^{l-1} v(s_{p+t},o)}{\ln\left(e^{-\lambda}\lambda^{l-1}/(l-1)!\right)}}{\prod_{t=0}^{l-1} c(s_{p+t})}.$$

Pair state propensity

Given a second strand starting at position *q*, the propensity *k(s, p, q, l, o*) of formation for a parallel pair state (*p, q, l, o*) from one or more copies of a polypeptide sequence *s* is calculated in a fashion similar to that above. (See Figure 3-5A for a complete visualization of the structure under consideration.) The calculation of *k* incorporates the single-strand propensity *h(s, p, l, o)* of the first strand, the composition propensity *g(s, q, l, o)* of the second strand, and the pairwise propensity *j(s, p, q, l, o)* of the two strands' adjacency. The pairwise propensity *j*, is calculated from the pairwise propensity table *W* by multiplying terms *w* for each pair of residues and dividing by the length-correction term. The inclusion of *h* in the calculation of *k* is necessitated by the form of *W*, which pre-supposes the formation of the first β-strand.

$$k(s,p,q,l,o)=h(s,p,l,o)\frac{j(s,p,q,l,o)}{g(s,q,l)}=\frac{\dfrac{\prod_{t=0}^{l-1} v(s_{p+t},o)}{\ln\left(\dfrac{e^{-\lambda}\lambda^{l-1}}{(l-1)!}\right)}\left(\dfrac{\prod_{t=0}^{l-1} w(s_{p+t},s_{q+t},o)}{\ln\left(\dfrac{e^{-\lambda}\lambda^{l-1}}{(l-1)!}\right)}\right)}{\prod_{t=0}^{l-1} c(s_{p+t})\left(\prod_{t=0}^{l-1} c(s_{q+t})\right)}$$

Maxima finding

The maxima finding subroutine of BETASCAN extracts the most likely strands and strand-pairs by asking if a single change to the strand or strand-pair would result in a higher probability of formation. Not all transitions between strand states or strand-pair states are physically realizable in one step. The constraints on the strand and pair spaces may be described as lattices, with nodes corresponding to each potential strand or strand-pair and edges corresponding to the conformational changes required to form one potential strand from another. Edges may be formed by the addition or removal of residues at either end, by the reversal of strand orientation (180° rotation around the long axis of the strand), and for strand-pairs, the shearing of the strands' interactions by one or two residues. The possible transitions, and the lattices so created, are depicted in Figures 3-5B and 3-5C.

The BETASCAN method assigns a propensity to each node of these lattices, with the highest score corresponding to the most likely strand or strand-pair. A hill-climbing method, which searches each node's adjacent neighbor for a higher score, is then executed across the entirety of strand and pair space. All nodes with at least one such neighbor are removed from consideration. The remaining sets of strand states and pair states are local propensity maxima in strand and pair space. Together with their propensity scores, these sets form the output of the BETASCAN algorithm.

To allow comparisons with other prediction methods and to highlight the most relevant strands and pairs, filtering was applied. Only those strands and pairs with positive log-odds propensity scores, indicating a propensity of formation greater than random sequence, were selected. For the results in the **Comparison** sections, a consistent set of strands was chosen by repeatedly selecting the highest scoring strand that was consistent with all previously selected strands until the list of potentially consistent strands with scores more likely than random was exhausted.

Test set construction

3-D crystal structures of the β-helices removed from the probability database (listed in Supplementary Table 3-S1) were downloaded from the Protein Data Bank [2]. The β-helix test set structures were chosen as non-redundant representatives of SCOP families, without substrates or other co-crystallized molecules. β-strands and strand-pairs were identified using STRIDE [152] as described in McDonnell et al. [117].

The all-β secondary structures of the input sequences were verified. For the β-helix sequences, 3-D X-ray crystallography was available to guarantee secondary structure details [116]. In addition, the sequences were analyzed by the secondary structure prediction program DSSP [154] to localize α-helical content.

Comparison calculations

BETASCAN, BETAWRAPPRO, BETAPRO, and PASTA were run on the 34 β-helix structures listed in Supplemental Table 3-S1. Because β-helix strands are, on average, just over four residues in length, BETASCAN runs were executed using maximum β-strand lengths of 3, 4, 5, 6, and 7, as well as with the default length of 13. A single consistent set of predicted strands was selected from the set of all predicted strands by repeatedly selecting the strand with the highest positive score that failed to overlap either any previously selected strand or any α-helix (as observed in crystal structure by DSSP [154]. The set of predicted strands was compared to the β-strands calculated from crystal structures by the program STRIDE [152] according to the settings of McDonnell et al. [117]. The STRIDE predictions were taken as the true positive β-strands for this class.

For each real strand, if at least one predicted strand overlapped more than 50%, a match was recorded. In addition to the fraction of matching crystal and predicted strands, statistics were collected on the number of matching residues and on the predictions of β-strand 'edges'. The N- and C-terminal ends of the crystal strand were compared, respectively, to the N- and C-terminal ends of the N- and C-most matching predicted strands. In most cases, only one predicted strand matched the crystal strand, and so the ends compared were the N- and C-terminal ends of the prediction.

To generate ROC and sensitivity/PPV curves (Figures 3-2C, 3-2D, and 3-2E), the output of BETASCAN was repeatedly analyzed with a lower-bound score cutoff, which was varied from 0 to +2 units. For the poor-pairwise-contact filter (Figures 3-2F, 3-2G, and 3-2H), ROC and sensitivity/PPV curves were generated as follows. For each strand in the BETASCAN singleton results, the scores of all strand-pairs sharing the first residue of the first pair (parameter *p*) with the predicted β-strand were summed. The β-strand was removed from the prediction if the summed score was less than the summed-score cutoff, which was varied from 0 to +40 units to produce the curves shown.

Comparison to BETAWRAPPRO results

The top hit of BETAWRAPPRO was taken as the prediction for each of the 23 structures; this yielded a set of 276 strands predicted by BETAWRAPPRO. A BETAWRAPPRO strand was taken to be a "correct prediction" if its N-terminal end was within 3 residues of a crystal structure strand as determined by the DSSP analysis found at the PDB website [2].

Comparison to BETAPRO, PASTA, PSIPRED, JPRED, SALSA, TANGO, and Zyggregator

BETAPRO, PASTA, SALSA, TANGO, PSIPRED, and JPRED were executed using all default settings. Zyggregator was used in fibrillar mode. To avoid bias and in keeping with author suggestions, no additional secondary structure descriptions or alignments were input to BETAPRO or JPRED. To overcome differences in scoring methods, predictions in Table 1 were rated as 'strong' (S), 'weak' (w), 'no prediction' (n), or no data available (x). A prediction was rated as 'strong' if more than 2/3 of the strand's length was predicted and if the internal rating system of the program (if present) scored any portion of the strand as greater than 50% of the peak prediction for that sequence. The prediction was rated 'weak' if the above conditions were not satisfied, but more than two residues of the strand were predicted at any confidence level. A prediction was indicated as (+) if the requirements for a weak prediction were met, but no separation existed between strand predictions.

Figures

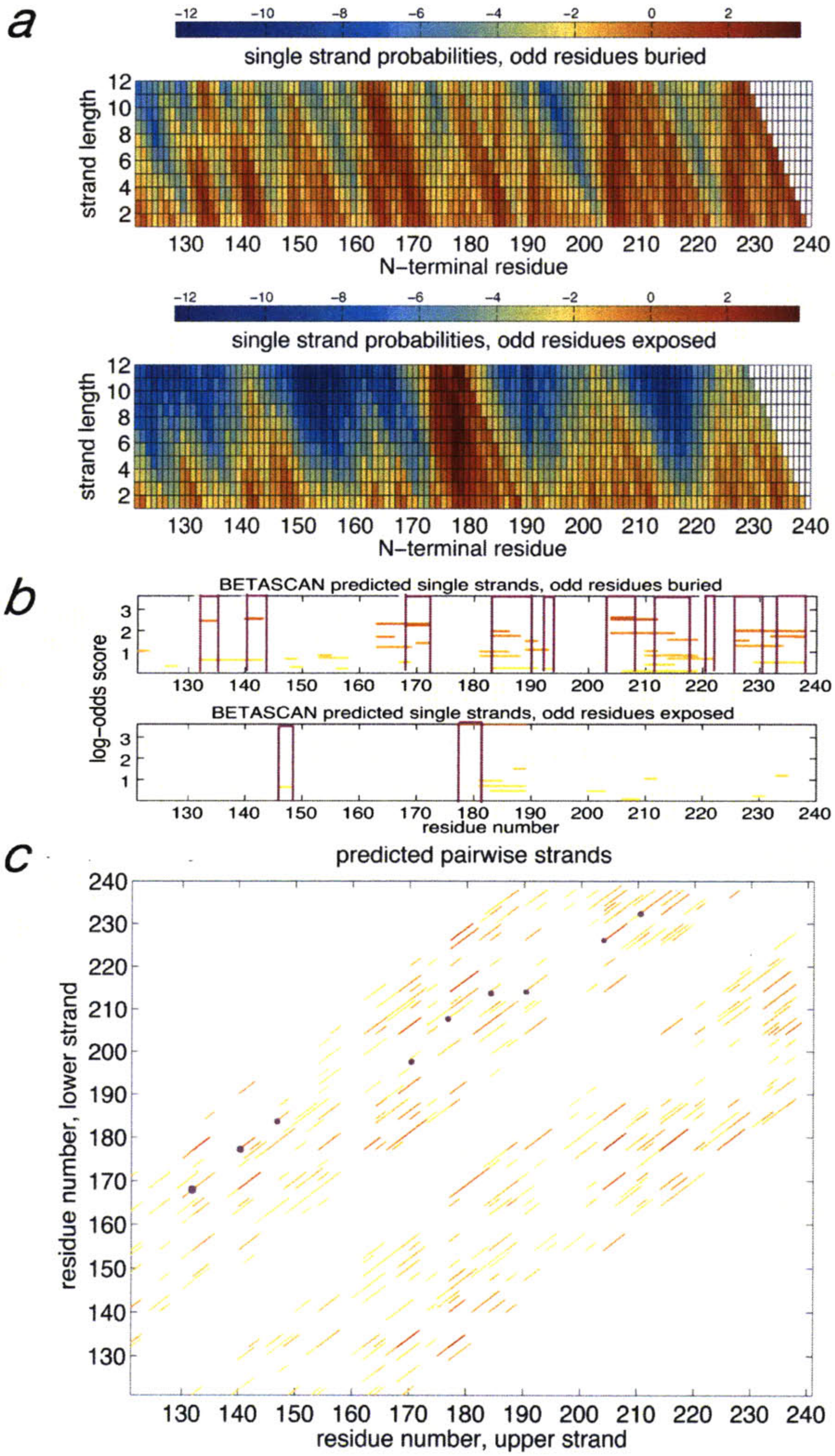

a
-12
-10
-8
-6
-4
-2
0
2
single strand probabilities, odd residues buried
strand length
12
10
8
6
4
2
130 140 150 160 170 180 190 200 210 220 230 240
N−terminal residue
-12
-10
-8
-6
-4
-2
0
2
single strand probabilities, odd residues exposed
strand length
12
10
8
6
4
2
130 140 150 160 170 180 190 200 210 220 230 240
N−terminal residue
b
BETASCAN predicted single strands, odd residues buried
log-odds score
3
2
1
130 140 150 160 170 180 190 200 210 220 230 240
BETASCAN predicted single strands, odd residues exposed
3
2
1
130 140 150 160 170 180 190 200 210 220 230 240
residue number
c
predicted pairwise strands
residue number, lower strand
240
230
220
210
200
190
180
170
160
150
140
130
130 140 150 160 170 180 190 200 210 220 230 240
residue number, upper strand

Figure 3-1, Sample output of the BETASCAN algorithm. The results for the β-helix domain of pectate lyase C are shown. *a,* heat-map of all β-strand probabilities. The horizontal axis indicates the N-terminal residue of potential β-strands, while the vertical axis indicates strand length. The upper and lower boxes display results for the two orientations of the strand. Colors indicate propensity of strand formation. Red indicates above-background probability, while blue indicates below-background probability. *b,* predicted most likely β-strands based on single strand probabilities. BETASCAN predictions are marked as horizontal lines, shading from red (maximum predicted score) to yellow (zero score, i.e., probability equal to background). The horizontal axis indicates the N-terminal residue of potential β-strands, while the vertical axis indicates the log-odds propensity. Overlapping strands represent alternate folding patterns with indicated likelihoods. Purple brackets indicate experimentally determined β-strands as derived from the PDB structure. *c,* predicted most likely strand-pairs based on pairwise probabilities. Purple dots indicate the N-terminii of experimentally determined strand-pairs as derived from the PDB structure.

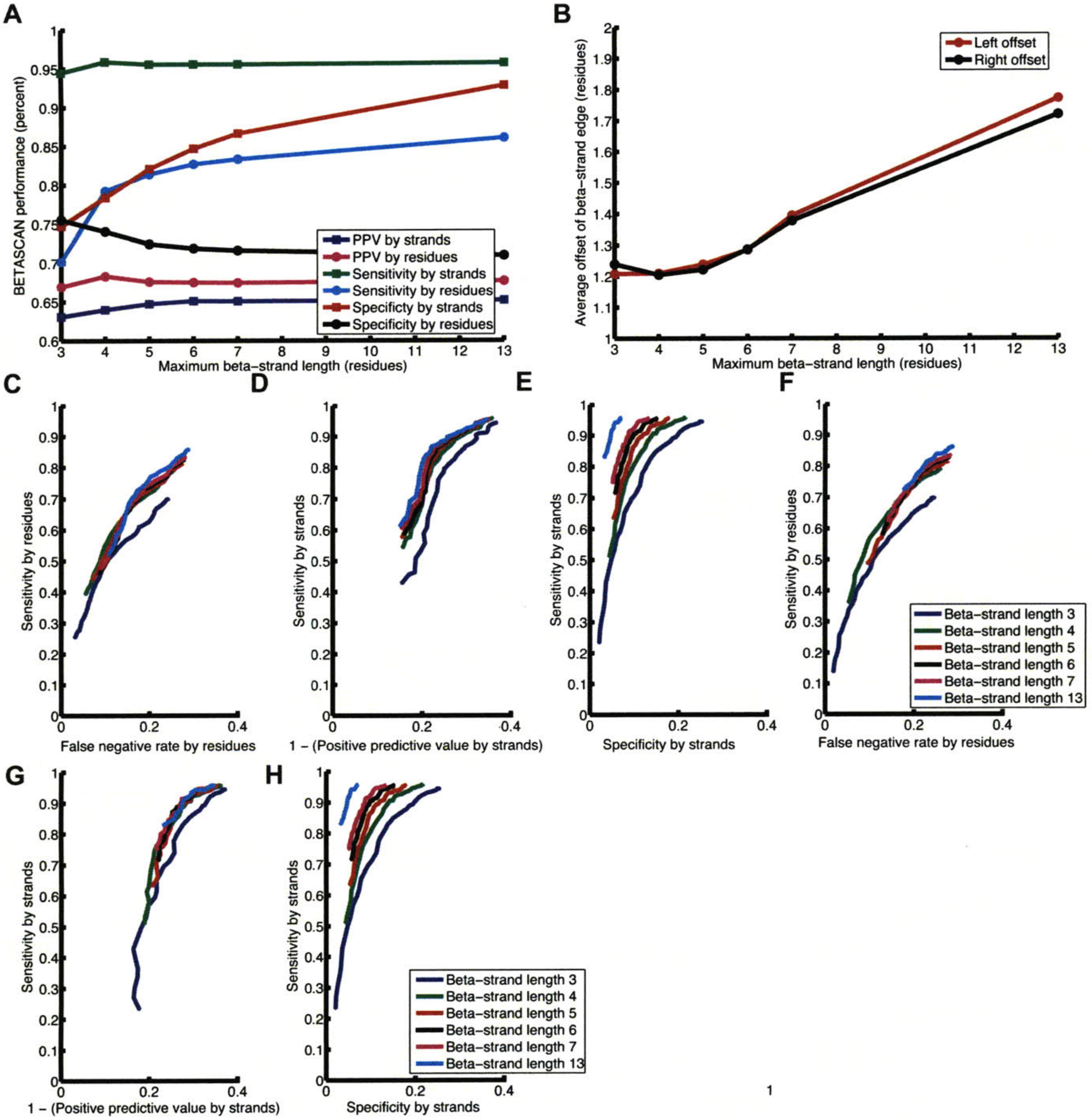

A
BETASCAN performance (percent)
Maximum beta–strand length (residues)
PPV by strands
PPV by residues
Sensitivity by strands
Sensitivity by residues
Specificty by strands
Specificity by residues
B
Average offset of beta–strand edge (residues)
Maximum beta–strand length (residues)
Left offset
Right offset
C
Sensitivity by residues
False negative rate by residues
D
Sensitivity by strands
1 – (Positive predictive value by strands)
E
Sensitivity by strands
Specificity by strands
F
Sensitivity by residues
False negative rate by residues
Beta–strand length 3
Beta–strand length 4
Beta–strand length 5
Beta–strand length 6
Beta–strand length 7
Beta–strand length 13
G
Sensitivity by strands
1 – (Positive predictive value by strands)
H
Sensitivity by strands
Specificity by strands
Beta–strand length 3
Beta–strand length 4
Beta–strand length 5
Beta–strand length 6
Beta–strand length 7
Beta–strand length 13
1

Figure 3-2, Statistics on BETASCAN accuracy in the set of β-helices. *a,* effect of maximum β-strand length on sensitivity and positive predictive value of BETASCAN, measured by strand and by residue. *b,* effect of maximum β-strand length on average absolute value difference in predicted and crystal-structure observed β-strand edges. *c-e:* effectiveness of BETASCAN singleton scores in β-structure prediction*; c,* ROC curve calculated residue-by-residue; *d,* graph of sensitivity against (1-PPV) calculated strand-by-strand; *e,* ROC curve calculated strand-by-strand. *f-h:* effectiveness of BETASCAN pairwise scores in β-structure prediction; *f,* ROC curve calculated residue-by-residue; *g,* graph of sensitivity against (1-PPV) calculated strand-by-strand; *h,* ROC curve calculated strand-by-strand.

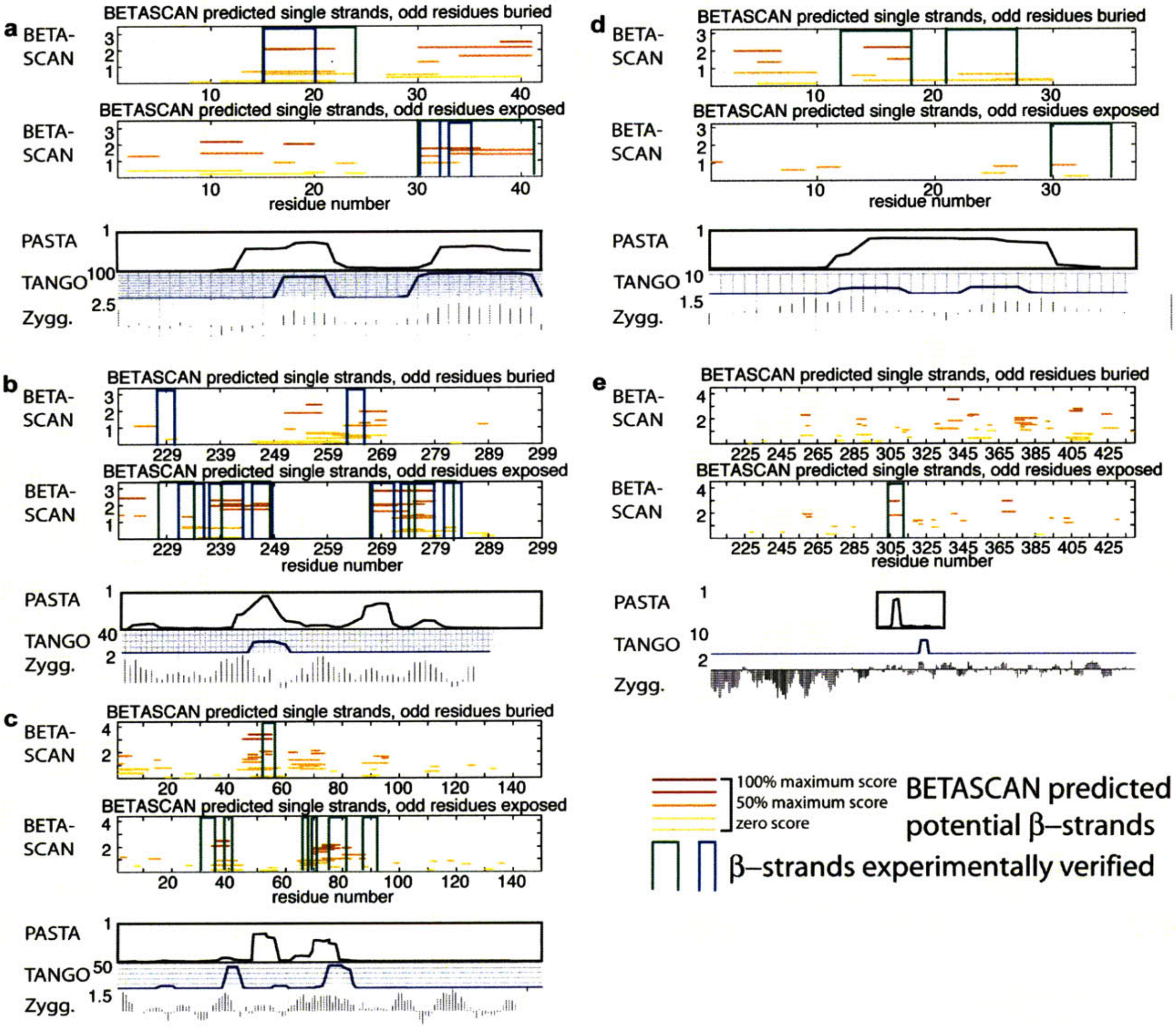

a
BETASCAN predicted single strands, odd residues buried
BETASCAN predicted single strands, odd residues exposed
BETA-SCAN
residue number
PASTA
TANGO
Zygg.
b
c
d
e
100% maximum score
50% maximum score
zero score
BETASCAN predicted potential β–strands
β–strands experimentally verified

Figure 3-3, BETASCAN output for amyloid and prion proteins with experimentally determined β-structures. Green vertical brackets indicate experimentally derived locations of β-strands; blue brackets indicate locations determined by a separate method. In the same manner as Figure 1b, BETASCAN predictions are marked as horizontal lines, shading from red (maximum predicted score) to yellow (zero score, i.e., probability equal to background). Overlapping lines indicate alternate folding patterns for the β-strands, with indicated probability. Two graphs are included to display the results for each orientation of the strand. For purposes of comparison, the set of highest-scoring non-overlapping strands in the BETASCAN single-strand prediction was taken as the predicted structure. Corresponding outputs of PASTA [142, 143], TANGO [138], and Zyggregator [139] are displayed below the BETASCAN results. Refer to Table 1 for a summary of the correspondences of these predictions. *a,* amyloid-β structure as determined by Luehrs et al. [126] (green) and Petkova et al. [80] (blue); *b,* het-S structure as determined by Ritter et al. [81] (green) and Wasmer et al. [82] (blue); *c.* alpha-synuclein structure as determined by Heise et al. [144]; *d,* amylin structure as determined by Kajava et al. [146]; *e.,* tau protein fragment PHF43 structure as determined by von Bargen et al. [145].

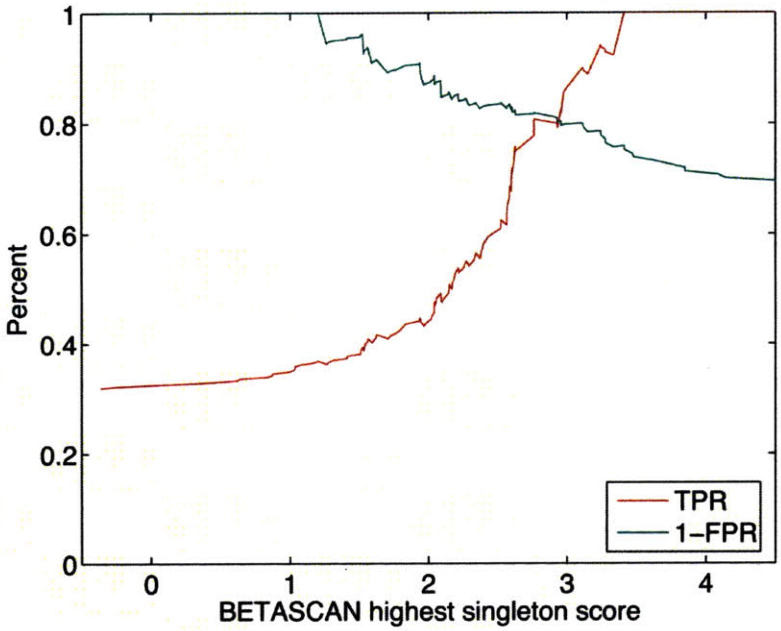


Figure 3-4, Sensitivtiy/specificity curve of BETASCAN highest singleton result for 120 non-redundant sequences experimentally observed for aggregation potential, collated by [138] and clustered by CD-HIT [151].

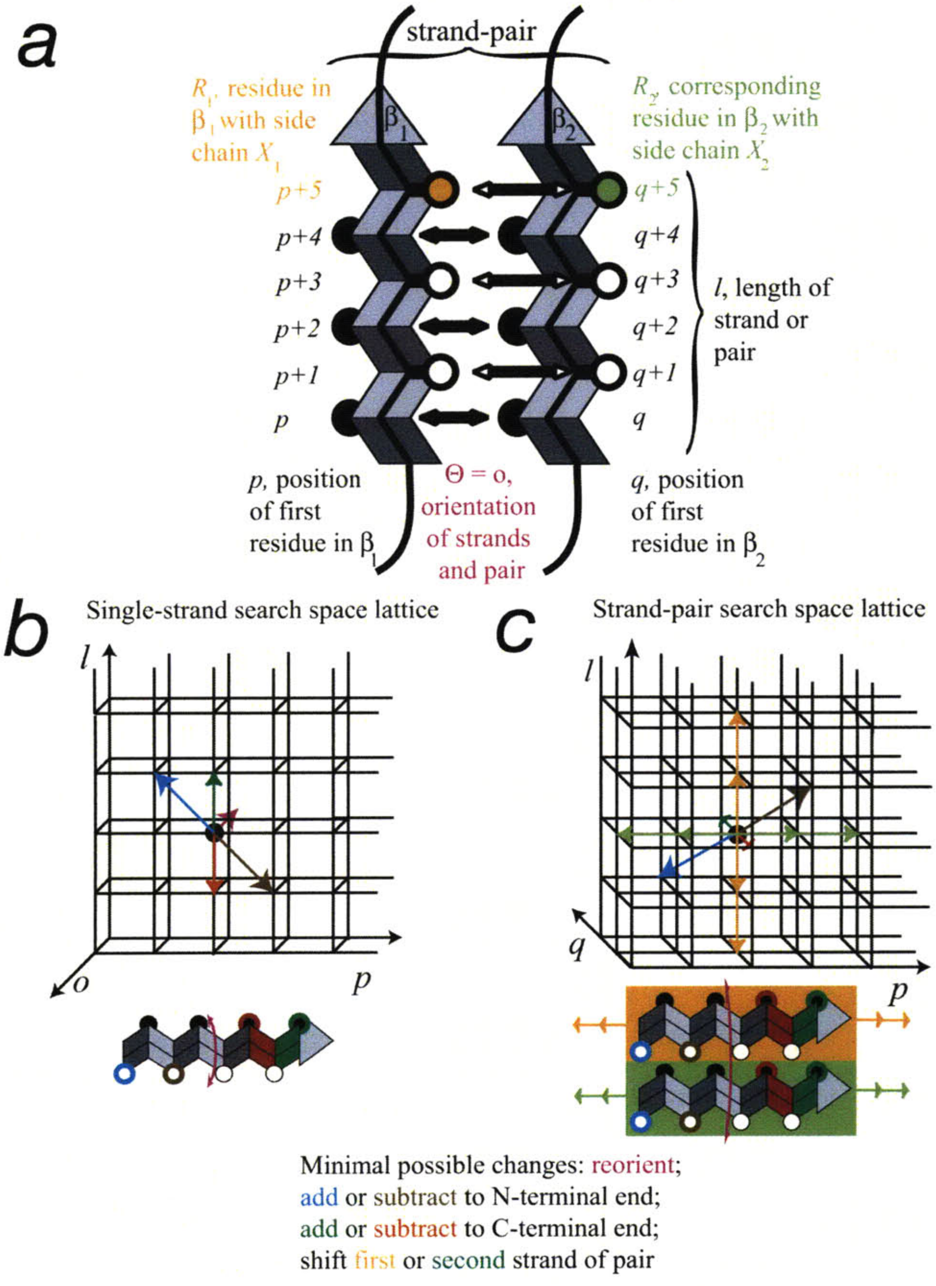

a
strand-pair
R_1, residue in β_1 with side chain X_1
β_1
β_2
R_2, corresponding residue in β_2 with side chain X_2
p+5
p+4
p+3
p+2
p+1
p
q+5
q+4
q+3
q+2
q+1
q
l, length of strand or pair
p, position of first residue in β_1
Θ = o, orientation of strands and pair
q, position of first residue in β_2
b
Single-strand search space lattice
l
o
p
c
Strand-pair search space lattice
l
q
p
Minimal possible changes: reorient;
add or subtract to N-terminal end;
add or subtract to C-terminal end;
shift first or second strand of pair

Figure 3-5, Relationships between physical features of β-sheet components and the definition of the computational search spaces. *a,* variable definitions as used in *Materials and Methods*. The two vertical β-strands form a single strand-pair, with odd residues labeled in white and even residues in black. The strands share the same orientation *o* and extend from *p* to *p+l* and from *q* to *q+l*. *b,* structure of the lattice of the β-strand search space defined by the variables *p* (location), *l* (length), and *o* (orientation). Changes in the parameters of a β-strand are physically possible in a single step only along the paths marked by arrows. The arrowheads therefore define the relative locations queried by the maxima-finding algorithm at each point. *c,* structure of the lattice of the strand-pair search space defined by the variables *p* (first strand location), *q* (second strand location), *l* (length) and *o* (orientation, not shown). In addition to the physically possible changes in *b*, shifts of one or two residues in the relative strand positions are possible. Arrowheads indicate the relative locations queried by the strand-pair maxima-finding algorithm for each point.

Tables

Table 3-1, Comparison of BETAWRAP results to previous algorithms. Letters indicate strength of prediction: S, strong (complete prediction); w, weak (missing >30% of length or <50% confidence); (+), prediction without strand boundaries; n, not predicted; x, data not available.

Protein	A-β		Het-s				α-synuclein							IAPP			Tau			
Strand	1	2	1	2	3	4	1	2	3	4	5	6	7	1	2	3	1	2	3	4
BETASCAN	**S**	**S**	**w**	**S**	**S**	**w**	**w**	**S**	**S**	**w**	**S**	**S**	**S**	**S**	**w**	**w**	**w**	**S**	**S**	**S**
BETAPRO	S	w	n	n	n	S	n	S	n	w	n	n	w	n	n	n	n	S	S	w
TANGO	S	S	n	S	n	n	n	S	w	w	w	S	n	w	w	n	n	S	n	n
Zyggregator	S	S	+	+	+	+	n	S	w	+	+	+	w	S	+	+	S	w	S	n
PASTA	S	S	n	S	w	n	n	w	S	w	w	S	n	+	+	n	S	S	S	w
SALSA	S	S	x	x	x	x	n	S	S	+	+	+	+	x	x	x	n	S	w	w
PSIPRED	S	S	n	w	w	w	n	S	w	n	n	n	n	n	n	n	n	n	S	n
JPRED	S	S	w	S	w	w	n	S	S	w	S	S	w	n	w	n	n	w	S	w

STITCHER: Flexible assembly of likely β-structures

Introduction

While the secondary structure of amyloid is known to be highly β-rich[32-34, 47, 125], structural determination has proven highly difficult, with only extremely short segments crystallized [51, 75] and a very few successful solid-state nuclear magnetic resonance (ssNMR) studies [80-82, 126, 127]. Due to the scarcity of direct evidence, the nature of amyloid and prion supersecondary structures and their relation to sequence have been highly contentious topics [49, 51, 128]. The parallel β-helices form a fold widely cited as one potential model for amyloid [52, 115, 136], while others favor a 'superpleated β-sheet' [146, 155, 156]. Complications include the morphological heterogeneity of amyloid structures suggested by EM imagery [52, 53] and the demonstration of prion 'strains' or 'variants' with differing growth and stability phenotypes [54-56]. In the case of the yeast prion protein Sup35, such variants have been demonstrated to maintain specificity through serial passage [55] and have been correlated with differences in conformation [57].

The bi-stable nature of amyloid prions, as well as the observation of heterogeneity and 'strains' in amyloid and prion folding, undermines the canonical viewpoint of 'one protein, one fold' long held by theorists of protein folding. Instead, a murky view arises of a set of stable valleys in a field of conformational configurations, within which variations are permitted around common or similar folding patterns. Enzymologists have long studied the variations in globular protein conformations caused by ligand binding, catalytic activity, presence of ions or cofactors. Amyloids embody a similar but larger set of variations.

In Chapter 3, the foundations were laid for a broader interpolation of β-protein structure from sequence in the program BETASCAN. Like BETAWRAP, BETASCAN calculated likelihood scores for potential β-strands and strand-pairs based on correlations observed in parallel β-sheets. A key and novel feature of BETASCAN was a maxima-finding algorithm that searched the strands and pairs with the greatest local likelihood for all of the sequence's potential β-structures. BETASCAN suggested multiple alternate folding patterns and assigned relative *a priori* probabilities to these strand-pairings. While sufficient to act as a detector of β-structure propensity in general and amyloidogenic potential in particular, BETASCAN did not filter enough alternate folding patterns to suggest the most likely overall folds of a β-structure. Thus, BETASCAN was insufficient to distinguish difficult cases of amyloid detection, such as the Het-S allele in *Podospora anserina,* and was unable to describe supersecondary structure more complex than strand-pairs.

The STITCHER method described in this chapter extends prediction of amyloid-like proteins by employing a combination of the flexible template, probabilistic prediction [113], and free-energy [112] methods for protein structural prediction. Since amyloid templates cannot be directly derived from known structures, the algorithm creates flexible templates based on an abstraction of the β-helix and superpleated-sheet folds by 'stitching' strand-pairs into β-sheets. Evaluation of the templates is then made by using a combination of free-energy methods and BETASCAN-derived probabilities. The free-energy methods account for the enthalpy of created hydrogen bonds and the entropy of linkers, while the probabilities describing the likelihood of β-sheet formation account for the specific sidechain-sidechain interactions that confer structural specificity. Of particular importance to achieving a sufficiently narrow search space for the flexible amyloid structural templates is the detection of stacking ladders, formed by the sidechain-sidechain stacking and bonding of glutamine, asparagine, tyrosine, and phenylalanine residues [77, 129, 157]. Once a set of high-scoring plausible structures is calculated, a consensus is established by polling for the inclusion of specific strand-pairs among the top *Z* structures. Portions of the structure are considered more likely to form if 80% or greater of the top *Z* structures agree on their inclusion. The STITCHER approach differs from the usual free-energy strategies in that neither an atom-by-atom calculation of structural free-energy nor a ensemble calculation spanning all possible free-energy spaces is attempted. Instead, the strand-pairs, linkers, stacks, and other structural elements are considered as discrete elements to be combined by the dynamic algorithm. This philosophy of establishing and then manipulating pre-defined structural components has been previously used successfully in other contexts [95].

Results

Detection of linear vs. β-helix-like amyloid potential

While amyloid structure is generally agreed to be overwhelmingly composed of β-structure, the applicability of the β-helix or superpleated-sheet models to any one amyloid has been contentious in the absence of readily available crystallization or NMR data. To address this difficulty, STITCHER was tested on amyloid beta, an amyloid with solved NMR structures [79, 80], allowing both superpleated sheet $(m = 1, i = j)$ and β helix $(m > 1)$ structures (see Symbols Used for definitions of all variables). The results are shown in Figure 4-1A.

In the case of amyloid beta, the highest-scoring structures all incorporate at least one $i = j$ strand-pair (the ten-residue β-strand from isoleucine 31 to valine 40 inclusive). Several of the highest-scoring structures include strands analogous to those in solved NMR structures, including strands beginning at tyrosine 10 (corresponding to [79]) and at leucine 17 (corresponding to [80]). However, the highest-scoring structures include a first strand with one residue shifted from a perfect in-register parallel alignment, in the region between histidine 13 and alanine 21 (sequence HHQKLVFFA). This region is known to have varying stability in alignment, especially at differing pH [158].

Het-s: detection of amyloidogenicity

A key ability in amyloid folding studies is the distinguishing of amyloidogenic from non-amyloidogenic sequences. To address this requirement, STITCHER was tested on the small-s and big-S variants of the Het-s mating compatibility factor. The small-s allele of this protein is known to form a prion [47] with multiple, partially conflicting determined solid-state NMR structures [81, 82]. In contrast, the big-S allele does not form an amyloid structure, despite having only three differing residues.

The results for Het-s and Het-S are displayed as Figure 4-2A and Figure 4-2B, respectively. Immediately evident is the greater connectivity of the Het-s predictions. Very few of the predicted Het-S strand-pairs are able to form multiple-sheet structures within the available constraints. On the other hand, Het-s has multiple potential connectivities. The strand-pair with the greatest occurrence, identified by N-terminal residues as isoleucine 14 – threonine 49, corresponds to the strands 1b, 2a, 3b, and 4a in the latest structure [82].

Sup35p: case study in three species

To compare the analysis of STITCHER in amyloids of related function, the sequences of the yeast prion Sup35p from *Candida albicans, Saccharomyces cerevisiae,* and *Yarrowia lipolytica* were analyzed. The results are available in Figure 4-3.

The clearest structural prediction is found in the results for the *C. albicans* sequence (Figure 4-3A), with only one consensus fold predicted for most of the amyloidogenic domain. The signal from each structural strand-pair is fairly weak, with most strands only of length 3 or less. Most of the amyloidogenic potential detected in this sequence comes from its long polyglutamine stretches. These are counted by STITCHER as loop-based stacking bonuses that contribute to structural stability. However, without sequence variation in these regions, they do not confer greater conformational specificity. Instead, tyrosine and phenylalanine ladders serve to align the structure, as previously verified experimentally [77].

The results for Sup35p *S. cerevisiae* (Figure 4-3B) and *Y. lipolytica* (Figure 4-3C) are more heterogeneous in fold prediction. Much of this heterogeneity is due to the recurrent repeats of the Sup35p sequence. For instance, the strand beginning at glycine 44 can favorably be paired with the repeats at glycines 68, 77, and 97. In the midst of this heterogeneity, the importance of consistently predicted strand-pairs is increased. Three pairs in the *S. cerevisiae* structure, identified by their N-terminal residues as asparagine 12 – glutamine 38, tyrosine 29 – tyrosine 49, and glutamine 38 - glutamine 62, are highly recurrent. These pairs correspond to the "head" area identified in previous experimental studies [57], where mutations are known to have large effects on amyloidogenic potential [133]. This contrasts with the increased variance in predictions for the region beyond residue 91, which corresponds to the "tail" region known to have structural heterogeneity associated with "strains" [57]. The *Y. lipolytica* structure (Figure 4-3C) shows similar traits, but with heterogeneity further increased by the irregularity of the sequence repeats.

Alpha-synuclein, Rnq1p results underscore the effects of ladders and single-residue repeats

To explore the utility of STITCHER to new amyloids, analysis was undertaken on human alpha-synuclein, a known amyloidogenic protein, and *S. cerevisiae* Rnq1p, a yeast prion (Figure 4-4). One putative structure has been proposed for alpha-synuclein [144], but confirmatory experimental evidence is yet to be forthcoming. No structure has yet been proposed for Rnq1p, although it is known that Rnq1p fibrils may serve as templates for Sup35p nucleation [159, 160].

Rnq1p structural prediction results (Figure 4-4A) are highly consistent and follow a pattern similar to that seen for *C. albicans* Sup35p. While nonspecific loop-based sidechain ladders of glutamine residues provide the bulk of structural stability, shorter strands including not only glutamine but phenylalanine, tyrosine, and asparagine ladders provide fold specificity. In the region of residues 53-175, the fold pattern is regular, albeit with large exterior loops exposing high glutamine and asparagine content. Thereafter, the increasing occurrences of polyasparagine and polyglutamine stretches in the sequence preclude specific structural alignments.

The results for alpha-synuclein (Figure 4-4B) reveal an overall consistency with local variations. The highest-scoring structures are composed of two sheets, the first with strands of length 3 and the second with strands of lengths 7-10. Interestingly, some strand pairings have several alternate alignments in nearly the same locale and with nearly equal scores. As with amyloid beta, this variability is associated with repeated residues found most often near the edges of the strands (valines 15 and 16, alanines 17-19; alanines 29 and 30; valines 48 and 49, glycines 67 and 68, valines 70 and 71, alanines 88-91, lysines 96 and 97).

Discussion

The STITCHER algorithm addresses the problem of amyloid structural prediction by generalizing the template method utilized by BETAWRAP and its successors. The BETAWRAP algorithm was applicable only to those β-helices which conformed to its rigid template. STITCHER utilizes the free-energy scoring method to 'stitch' a template based on the BETASCAN strand-pair locations and scores, stacking patterns, and linker distances. Within the constraints of the rung and sheet patterns of the superpleated-sheet and β-helix folds, a range of potential strand and loop lengths and locations are dynamically analyzed and compared. Because of both the uncertainty in free-energy parameters and the heterogeneity of amyloid folds, the top fifty fold solutions are presented and polled for strand-pair inclusion. While the highest-scoring fold is frequently correct in many particulars, the best results are achieved by assembling consensus structure(s) from the most frequently occurring strand-pairs among the top solutions. The exploration of the local folding space permits a more robust and complete description of the structures available to the amyloidogenic protein than a single, so-called "optimal" structure. At the same time, the consensus structure is distinct from an "ensemble" structure in that it is composed solely of separately evaluated structures, any of which could be a stable local minimum in conformational space. Therefore, the consensus reflects not an "average" structure, but the degree and location of similarity between the possible structures of the amyloid.

The results for alpha-synuclein, Rnq1p, and to some extent the various Sup35p proteins highlight the roles of single-residue repeats, motif repeats, and sidechain stacking ladders in amyloid structure. Single-residue repeats may contribute to the stability of a structure by forming stacking ladders, especially in the cases of polyglutamine and polyasparagine. However, this stability increase comes at the expense of specific folding in the repeat region, as the importance of aligning any particular pair of residues is reduced. The same effect is seen in a reduced form for repeats composed of multiple-residue motifs, such as in *S. cerevisiae* Sup35p. Specificity of folding for repeat-heavy amyloidogenic proteins appears to be conveyed through two other features of a structure: length minimization of the intervening linker loops, and formation of strand-pairs with rarer residues and stacking ladders.

In the case of alpha-synuclein, the double histidine and double phenylalanine may contribute to the added stability of the shifted strand-pair over the in-register alignment, as well as the avoidance of close proximity for the charged lysine 16. However, the differential in stability would likely be dependent on the ionization state of histidines 13 and 14. This state would also contribute to the heterogeneity of amyloid beta fibrils and complicate structural determination efforts. Given the known sensitivity of amyloid beta folds to pH changes and the presence of variable-ionization histidines, previous structural determination efforts may be biased towards preparations that discourage formation of the shifted strand-pair in favor of the in-register alignment.

In the cases of known amyloid structures, STITCHER is able to extract and present the essential elements of experimentally determined folds. All experimentally based structural models of amyloid beta and Het-s protein are contained in the top predictions in Figures 4-1 and 4-2. Conversely, the results for Het-S protein convey its non-amyloidogenic nature, despite high β-propensity and a sequence only three mutations away from the Het-s prion sequence. While more robust analysis and verification awaits additional amyloid structure determination, the STITCHER methodology appears to be a valuable addition to the growing number of amyloid detection algorithms. It should be noted that the particular fold taken by an amyloid is sensitive to environmental conditions, including pH, concentration of protein, and presence of chaperone proteins. Thus, the structure with the highest STITCHER score may not be that taken by the protein under experimental conditions, as in the case of Het-s.

The low sequence homology, low sequence complexity, and high heterogeneity of amyloid proteins only underscore the adaptable nature of the flexible-template approach. With the identification of further sets of flexible constraints on β-rich folds, the STITCHER methodology should be readily extensible to other varieties of all-β protein structures. While some interpretation and experimental verification is necessary for complete understanding of amyloid folding, the identification of the range of most likely folds should greatly enhance the further computational and experimental investigation of amyloid and prion proteins.

Materials and Methods

Algorithmic strategy

The greatest problem in protein structure prediction is the reduction of possible conformation patterns from a mind-boggling potential space to the few viable and stable conformations. In the present case, the restriction to parallel, all-β structures massively reduces the conformational space and simplifies the prediction problem. The maxima-finding algorithm of BETASCAN reduced the problem further by identifying locally probable strands and strand-pairs. Nonetheless, far more β-strands are deemed 'probable' by BETASCAN than could be assembled into any one fold.

The challenge of 'stitching' an all-β fold from strand-pairs originates in the observation that the most stable structure may not consist of that formed from only the most likely strand-pairs. Once the conformational space is restricted by one strand-pair, the conditional probability of formation for other strand-pairs may be increased. Therefore, even low-probability strand-pairs need to be considered as potential components of a fold. However, the number of possible folds increases exponentially with each additional component. The conformational space must be reduced to a computationally tractable size without excluding components which may form part of near-optimal structures. Two strategies are here employed. First, constraints are placed on the structure space to fit observations and suspected characteristics of the structure. Then, the remaining structure space is subdivided into rungs, sheets, and ultimately units of strand-pairs and linker loops, whose contributions to fold stability may be assessed piecewise. Finally, a dynamic programming algorithm is used to assess the stability of complete structures by assembling the results of strand and rung sub-calculations. Once complete structures are assembled and ranked by stability, the top structures are polled to assess their agreement or disagreement for specific strand-pairs. As more top-scoring structures contain a specific strand-pair or set of strand-pairs, the more likely it becomes that these structural component(s) will be found in experimental studies of the amyloid.

Fold constraints and parameters

The target conformational space of STITCHER is the space of amyloids composed of parallel β-sheets. Following the conventions of previous authors [115-117, 161] and the evidence provided by known amyloid structures [79-82, 155], the amyloid is modeled as a set of n sheets arranged with sequential rungs. Each rung contains n strands, with the strands contributing to the β-sheet. For a structure of m rungs, therefore, there will be mn strands and $mn - 1$ *linkers* connecting strands to each other. Strand-pairs $\{i, j \geq i, L > 1\}$, formed by two strands starting at residues *i* and *j* stacked atop one another in a β-sheet for the next L residues, exist between every adjacent rung. There are therefore $(m-1)n$ of these strand-pairs in each putative structure. A complete amyloid protofibril is believed to be composed of many copies of such structures, each aligned either head-to-tail or head-to-head, tail-to-tail [57].

The regularity of the β-helix motif markedly simplifies the conformational space to be probed. Bounds on the parameters m and n further reduce the number of choices to be made in selecting plausible structures. The two models of amyloids in the literature may be described by constraints on parameters. The β-helix fold requires $m \geq 1, n \geq 2$ [81, 126, 147]; nearly all observed cases in nature, including solved structures of amyloids, are either $n = 2$ or $n = 3$. The superpleated β-sheet model is formed by parameters $m = 1, n \geq 2$. This model can be viewed as a special case of β-helices stacked head-to-tail, where every copy of the amyloid protein forms a single rung of *n* β-strands. In the case of $m = 1$, therefore, every strand-pair consists of two identical copies of a strand, and $i = j$ for all strand-pairs. Therefore, two possible sets of parameters were considered for analysis: $n = \{1,2,3\}, m \geq 1$ and $m = 1, n \geq 1, i = j$.

Finally, the space of potential strand-pair combinations is restricted by the geometry of all-β structures. While an amyloid has a quaternary structure resembling an undulating planar surface, the backbones of $m > 1$ amyloids must loop back to a location near their starting point at the end of each rung. This requires, geometrically, that the distance from the end of the first strand of any strand-pair to the start of the second be longer than the strand itself: $(j - i) > 2L$. Likewise, observations of crystal and NMR structures indicate that the shortest, tightest turn possible in a compact β-helix is three residues; therefore, for $m > 1$, linkers must be at least three residues long.

Free-energy scoring function

The structures which conform to the restrictions outlined above are scored using a formula that includes the BETASCAN scores of their strand-pairs, as well as bonuses reflecting the stabilizing influence of non-backbone hydrogen bonding and the entropic cost of restricting backbone movement into the loops of strands. The problem of weighting the contributions of these various effects to fold stability was solved by summing estimates of their free-energy contributions to stability. Following the pattern of Zhang *et al.* [162], the Gibbs free energy delta G of the change from unfolded to folded state is

$$\Delta G = \Delta E_c + \Delta E_{el} - T\Delta S_{bb}^{prot}$$

where ΔE_c is the contact enthalpy of placing residues together, ΔE_{el} is the electrostatic energy associated with ionic interactions, and $T\Delta S_{bb}^{prot}$ is the entropy of folding. Amyloids are typically poor in charged amino acids, and the stronger partial dipoles associated with other amino acids are usually incorporated into hydrogen bonding. In this analysis, the electrostatic interaction is therefore only considered with reference to hydrogen bonds, and ΔE_{el} is set to zero.

The contact energies can be further decomposed into energies contributed by backbone-backbone, backbone-sidechain, and sidechain-sidechain interactions:

$$\Delta G = \Delta E_{bb-bb} + \Delta E_{bb-sc} + \Delta E_{sc-sc} - T\Delta S_{bb}^{prot}$$

Role of BETASCAN scores

In the case of an all-β structure, the backbone consists of β-strands, with a quasi-linear arrangement constrained by the hydrogen bonds to other strands, and linkers, which are only constrained by the strands at their beginning and end. There is one hydrogen bond per residue in the length of each strand-pair, and an equivalent entropy loss for the constraint it imposes on backbone flexibility. Therefore we separate the entropy into linker and strand terms, and rearrange to express them with reference to length:

$$\Delta G = (\Delta E_{bb} - T\Delta S_{str.}) + \Delta E_{bb-sc} + \Delta E_{sc-sc} - T\Delta S_{link}$$

$$\Delta G = (\Delta E_{H-bond} - T\Delta S_{bb-res.})L_{str.} + \Delta E_{bb-sc} + \Delta E_{sc-sc} - T\Delta S_{link}$$

Side-chain/backbone interactions are a primary determinant of β-sheet propensities, both through van der Waals interactions [163] and entropy of salvation [164]. These factors explain much of the known relative affinities of amino acids [165], although these propensities must be interpreted in context [166]. The BETASCAN algorithm uses these propensities to estimate relative probabilities of formation of a strand-pair, with normalization for length to allow comparison of strands with different lengths. These relative probabilities are related to the energy of formation of the β-strand, according to the Boltzmann relationship

$$\Delta G_{\beta-form.} = k\ln\left(\frac{P_{\beta}}{P}\right)$$

where P_{β}/P is the ratio of potential β-containing conformations to all potential conformations. Since the BETASCAN score is a log-odds ratio estimating the relative probability of the β-containing conformation, it can be surmised as proportional to this energy. The energies from the direct hydrogen bonds made by all residues in the strand may be combined with the side-chain effects and estimated by the BETASCAN score for the entire strand:

$$\Delta G = \sum_{strands} [(\Delta E_{H-bond} + \Delta E_{bb-res} - T\Delta S_{bb-res.})L_{str.}] + \Delta E_{sc-sc} - T\Delta S_{link}$$

$$\Delta G = \sum_{strands} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\, res.}}\right) L_{str.}\right] + \Delta E_{sc-sc} - T\Delta S_{link}$$

Side-chain stability bonuses

Side-chain/side-chain energies include hydrogen bonding in the case of asparagine and glutamine stacking [129], pi-bond orbital stacking in the case of tyrosine and phenylalanine [167], and van der Waals interactions between side-chains. The first two contribute bonus stability beyond that calculated by BETASCAN. The last has been shown to be very small (less than 0.20 kcal/mol [162]) and can be disregarded. We therefore set:

$$\Delta E_{sc-sc} = \sum_{X=\{Q,N,Y,F\}} (n_{XX})\Delta E_{XX-stack}$$

Entropic penalties

Finally, we consider the entropy of the linkers, those free loops of peptide constrained at either end by attachment to β-strands. The problem of calculating this entropy is a subset of the general problem of polymer condensation entropy [168] and bears remarkable similarity to that of disulfide bond entropy [169]. The entropy may be calculated as

$$\Delta S = -R \ln\left(\frac{3}{\left(2\pi {l_{aa}}^{2} L_{link}\right)^{3/2}}\right) v_{ends}$$

where R is the gas constant, l_{aa} the length from α-carbon to α-carbon, 3.8 angstroms, L_{link} is the number of residues in the linker, and v_{ends} is the volume the ends of the linker may occupy. A hydrogen bond is approximately the same length as the distance between sulfide groups in a disulfide bond, namely 4.8 angstroms. The entropy calculation, using these values, may be simplified to

$$T\Delta S = -2.1 \frac{\text{kcal}}{\text{mol} - \text{res}} - \frac{3}{2} R \ln L_{link}$$

for $T \approx 300\ \text{K}$. Because we are comparing structures each known to have linkers, we disreagard the constant term and estimate simplify further to yield the relative entropy of a linker,

$$T\Delta\Delta S = (0.9 \frac{\text{kcal}}{\text{mol} - \text{res}}) \ln L_{link}$$

This formula is used in two different ways: to calculate the entropic penalty of adding a rung, and to calculate the entropic penalties accruing to forming strands inside the rung. The difference is in the value of L_{link}. For a rung penalty, we assess the entropy of forming a loop from a free polypeptide chain without regard to strand-pairs (as the strand-pairs cannot form until the chain is in proximity to itself). In this case, $L_{link}^{rung} = j_1 - i_1$, the difference between the N-terminii of the two strands in the first strand-pair of the rung. For linkers within the rung connecting two strand-pairs *f* and *g*, the length of the linker is the number of residues between the strand-pairs, counting both the upper and lower chains of the pair:

$$L_{link}^{f-g} = i_g - (i_f + L_f) + j_g - (j_f + L_f)$$

The form of the scoring function for STITCHER may now be fully described as

$$\Delta G = \sum_{strands} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\, res.}}\right) L_{str.} \right] + \sum_{X=\{Q,N,Y,F\}} (n_{XX}) \Delta E_{XX-stack}$$

$$- \sum_{rungs} \left(0.9 \frac{\text{kcal}}{\text{mol} - \text{res}}\right) \ln L_{link}^{rung} - \sum_{links} \left(0.9 \frac{\text{kcal}}{\text{mol} - \text{res}}\right) \ln L_{link}^{f-g}$$

Energy weights

To make this function calculable, we must estimate the weights $\overline{\Delta G_{\beta\, form.}^{per\, res.}}$ and $\Delta E_{XX-stack}$. Experimental data [162-164, 166] suggests the free energy of β-strand formation per residue to be approximately 1 kcal/mol-res, a combination of the enthalpy of the hydrogen bond and the entropy of salvation as influenced by side-chains. This is a somewhat rough estimate due to context-dependency [166]. For the bonuses and penalties, we assess the contribution of additional hydrogen bonds to the free energy. The free energy of the hydrogen bond is again offset to some degree by salvation, though not as strongly as for the backbone. The rough bonus weights $\Delta E_{QQ-stack} \approx 1 \frac{\text{kcal}}{\text{mol-res}}$, $\Delta E_{NN-stack} \approx 2 \frac{\text{kcal}}{\text{mol-res}}$, $\Delta E_{YY-stack} \approx 1 \frac{\text{kcal}}{\text{mol-res}}$ were used for this study. The extra weighting of *NN* over *QQ* is justified in two ways. First, asparagine (*N*) has a shorter distance from backbone to amide than does glutamine *(Q)*. Additionally, experimental data [157] demonstrate that in at least one prion, replacing all glutamines with asparagines provides double the stability of the prion fold as compared to replacing all asparagines with glutamines. These estimated energy weightings may become more accurate as calorimetry of sidechain-sidechain interactions becomes available.

Evaluation of score

The STITCHER algorithm uses a dynamic programming algorithm to evaluate estimated ΔG for the combinations of strand-pairs that can be combined into templates matching the parameters described above. To do so, the calculation of ΔG is subdivided into calculation by rungs. The total free energy change can then be calculated by summing the stability contribution of any rung r containing strands $\{r_1 \dots r_n\}$ and linkers $\{r_{link}^{1-2} \dots r_{link}^{(n-1)-n}\}$, with a linker r_{link}^{0-1} to the previous rung, as

$$\Delta G = \sum_{r=\{1 \dots m\}} [\Delta G_r]$$

$$\Delta G_r = \sum_{str=\{r_1 \dots r_n\}} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\ res.}} \right) L_{str} \right] + \sum_{X=\{Q,N,Y,F\}} (n_{XX}^{r}) \Delta E_{XX-stack}$$
$$- \sum_{g=\{1 \dots n\}} \left(0.9 \frac{\text{kcal}}{\text{mol} - \text{res}} \right) \ln\left(L_{link}^{(g-1)-g} \right)$$

If the stacking bonuses, a directly sequence-dependent calculation, are considered apart from the strand and linker scores, which are only indirectly sequence-dependent, then the rung calculation can be partially separated into strand calculations:

$$\Delta G_r = \sum_{g=\{1\ldots n\}} \left[Score\left(\overline{\Delta G_{\beta\, form.}^{per\ res.}}\right) L_{r_g} - \left(0.9 \frac{\mathrm{kcal}}{\mathrm{mol-res}}\right) \ln\left(L_{link}^{(g-1)-g}\right) \right] + \sum_{X=\{Q,N,Y,F\}} (n_{XX}^{r}) \Delta E_{XX-stack}$$

By calculating the free-energy scores of strand-pairs and linkers as subproblems of rung scoring, and rung scores as subproblems of structure assembly, the dynamic programming method can be used to iteratively calculate the *Z* structures with the highest score by tracing back through internally consistent partial structures that do not violate the defined fold constraints.

Consensus evaluation and output

The consensus of the *Z* highest structures (with default *Z* of 50) is assessed by scanning over all structures for included strand-pairs by the locations (i, j) of their termini. If the number of strand-pairs in the *Z* highest structures with N-termini of $(i \pm 2, j \pm 2)$ total more than 80% of Z, the location is noted as a consensus structure element, and the ranges of i, j, L, and strand-pair score over the strand-pairs in the $(i \pm 2, j \pm 2)$ region are output. For display purposes, the strand-pairs with matching lower and upper strands are aligned vertically to reconstruct predicted β-sheets. The output of STITCHER includes the set of *Z* predicted structures, a diagram of local structure space, and the top-scoring and consensus structures.

Figures

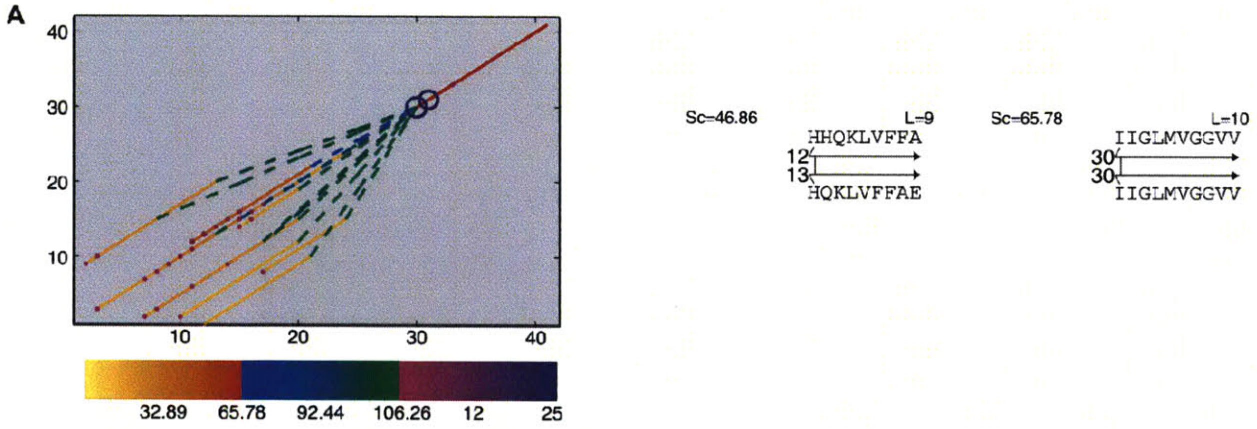

A
40
30
20
10
10
20
30
40
32.89
65.78
92.44
106.26
12
25
Sc=46.86
L=9
HHQKLVFFA
12
13
HQKLVFFAE
Sc=65.78
L=10
IIGLMVGGVV
30
30
IIGLMVGGVV

Figure 4-1, STITCHER results for, *A,* amyloid beta. At left, a contact map of the fifty highest-scoring folds. The horizontal and vertical axes indicate, respectively, the residue numbers (counted from the N-terminus) of the lower and upper strands in each strand-pair of the structures. Starting locations of strand-pairs are indicated by circles, with size of circle (small to large) and color of circle (magenta to black, right-hand color spectrum) indicating the strength of the consensus vote of the top fifty structures for that strand-pair. The strand-pairs are drawn along their length in shades of orange, with stability increasing from yellow to red (left-hand color spectrum). Fold structures are indicated by the dotted lines connecting strand-pairs into rungs and sheets. Structure scores are indicated by shades with stability increasing from blue to green (center color spectrum).

At right, the highest-scoring fold. Each strand-pair is denoted by its score (Sc) and its length (L). To the extent possible, rung-pairs proceed from left to right and sheets from top to bottom. Numbers to the left of the strands indicate the number of the residue immediately preceding the N-terminus of the strand. Slanted lines indicate the first residue of the strand, arrowheads the last residue of the strand, and connecting line(s) indicate the possible residue-pairing(s) of the first residues of the strands.

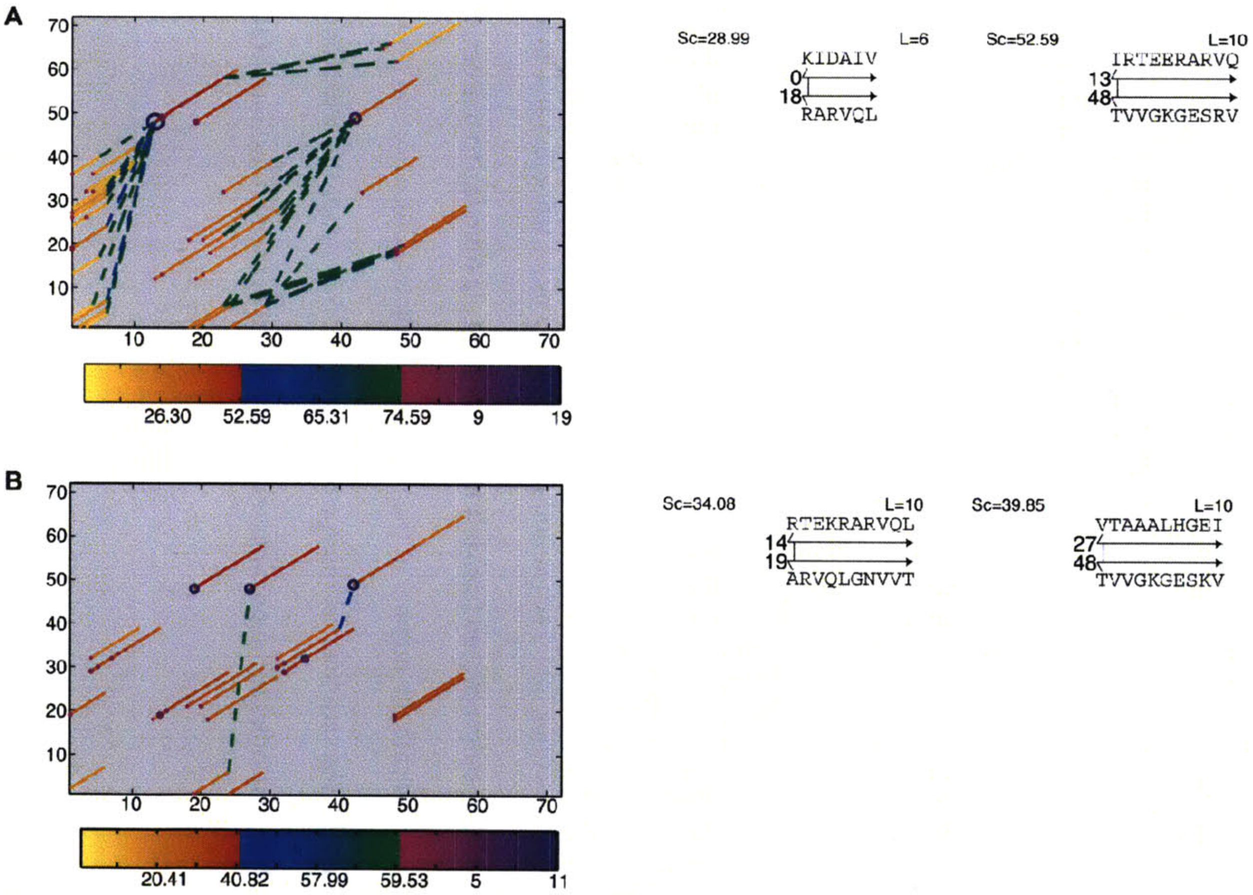
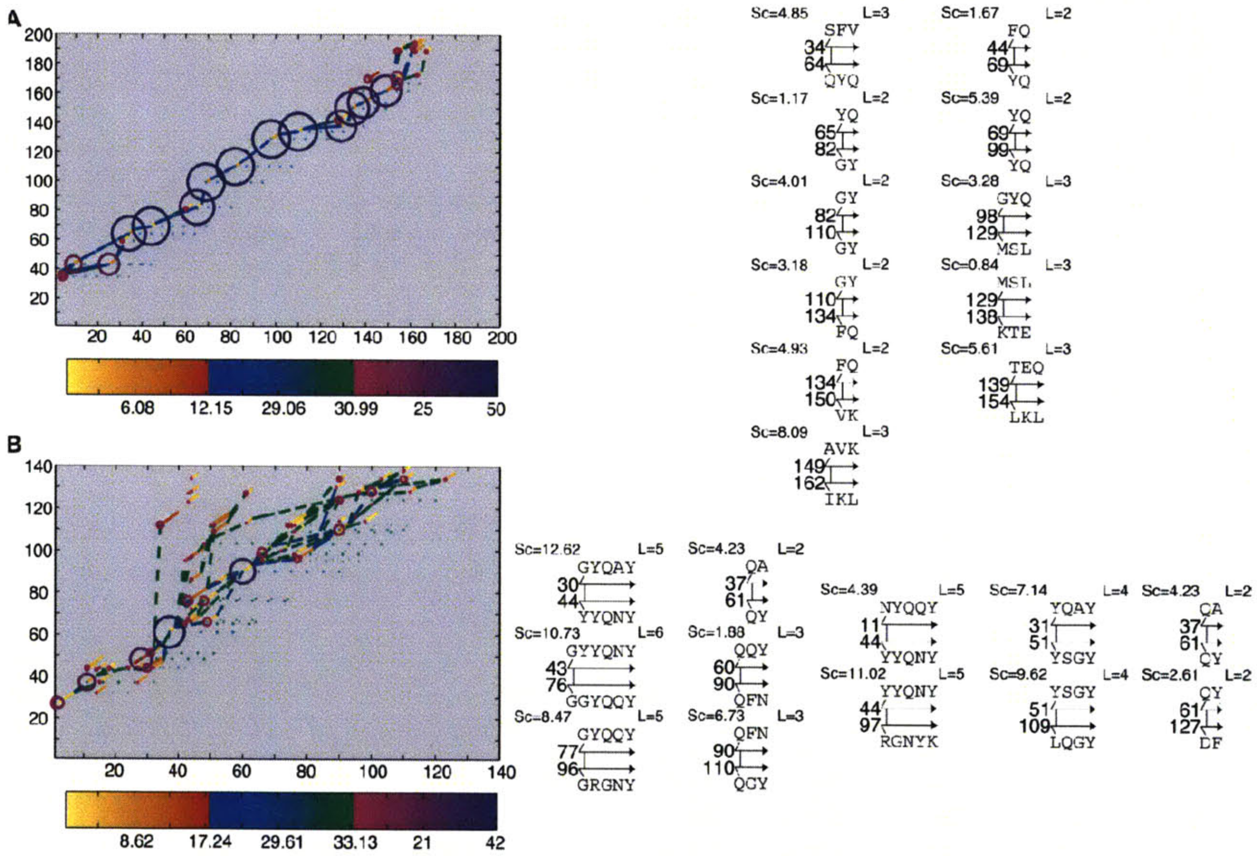

A
70
60
50
40
30
20
10
10
20
30
40
50
60
70
26.30
52.59
65.31
74.59
9
19
Sc=28.99
L=6
KIDAIV
0
18
RARVQL
Sc=52.59
L=10
IRTEERARVQ
13
48
TVVGKGESRV
B
70
60
50
40
30
20
10
10
20
30
40
50
60
70
20.41
40.82
57.99
59.53
5
11
Sc=34.08
L=10
RTEKRARVQL
14
19
ARVQLGNVVT
Sc=39.85
L=10
VTAAALHGEI
27
48
TVVGKGESKV

Figure 4-2, STITCHER results for the two alleles of the *Podospora anserina* mating compatability protein: *A,* Het-s; *B,* Het-S. Contact maps, at left, and top-scoring structures, at right, are as described in the caption to Figure 4-1.


A
200
180
160
140
120
100
80
60
40
20
20
40
60
80
100
120
140
160
180
200
6.08
12.15
29.06
30.99
25
50
Sc=4.85 L=3
SFV
34
64
QYQ
Sc=1.67 L=2
FQ
44
69
YQ
Sc=1.17 L=2
YQ
65
82
GY
Sc=5.39 L=2
YQ
69
99
YQ
Sc=4.01 L=2
GY
82
110
GY
Sc=3.28 L=3
GYQ
98
129
MSL
Sc=3.18 L=2
GY
110
134
FQ
Sc=0.84 L=3
MSL
129
138
KTE
Sc=4.93 L=2
FQ
134
150
VK
Sc=5.61 L=3
TEQ
139
154
LKL
Sc=8.09 L=3
AVK
149
162
IKL
B
140
120
100
80
60
40
20
20
40
60
80
100
120
140
8.62
17.24
29.61
33.13
21
42
Sc=12.62 L=5
GYQAY
30
44
YYQNY
Sc=4.23 L=2
QA
37
61
QY
Sc=10.73 L=6
GYYQNY
43
76
GGYQQY
Sc=1.88 L=3
QQY
60
90
QFN
Sc=8.47 L=5
GYQQY
77
96
GRGNY
Sc=6.73 L=3
QFN
90
110
QCY
Sc=4.39 L=5
NYQQY
11
44
YYQNY
Sc=11.02 L=5
YYQNY
44
97
RGNYK
Sc=7.14 L=4
YQAY
31
51
YSGY
Sc=9.62 L=4
YSGY
51
109
LQGY
Sc=4.23 L=2
QA
37
61
QY
Sc=2.61 L=2
QY
61
127
DF

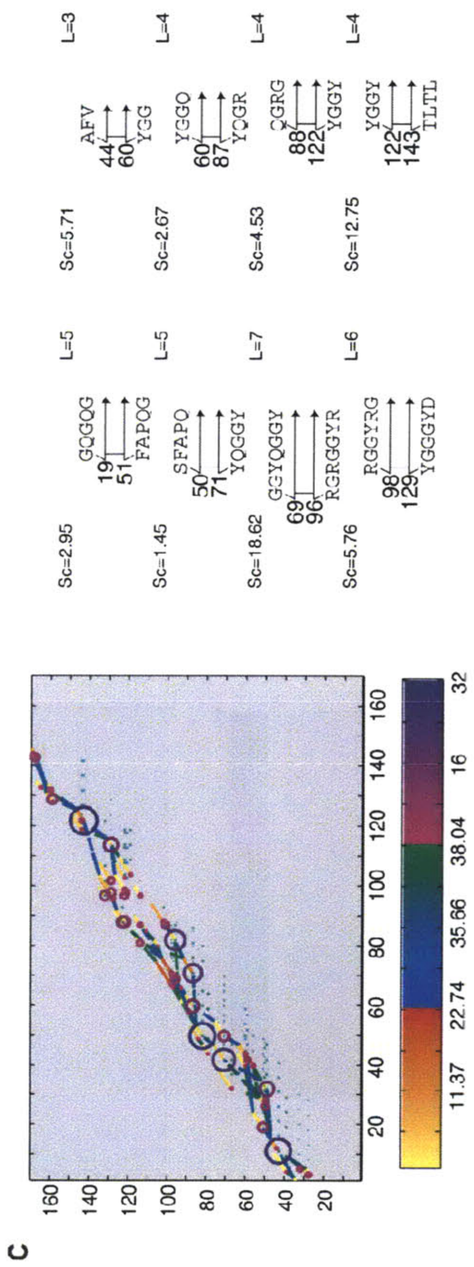


Figure 4-3, STITCHER results for Sup35p in three species. *A, C. albicans*, *B, S. cerevisiae*, and *C, Y. lipolytica*. At left, a contact map of the fifty highest-scoring folds. At right, the top-scoring structure. For *S. cerevisiae*, the highest- scoring structures for $n = 2$ and $n = 3$ are presented. For description of colors, numbers, and lines, see the caption to Figure 4-1.

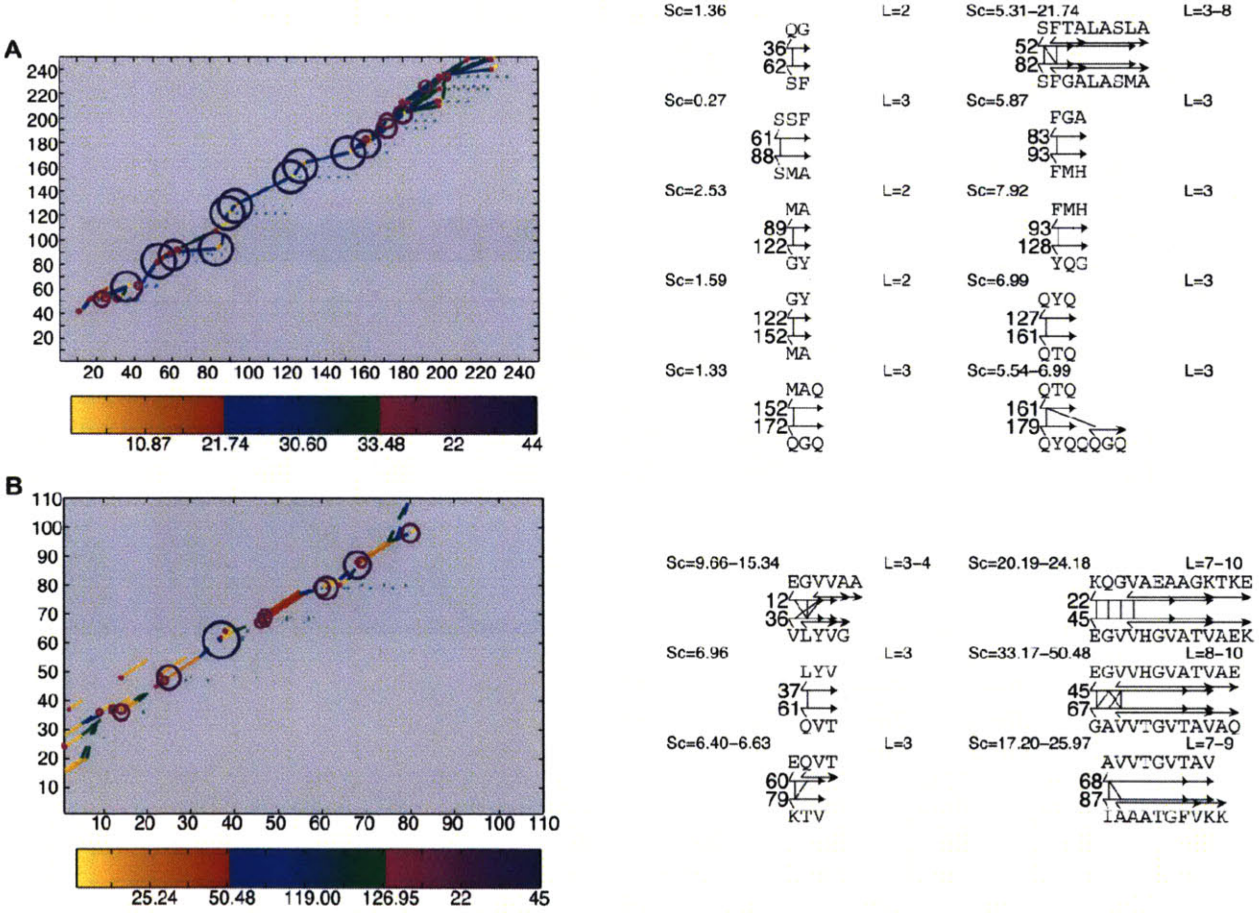

A
240
220
200
180
160
140
120
100
80
60
40
20
20 40 60 80 100 120 140 160 180 200 220 240
10.87 21.74 30.60 33.48 22 44
Sc=1.36 L=2
QG
36
62
SF
Sc=5.31–21.74 L=3–8
SFTALASLA
52
82
SFGALASMA
Sc=0.27 L=3
SSF
61
88
SMA
Sc=5.87 L=3
FGA
83
93
FMH
Sc=2.53 L=2
MA
89
122
GY
Sc=7.92 L=3
FMH
93
128
YQG
Sc=1.59 L=2
GY
122
152
MA
Sc=6.99 L=3
QYQ
127
161
QTQ
Sc=1.33 L=3
MAQ
152
172
QGQ
Sc=5.54–6.99 L=3
QTQ
161
179
QYQGQGQ
B
110
100
90
80
70
60
50
40
30
20
10
10 20 30 40 50 60 70 80 90 100 110
25.24 50.48 119.00 126.95 22 45
Sc=9.66–15.34 L=3–4
EGVVAA
12
36
VLYVG
Sc=20.19–24.18 L=7–10
KQGVAEAAGKTKE
22
45
EGVVHGVATVAEK
Sc=6.96 L=3
LYV
37
61
QVT
Sc=33.17–50.48 L=8–10
EGVVHGVATVAE
45
67
GAVVTGVTAVAQ
Sc=6.40–6.63 L=3
EQVT
60
79
KTV
Sc=17.20–25.97 L=7–9
AVVTGVTAV
68
87
IAAATGFVKK

Figure 4-4, STITCHER results for, *A, S. cerevisiae* Rnq1p, and *B,* α-synuclein. At left, a contact map of the fifty highest-scoring folds. At right, a consensus structure assembled from all clusters of strand-pairs $\{(i_o - 2) \geq i \geq (i_o + 2), (j_o - 2) \geq j \geq (j_o + 2)\}$ found in $> 80\%$ of the top fifty highest-scoring folds. The range of strand positions and lengths is indicated by the shortest and longest possible strand arrows, drawn at the most N- and C-terminal possible locations. Connecting lines indicate the possible pairings of the first residues of the strands making up the strand-pair. The range of lengths and scores for a cluster of possible strand-pairs is indicated at L and Sc, respectively, for each cluster. The numbers to the left of the strands indicate the residue immediately before the leftmost possible residue in each strand (i.e., the residue before the first written above and below the strand arrows). For description of colors and numbers, see the caption to Figure 4-1.

HELIXCAP: Inhibition of amyloidogenesis

Introduction

Parallel β-helices (Figure 5-1) are a subdivision of the protein all-β folds, as defined by the Structural Classification Of Proteins (SCOP) database [3, 170] .The defining attribute of the β-helix structure is the regular nature of its *rungs*, which consist of β-strands arranged in a sequential and repeating order. The combination of repetitive order and a helical conformation creates two or more parallel β-sheets around a central core. Although sometimes interrupted by loops containing other elements of secondary structure, the β-helix as a domain is structurally and geometrically regular. This regularity is maintained despite great disparity in primary sequence [116, 171, 172].

The plant virulence factor pectate lyase C was the first parallel β-helix fold to be identified [173]. Identifications of similar proteins followed [174, 175] and led Yoder and Jurnak to describe a family of pectate lyase-like folds [170]. In the pectate lyase family of β-helices, characteristic right-handed turns and three-stranded rungs create three β-sheets in the complete fold. Further families with other folds were identified. UDP-N-acetylglucosamine acyltransferase typifies left-handed three-stranded β-helices [171, 176]. Likewise, *Pseduomonas aeruginosa* alkaline protease [177] is an example of a two-stranded right-handed β-helix.

Richardson and Richardson [178] noted that β-helices typically have a loop at either end of the structure that serves as a *β helix cap*. This cap is amphipathic, with one side exposed to solvent and the other to the hydrophobic core. The cap thus protects the hydrophobic core of the β-helix from solvent exposure. Richardson and Richardson, among others [179], have speculated that the cap also acts to prevent aggregation in β-helix proteins. These studies point to evidence that many agglutinative proteins, including many prion and amyloid proteins, are repeating and indefinitely extendable β-sheets. Without an element that interrupts β-β contacts at the ends of β-helices, multiple strands of β-helices could dock and form multimeric fibers.

In this chapter, we present an in-depth study on caps in β-helices. Despite the interest in β-helices as potential models of prions and aggregative proteins [77, 180], no in-depth study of β-helix caps has been made. The available structures in the Protein Data Bank (PDB) [2] can be classified by β-helix cap fold to reveal the α-helix cap and 'visor' cap families. Next, we show that these cap families are cross-correlated with the established β-helix families. Finally, a more in-depth study of the sequences of α-helix caps is also made, culminating in a predictive HMM tool that can identify α-helix family caps for β-helices.

Results

Types of caps

In reviewing the known PDB structures of β-helix proteins, several recurrent cap motifs became evident. These motifs are here denoted as the α-helix cap, subdivided into left- and right-handed motifs, and the 'visor' cap, subdivided into the previous-strand and cross-helix motifs. Together, these motifs comprise 83% of the caps in the non-redundant sample of structures described in Table 5-1. In addition to these recurrent motifs, 9 structures use unique, structurally unrelated, and motif-less loops as their caps. One structure, the tailspike of the *Salmonella* phage P22 [174], has a cap domain in lieu of a cap motif. Composed of strands from three identical monomers, a second, interleaved β-helix domain abuts the C-terminal end of the monomeric β-helix domain. The interleaved helix serves as a cap for all three monomeric β-helices.

There is a distinct bias in cap location. N-terminal caps are most often found in the data set to be α-helix caps, and vice versa. Similarly, C-terminal caps are most often found be to 'visor' caps, and vice versa. However, counterexamples exist for both major types of caps. 1THJ, a carbonic anhydrase [181], bears a C-terminal right-handed α-helix cap. Likewise, 1HF2, the MINC protein [182], bears an N-terminal 'visor' cap.

α-helix cap

Structural alignments of α-helix caps are depicted in Figure 5-2. An α helix cap consists of an α helix, a cap loop, and an additional element of secondary structure. The additional element may be a β-strand, as with 2PEC [120] and 1KK6 [183], a loop, as with 1CZF [184], or an α helix, as with 1G95 [185]. The cap helix and additional element form an approximate plane, perpendicular to the long axis of the β-helix (see Figure 5-2B and 5-2D). The cap helix is slightly offset from the long axis and interfaces with the hydrophobic residues in the first rung of the β-helix. Observed examples of α-helix caps (see Table 5-1) have α-helices of 5 to 21 residues in length. The additional element and cap helix lie on the opposite sides of the long axis. However, the additional element is antiparallel to the cap helix and more distant from the long axis.

Typically, the cap elements are the only secondary structures in the immediate vicinity of the N-terminal region of the β-helix. The polypeptide chain immediately N-terminal to the cap either interacts with one face of the β-helix or interacts directly with the cap elements. There is some evidence [186] that elements distal to the cap may act to immobilize cap elements and hold them in place.

The contacts of the cap elements with the exterior side chains of the first rung vary according to the type of helix. In right-handed helices, the cap helix is often both parallel and proximate to the B2 strand of the first rung (see labels, Figure 5-1). For these proteins, interactions are possible between cap helix residues and B2 exterior side chains. The additional element, which is not an α helix in observed right-handed helices, interacts with the B1 and B3 strands of the first rung. In left-handed helices, the cap helix often lies perpendicular to one of the three β strands. For these structures, only one or two cap helix residues are in sufficient proximity to permit contact with exterior side chains. The additional element interacts with a second β strand.

α-helix caps are found in both right-handed (Figure 5-2A) and left-handed (Figure 5-2B) versions, irrespective of the parity of the helix it caps. The population of left-handed α-helix caps is both larger and more diverse than that of right-handed α-helix caps (3 vs. 14 families). In addition, the left-handed α-helix caps have less structure or sequence conservation in the cap loop than right-handed α-helix caps.

'Visor' caps

'Visor' caps are a set of motifs usually, though not exclusively, found at the C-terminal ends of β-helices. A representative assortment of 'visor' caps is shown in Figure 5-3. 'Visor' caps vary widely in their topology but share a common feature. In each 'visor' cap motif, the cap loop crosses and interacts with one of the three β-strands. The motifs may be classified according to the location of the crossing as previous-strand or cross-helix motifs.

Unlike α-helix caps, 'visor' caps may not completely prevent exposure of the core of the β-helix to solvent. The previous-strand caps of left-handed β-helices (Figure 5-3A) only interact with two of the four core hydrophobic residues in the last rung.

The most common 'visor' cap motif is a *previous-strand* motif, where the cap loop is directly connected by a turn to the last strand in the β-helix. The turn element of a previous-stand motif is located below a β-strand, thus making the last strand shorter than those preceding it. Depending on the angle between the last strand and the cap loop at the crossing, this turn is between 180 and 270 degrees. The cap loop crosses under and interacts with the previous strand. The interaction may involve many amino acids and resemble a hairpin loop, such as 1KQA [187] (Figure 5-3A). Contrariwise, the interaction may involve only one or two pairs of amino acids, such as 1IDK [188] (Figure 5-3B).

The opposite motif is a *cross-helix* motif, in which the cap loop crosses one of the other two strands of the helix. Cross-helix motifs are found in both left-handed β-helices (Figure 5-3C, 1G95 [185]), and right-handed helices (Figure 5-3D, 1DBG [189]). In contrast to previous-strand motifs, cross-helix motifs tend to originate below turns in the β-helix. The exception is 1HF2 [182], which originates after a shortened strand. The angle formed by the turn between the last strand and the cap loop is less than 180 degrees. The crossing of a cross-helix motif is close to 90 degrees and creates interactions between only one or two pairs of amino acids.

A few 'visor' cap motifs, such as that for pectate lyase C (Figure 5-3E, 2PEC [120]), include an α-helix. These motifs are distinguished from the α-helix motifs by the relative location of the helix. In a 'visor + α' motif, the α-helix interacts with one strand of the β-helix and not with the hydrophobic core. The α-helix is located between the last strand of the β-helix and the cap loop. Because of the distance thereby introduced between the last strand and the cap loop, all 'visor + α' motifs are cross-helix motifs.

Common sequence features of α-helix cap motifs

Figure 5-4 depicts a composite sequence alignment of the α-helix caps based on structural considerations (see Methods). The α-left (pectate lyase), α-right (pectate lyase), and α-right (left-handed β-helix) helix caps were aligned as separate groups. Each alignment displays the structural elements noted by visual observation and structural alignment: an additional element and a consensus α-helix immediately adjacent to the β-helix. The α-helices are broadly amphipathic, with a hydrophobic side facing the β-helix and a polar side facing solvent.

In contrast, 'visor' caps do not display significant similarities in sequence. These caps' sequences vary in length and sequence composition. There is little or no correlation between structural alignment and sequence composition. Instead, some 'visor' cap families share more distant structural elements. For instance, almost half of the members of the pectate lyase superfamily have a secondary structure element after the C-terminal 'visor' cap, aligned parallel to the axis of the β-helix and located beside the B2 β-sheet of the helix.

Hidden Markov model of α-helix cap sequences

Our initial data set contained 44 β-helix proteins, each representing one SCOP family of β-helix structures. This set was divided into classes by N-terminal cap structure, as follows: α-left cap, 4 structures; α-right from pectate lyase superfamily, 15 structures, α-right from left-handed superfamily, 7 structures; double-α, 2 structures; visor, 5 structures; loop cap, 4 structures; and cap missing from structure, 6 structures (see Table 5-1). Of these, the α-left and α-right structures were sufficiently similar to allow structure and sequence alignment (Figure 5-4); the composite structure and sequence alignments therefore comprised 26 proteins.

The sequence and structure patterns of these α-helix caps were used to create a hidden Markov model (HMM) [190]. A logo of the HMM, as generated by Logomat-M [191], is shown in Figure 5-5A. The most prominent features of the model are the high incidence of hydroxyl residues around the α-helix (residues 7, 9, 11, and 12) and the tendency towards small hydrophobic residues in the β-strand (residues 15-20).

Predictive model of α-helix cap structure

In order to produce a model with greater statistical support, BLAST [192] was used to search the GenBank [193] protein database for sequence homologues of the 26 α-left and α-right structures, resulting in an expanded database of 1057 sequences. These sequences were aligned by *hmmalign* to the HMM described above and used to generate a second HMM, depicted in logo format in Figure 5-5B. This second model displays a complex pattern with signal distributed throughout, except for low contributions from the turn positions 3-5, 21, and 23. Compared to the first model, there are stronger signals in positions 7 and 9 for serine and threonine residues. In addition, the β-strand (residues 13-17) displays a slight but continuous preference for alanines. In positions 18 and 19, where the β-strand crosses the first rung below, the preference switches to bulky hydrophobics (isoleucine, leucine, proline).

Predictive capabilities of α-cap model

In order to validate our HMM-based model of the α helix cap, several target sets were analyzed. First, as a negative control, the October 12, 2007 copy of the sequences of structures in the Protein Data Bank (PDB) [2] with all β-helices removed (the "PDB^{-}" data set) was analyzed. None of the 18,659 sequences in this set resulted in an *hmmsearch* score above threshold. As a positive test and a demonstration of model robustness, leave-one-out cross-validation was performed on sequence-similarity clusters of the 1083 source sequences. To ensure that performance was not due to clustering parameters, validation was performed on clusters with 25% and with 75% sequence similarity. At 25% cluster similarity, the model detected 757 caps (70%), while at 75% similarity, 943 caps (87%) were detected. Finally, to guard against the possibility of overtraining, a model generated without 24 initial sequences was tested on them; 22 of the 24 sequences were detected (93%).

To evaluate the predictive performance of the HMM on a plausible target set, a set of amyloidogenic sequences not known to have helix caps was assembled (Table 5-2) and analyzed with *hmmsearch*. Of these sequences, only the huntingtin analogue of *Xenopus tropicalis* residues 1097-1109 scored above threshold. However, as only a 12-residue portion of the model was matched in a region not suspected of amyloidogenicity (i.e., the polyglutamine sequence found in the first 100-200 residues of known huntingtin sequences), this result may be a spurious hit.

Discussion

The identification of helix caps as a key element of the β-helix fold advances our understanding of the mechanisms used to inhibit and/or control aggregation, and particularly amyloid formation. As has been previously noted by other authors, the wraps of the β-helices are quite similar to the resolved and theorized structures of many amyloid protofibrils. Two major forces act to bring together the monomer β-strands of amyloids: the hydrophobic effect and the hydrogen-bonding patterns of β-sheets. Secondary effects that have been observed stabilizing amyloids, such as tight side-chain packing, side-chain to side-chain hydrogen bonding and packing interactions, and side-chain to backbone hydrogen bonding, require the alignment of the β-sheets to form. The unifying structural functions of the helix caps appear to be to block the hydrophobic effect by shielding the β-helix core from solvent, and to preclude extension of one or more β-sheets via a stably folded physical obstruction.

Because of the broad nature of these structural functions, evolution appears to have selected for a range of different solutions to the helix-capping problem. The diversity of 'visor' cap shapes, as well as unique non-secondary-structure caps and the interleaved β-strands of P22 tailspike, illustrate that no one supersecondary structure or motif must necessarily lie at the ends of a β-helix domain. It was therefore something of a surprise that a significant number of β-helices, including but not limited to the pectate lyase superfamily, do indeed have a loosely conserved motif of secondary structure elements. These β-helices are low in sequence homology. The multiple α-left and α-right folds, as well as the double-α motif seen in the left-handed β-helix superfamily, suggest the possibility of either the loose conservation of an ancient motif or convergent evolution in the design of β-helix caps.

The HMM-based HELIXCAP detector for α-helix-containing β-helix folds presented here should prove useful in future studies of β-helices. Beyond its immediate role of detecting α-helix cap motifs similar to those noted here, we suggest its use as a scanner of genomic data to identify hitherto unidentified β-helices. Because of the low sequence homology of β-helices, their detection from sequence data has often proven difficult. Efforts such as BETASCANPRO [117] have been only partially successful, depending on particular features of the wrap design. While the HELIXCAP detector is also limited, identifying only caps with the α-motif, the range of structures detected is different than that of BETASCANPRO and may be useful in identifying targets for or in conjunction with that or other β-helix detectors.

Finally, the α-helix motif discovered here suggests a possible control mechanism for amyloids. If a cap motif is proximate in sequence to the amyloidogenic domain of a protein, it would serve to forestall the polymerization of its neighboring sequence during initial folding. If the cap is then moved by conformational change, unfolded, or removed by proteolytic action, it would be unable to further prevent amyloid propagation. This hypothesis was tested by scanning a set of known and well-characterized amyloids, some known to contain α-helices, with the HMM-based detector. While no evidence for this mechanism was found, it remains an intriguing possibility either to be found naturally or to be engineered in future amyloidogenic protein constructs.

Methods

Structural and sequence alignments

β-helix families and superfamilies were determined according to SCOP. One PDB structure per SCOP family was downloaded and used for analysis; the latest structure without ligands, heavy atoms, or other molecules incorporated into the crystal was used as representative. The structures were grouped by SCOP superfamily. The full list of structures used is found in Table 5-1 [120, 170, 171, 174, 176, 177, 181-185, 187-189, 194-212].

Structural alignments were generated using DALI [213]. All generated alignments were of secondary structures in the β-helix cap and turns connecting these secondary structures. Alignments were first made in a pairwise fashion to a template structure: 1DBG (α-left helix caps), 1GQ8 (α-right helix caps, pectate lyase superfamily), 1G95 (α-right helix caps, left-handed superfamily), 1JTA (previous-strand visor), 1QCX (cross-strand visor), 2PEC (visor + α). The rotation and translation matrices derived from the DALI alignment were applied to the original PDB files. The rotated PDB structures were combined to produce a general alignment. The orientations of the secondary structure elements, as well as the long axes of the β-helices, were examined to confirm the accuracy of the alignments. Images of the caps were generated using iMol 0.3 [214]. Sequence alignments for each superfamily were derived from the DALI alignments. The general sequence alignment of α caps was arranged to optimize correspondence of secondary structure elements as determined by PSIPRED [150] and turns as determined by inspection of DALI alignments.

Hidden Markov model generation and testing

HMMer [190] was used to compile and calibrate hidden Markov models (HMMs). An initial seed model was generated from the sequence alignment of α caps using *hmmbuild.* Because of the small number of sequences in the alignment, this model was considered to be insufficient for statistical validation. Therefore, a larger model was constructed as follows. A database of β-helix sequences was generated by using BLAST [192] to search the NIH Entrez nonredundant database for matches to sequences in the alignment of Figure 4, with a minimum E-value of $1.0\text{x}10^{-60}$. 1084 sequences were included in the database. *hmmalign* was used to align the database to the initial seed model, and *hmmbuild* was run on the alignment to build the large model. To use the large model for detection of an α-helix cap, *hmmsearch* was executed, using the forward (as opposed to the default Viterbi) algorithm to calculate the score. E-values of 0.5 and above were considered as positive hits. The forward algorithm was chosen because of the high specificity and low sensitivity of the detector; Eddy [190] noted that the forward algorithm is more sensitive to subtle patterns.

An updated version of the PDB^{-} data set from BETAWRAP [116] was used as a negative control data set. The PDB^{-} database used for HELIXCAP is the non-redundant database of all sequences in the October 12, 2007 release of the Protein Data Bank [2] with corresponding elements in SCOP [3], excluding all sequences identified as β-helices. PDB^{-}-HELIXCAP contains 18,659 sequences. For a positive control, cross-validation was conducted as follows. The database of β-helix sequences was clustered using BLASTCLUST [192], creating 69 clusters with greater than 25% sequence identity and 498 clusters with greater than 75% sequence identity. For each cluster, a corresponding cross-validation model was constructed. Cross-validation models were constructed in the same manner as the full model, except that the sequences in the cluster were removed from the database before alignment. Each cluster was then analyzed using its corresponding cross-validation model. Results were pooled for statistical analysis.

Figures

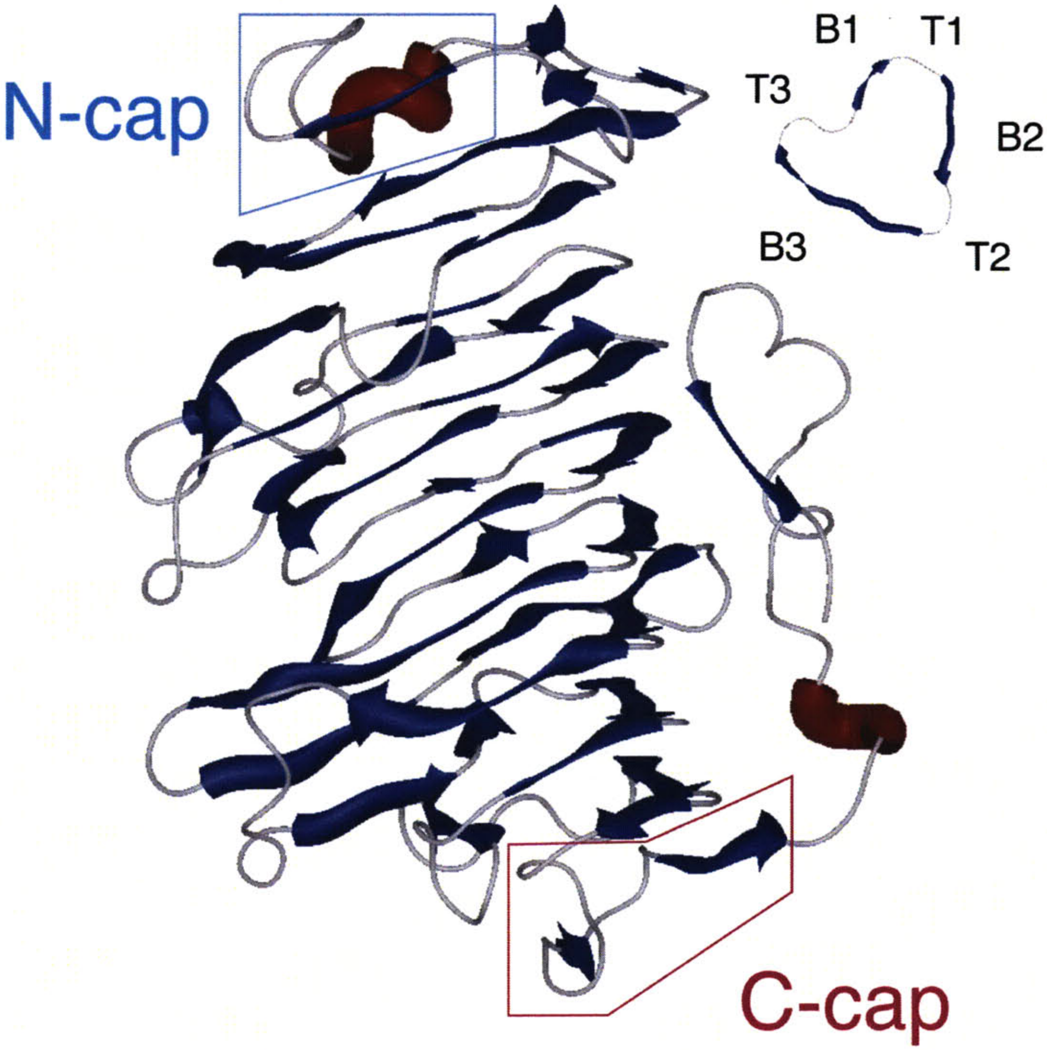


Figure 5-1. Pectin methylesterase, 1GQ8; a typical β-helix. 1GQ8 is a right-handed β-helix, three-sided, with a right-handed α-helix cap at its N-terminus and a previous-strand 'visor' cap at its C-terminus. Inset, the assignment of β-strand and turn names in a β-helix rung.

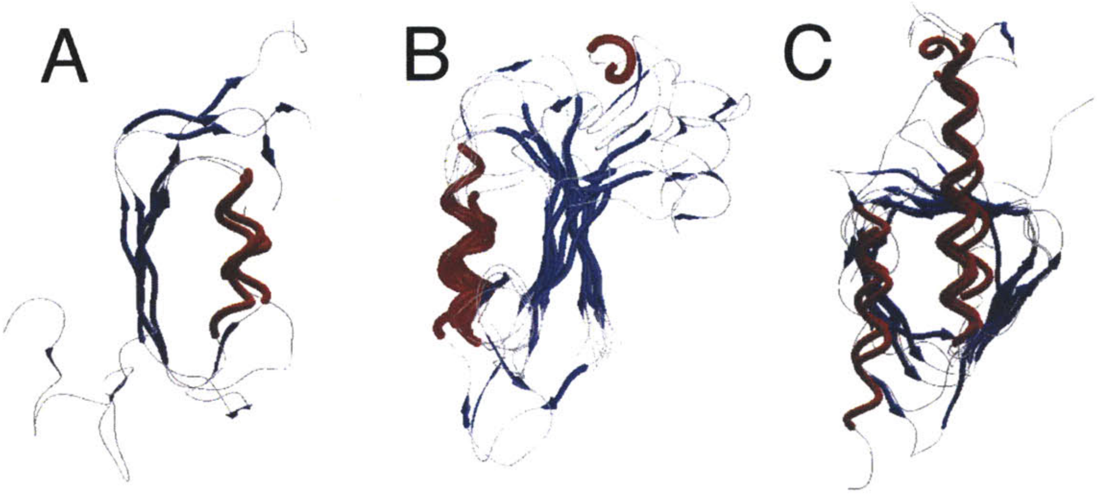


Figure 5-2. α helix caps. Residues more than three residues distal to the cap, or beyond the first rung of the β-helix, are omitted for clarity. A and B depict caps in the pectate lyase superfamily; C depicts caps in the left-handed β-helix families. A, right-handed α-helix caps, top view; B, left-handed α-helix caps, top view; C, right-handed α-helix caps, top view. All views were produced by SWISS-PDB [215].

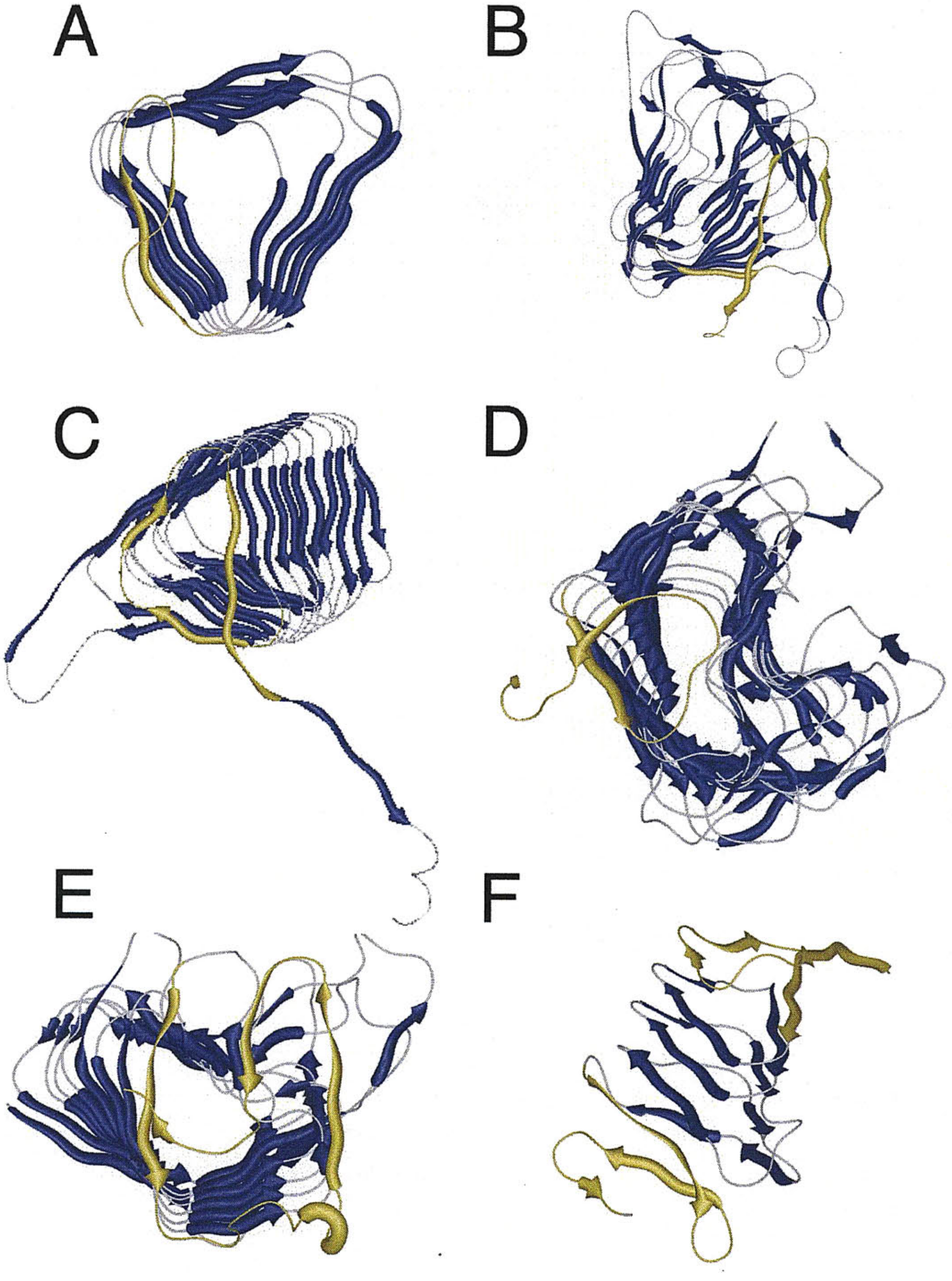
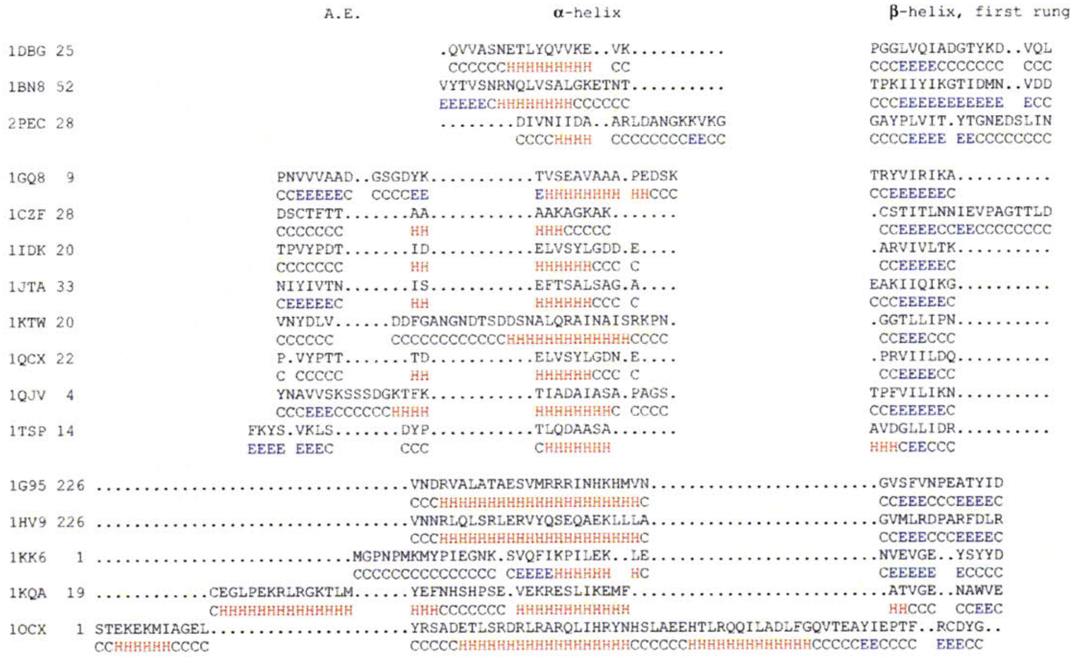

A
B
C
D
E
F

Figure 5-3. The visor family of caps. Visors are heavily variable; these images depict a representative cross-section of folds. For visibility, N-caps and loops outside the β-helix have been removed. From top left: A, 1KQA, a previous-strand visor on a left-handed β-helix; B, 1IDK, a previous-strand visor on a right-handed β-helix; C, 1G95, a cross-helix visor on a left-handed β-helix; D, 1DBG, a cross-helix visor on a right-handed β-helix; E, 2PEC, a visor + α visor on a right-handed β-helix; F, 1HF2, a structure with visor caps at both the N- and C-terminal ends.



```
                                 A.E.                   α-helix                            β-helix, first rung

1DBG 25                                   .QVVASNETLYQVVKE..VK..........                 PGGLVQIADGTYKD..VQL
                                           CCCCCCHHHHHHHHH  CC                           CCCEEEECCCCCCC  CCC
1BN8 52                                   VYTVSNRNQLVSALGKETNT..........                 TPKIIYIKGTIDMN..VDD
                                          EEEEECHHHHHHHHCCCCCC                           CCCEEEEEEEEEEE  ECC
2PEC 28                                   ........DIVNIIDA..ARLDANGKKVKG                 GAYPLVIT.YTGNEDSLIN
                                                  CCCCHHHH  CCCCCCCCEECC                 CCCCEEEE EECCCCCCCC

1GQ8   9                    PNVVVAAD..GSGDYK...........TVSEAVAAA.PEDSK                   TRYVIRIKA..........
                            CCEEEEEC  CCCCEE           EHHHHHHHH HHCCC                   CCEEEEEEC
1CZF 28                     DSCTFTT.......AA...........AAKAGKAK.......                   .CSTITLNNIEVPAGTTLD
                            CCCCCCC       HH           HHHCCCCC                           CCEEEECCEECCCCCCCC
1IDK 20                     TPVYPDT.......ID...........ELVSYLGDD.E....                   .ARVIVLTK..........
                            CCCCCCC       HH           HHHHHHCCC C                        CCEEEEEC
1JTA 33                     NIYIVTN.......IS...........EFTSALSAG.A....                   EAKIIQIKG..........
                            CEEEEEC       HH           HHHHHHCCC C                       CCCEEEEEC
1KTW 20                     VNYDLV......DDFGANGNDTSDDSNALQRAINAISRKPN.                   .GGTLLIPN..........
                            CCCCCC      CCCCCCCCCCCCHHHHHHHHHHHHHCCCC                     CCEEECCC
1QCX 22                     P.VYPTT.......TD...........ELVSYLGDN.E....                   .PRVIILDQ..........
                            C CCCCC       HH           HHHHHHCCC C                        CCEEEECC
1QJV   4                    YNAVVSKSSSDGKTFK...........TIADAIASA.PAGS.                   TPFVILIKN..........
                            CCCEEECCCCCCHHHH           HHHHHHHHC CCCC                    CCEEEEEEC
1TSP 14                  FKYS.VKLS.......DYP...........TLQDAASA.......                   AVDGLLIDR..........
                         EEEE EEEC       CCC           CHHHHHHH                          HHHCEECCC

1G95 226 .................................VNDRVALATAESVMRRRINHKHMVN.......................GVSFVNPEATYID
                                          CCCHHHHHHHHHHHHHHHHHHHHHC                       CCEEECCCEEEEC
1HV9 226 .................................VNNRLQLSRLERVYQSEQAEKLLLA.......................GVMLRDPARFDLR
                                          CCCHHHHHHHHHHHHHHHHHHHHHC                       CCEEECCCEEEEC
1KK6   1 ...........................MGPNPMKMYPIEGNK.SVQFIKPILEK..LE.......................NVEVGE..YSYYD
                                    CCCCCCCCCCCCCCC CEEEEHHHHHH  HC                       CEEEEE  ECCCC
1KQA  19 ............CEGLPEKRLRGKTLM......YEFNHSHPSE.VEKRESLIKEMF..........................ATVGE..NAWVE
                     CHHHHHHHHHHHHHH      HHHCCCCCCC HHHHHHHHHHHH                          HHCCC  CCEEC
1OCX   1 STEKEKMIAGEL.....................YRSADETLSRDRLRARQLIHRYNHSLAEEHTLRQQILADLFGQVTEAYIEPTF..RCDYG..
         CCHHHHHHCCCC                     CCCCCHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHCCCCCEECCCC  EEECC
```

Figure 5-4. Alignment of α-helix caps. At top, α-left caps from the pectate lyase superfamily; middle, α-right caps from the pectate lyase superfamily; at bottom, α-right caps from the left-handed superfamily. The α-helix and additional element (A.E.) are labeled.

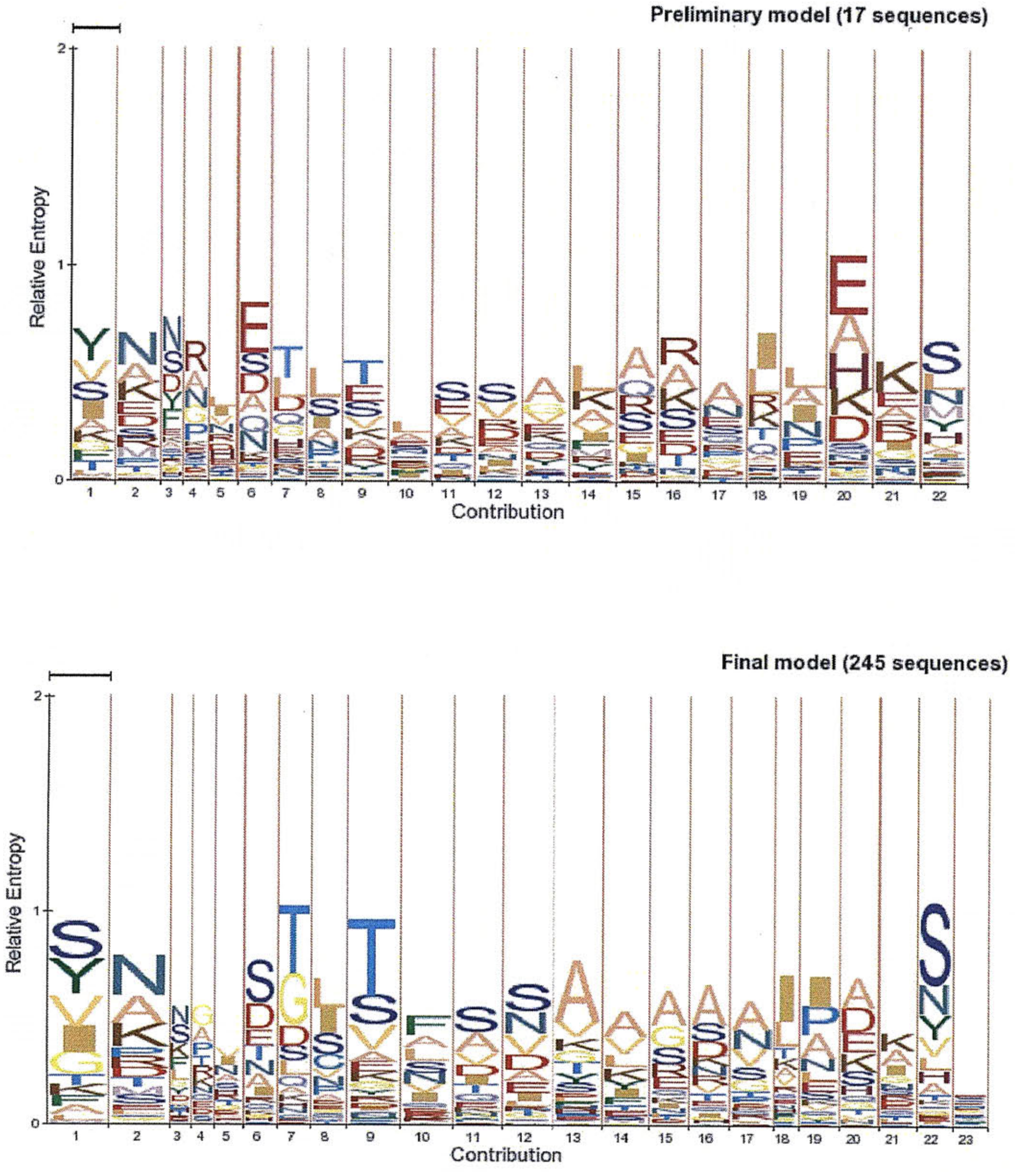


Figure 5-5. HMM-Logo representations [191] of the α-helix-cap predictive model. A, the initial model constructed from 26 aligned crystal structures. B, the augmented model constructed from 1083 sequences aligned to the initial model.

Tables

Table 5-1: Classification of β-helix caps by type.

PDB ID	Protein name	N-cap	Residues	C-cap	Residues
Pectate lyase superfamily					
1BHE	Polygalacturonase	α-right	19-51	Previous-strand visor	364-376
1BN8	Pectate lyase	α-left	32-48	Previous-strand visor	354-368
1CZF	Endopolygalacturonase II	α-right	28-45	Cross-helix visor	347-362
1DAB	P.69 pertactin virulence factor	*		Loop	520-528
1DBG	Chondroitinase B	α-left	26-40	Cross-helix visor	421-428
1EE6	Pectate lyase	*		*	
1GQ8	Pectin methylesterase	α-right	8-32	Previous-strand visor	270-286
1HG8	Polygalacturonase	α-right	25-43	Cross-helix visor	357-373
1IA5	Polygalacturonase	α-right	1-24	Cross-helix visor	323-339
1IDK	Pectin lyase A	α-right	23-38	Cross-helix visor	317-324
1JTA	Pectin lyase A	α-right	34-52	Previous-strand visor	320-329
1K5C	Endopolygalacturonase I	α-right	1-17	*	
1KTW	Iota-carrageenase	α-left	46-79	Loop	438-450
1NHC	Endopolygalacturonase I	α-right	33-51	Cross-helix visor	352-368
1OGM	Dextranase	Cross-helix visor	215-231	Previous-strand visor	544-574
1PCL	Pectate lyase E	α-right	27-44	Visor + α	280-324
1QCX	Pectin lyase B	α-right	23-39	Cross-helix visor	317-328

1QJV	Aspartyl esterase	α-right	29-56	Previous-strand visor	330-344
1RMG	Rhamnogalacturonase A	α-right	19-51	Cross-helix visor	343-364
1RU4	Pectate transeliminase	α-right	38-72	Cross-helix visor	349-360
1RWR	Filamentous hemagglutinin FhaB	Cross-strand visor	1-28	*	
1TSP	P22 tailspike	α-right	124-132	Interleaved β-helix[§]	
2PEC	Pectate lyase C	α-left	19-40	Visor + α	307-314
Left-handed superfamily					
1EWW	Spruce budworm antifreeze	Loop	1-18	*	
1G95	GLMU	α-right	239-255	Previous-strand visor	426-439
1HV9	UDP-N-acetylglucosamine pyrophosphorylase	α-right	239-256	Previous-strand visor	426-440
1KK6	Streptogramin A acetyltransferase	α-right	16-33	Previous-strand visor	155-169
1KQA	Galactoside O-acetyltransferase	α-right	22-54	Cross-helix visor	172-184
1LXA	UDP-N-acetylglucosamine acyltransferase	*		Visor + loop	185-197 260-262
1M8N	Isoform 501, spruce budworm antifreeze	*		Previous-strand visor	103-121
1OCX	Maltose O-acetyltransferase	α-right	19-53	Previous-strand visor	169-183
1SSM	Serine acetyltransferase	α-right	102-137	Previous-strand visor	234-240
1T3D	Serine acetyltransferase	α-right	105-142	Cross-helix visor	233-247
1TDT	Tetrahydrodiplicolinate N-	Loop	87-107	Cross-helix visor	249-256

	succinyltransferase				
1THJ	Carbonic anhydrase	Loop	8-13	α-right	171-185
1XAT	Hexapeptide xenobiotic acetyltransferase	Visor	11-31	Previous-strand visor	151-164
Non-pectate-lyase superfamily					
1EA0	Glutamase synthase	Double α-right	1210-1236	Visor + α	1400-1415
1EZG	*Tenebrio molitor* antifreeze	*		*	
1HF2	MINC	Visor	96-110	Cross-helix visor	187-202
1K4Z	CAP-1	*		Cross-helix visor	1492-1508
1LLZ	Glutamate synthase	Double α-right	1248-1300	Visor + α	1436-1451
Other families					
1K28	T4 cell-puncturing device	β-sheet[§]	389-433	*	
1WMR	Isopullulanase	Visor	200-214	Loop	535-564
2ARA	ARAC	Loop	110-118	*	

* Cap not in crystal structure.

[#] Residue numbers given from start of sequence present in crystal structure.

[§]Part of an adjoining domain.

Table 5-2: List of amyloidogenic proteins tested with helix-cap detector.

Protein name	Organism
Amyloid beta	*Homo sapiens*
Het-s	*Podospora anserine*

Amylin (IAPP)	*Homo sapiens*
Alpha-synuclein	*Homo sapiens*
Tau fibrillar protein	*Homo sapiens*
Curli A (csgA)	*Escherichia coli* K12
Curli B (csgB)	*Escherichia coli* K12
PrP	*Homo sapiens*
Huntingtin	*Homo sapiens*
Huntingtin	*Canis lupus familiaris*
Huntingtin	*Mus musculus*
Huntingtin	*Rattus norvegicus*
Huntingtin	*Sus scrofa*
Huntingtin	*Gallus gallus*
Huntingtin	*Danio rerio*
Huntingtin	*Takifugu rubripes*
Huntingtin	*Pan troglodytes*
Huntingtin analog gi|47220259	*Tetraodon nigroviridis*
Huntingtin analog HD-prov	*Xenopus tropicalis*
Huntingtin analog gi|72175056	*Strongylocentrotus purpuratus*
Huntingtin analog ENSANGP00000020200	*Anopheles gambiae*
Huntingtin analog CG9995-PA	*Drosophila melanogaster*
Huntingtin analog DDB0206578	*Dictyostelium discoideum*
Sup35p (ERF3)	*Saccharomyces cerevisiae*
ERF3	*Candida glabrata* CBS138
ERF3	*Zygosaccharomyces rouxii*

EFR3	*Kluyveromyces lactis*
ERF3 (AGL145Wp)	*Ashbya gossypii*
ERF3	*Candida albicans*
ERF3	*Candida albicans* SC5314
ERF3	*Debaryomyces hansenii*
ERF3	*Debaryomyces hansenii* CBS767
Rnq1p	*Saccharomyces cerevisiae*
New1p	*Saccharomyces cerevisiae*
Ure2p	*Saccharomyces cerevisiae*
Acyl phosphatase	*Homo sapiens*
Beta-microglobulin	*Homo sapiens*
Cro repressor	Phage 434
Myoglobin	*Physeter macrocephalus*
Myohemerythrin	*Themiste zostericola*
Plastocyanin	*Phaseolus vulgaris*
Bovine pancreatic trypsin inhibitor	*Bos Taurus*
Ribosomal protein L9	*Bacillus stearothermophilus*
Glutathione S-transferase	*Homo sapiens*
Spectrin SH3	*Gallus gallus*
Ada-2h	*Homo sapiens*
Ara	*Arabidopsis thaliana*
Com-A	*Bacillus subtilis*
CheY	*Escherichia coli*
Flavodoxin	*Clostridium mp*

P21-ras	*Homo sapiens*
PL-B1	*Peptostreptococcus magnus*
Protein G	*Streptococcus sp.* group g

Conclusions and Future Paths

The prediction and understanding of overall β-structure remains a challenge. Nonetheless, the results of the BETASCAN-HELIXCAP-STITCHER set of algorithms demonstrate that the challenge can be selectively overcome. With sufficient restriction of conformational space and utilization of the cooperative and piecewise assembly of protein structure from simpler elements, Levinthal's paradox becomes merely a surmountable barrier to computational analysis. A day can be foreseen when flexible templates, free-energy assessments, and probabilistic prediction are combined in new ways to solve even larger subsets of the general protein folding problem.

Several obstacles remain to the full generalization of the "divide and conquer" method. Attempts to provide BETASCAN with the ability to analyze both parallel and antiparallel strand-pairs foundered on the inability to compare the two types of β-alignment on an equal basis. While estimates of antiparallel frequency with regard to length exist, no empirically based formula could be found to estimate the relative probability of a parallel configuration with that of an antiparallel configuration that also preserved the length normalization.

The STITCHER algorithm was designed to permit the further extension of sheets between monomers to model the assembly of full amyloid fibrils. While this can be accomplished with some deft manipulation of input for interfaces composed of parallel β-strands, the antiparallel cases would depend on the yet-to-be-implemented antiparallel BETASCAN scoring. Automation of analysis for the various parallel and antiparallel interface cases would be a welcome and useful addition to the STITCHER protocol.

Finally, the HELIXCAP algorithm provides not only a useful negative detector to rule out amyloidogenicity in protein sequences, but enhances our understanding of the types and effects of β-helix caps. Experiments are now ongoing to determine if removal of the N-terminal cap of pertactin has the predicted dimerizing effects. Preliminary experiments with the removal of the entire N-terminal half of pertactin have indeed resulted in a crystallizable dimer of the C-terminal halves of the pertactin β-helix. With further understanding of the binding of helix caps to the first rung of the β-helix, there arises the possibility of a new class of pharmaceutical inhibitors intended to mimic the helix-capping interaction. Such inhibitors could be designed to either stabilize amyloidogenic proteins against fibril formation or to retard the growth of fibrils already present.

Sixty years after the first elucidation of secondary structures in proteins, the paths forward to a full understanding of the workhorse molecules of life on Earth remain murky. The amyloid and prion proteins, a puzzling mystery of folding vital to medicine and biology, present a new perspective on protein folding that both challenge and affirm long-held assumptions about the causes of and influences upon the specificity of fold conformations. By subdividing the folding problem into tractable components, utilizing our full computational and algorithmic capabilities, and intuiting the means to assemble full structures from the most likely elements, parts of the amyloid/prion folding problem can now be attacked productively. A way now lies open to further understanding by both experimentalists and computational biologists of the fundamental properties of these most intriguing molecules.

References

1. Pauling, L. and R.B. Corey, *The Pleated Sheet, A New Layer Configuration of Polypeptide Chains.* Proceedings of the National Academy of Sciences of the United States of America, 1951. **37**(5): p. 251-256.
2. Berman, H.M., et al., *The Protein Data Bank.* Nucleic acids research, 2000. **28**(1): p. 235-42.
3. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32(Database issue)**: p. D226-9.
4. Hunter, P., *Into the fold. Advances in technology and algorithms facilitate great strides in protein structure prediction.* EMBO Rep, 2006. **7**(3): p. 249-52.
5. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
6. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.* Proceedings of the National Academy of Sciences of the United States of America, 1961. **47**(9): p. 1309-1314.
7. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.
8. Levinthal, C., *Are there pathways for protein folding.* J. Chim. Phys, 1968. **65**(1): p. 44-45.
9. Eisenberg, D., *The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11207-10.
10. Pauling, L., R.B. Corey, and H.R. Branson, *The Structure of Proteins.* Proceedings of the National Academy of Sciences of the United States of America, 1951. **37**(4): p. 205-211.
11. Blake, C.C.F., et al., *Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 [angst] Resolution.* Nature, 1965. **206**(4986): p. 757-761.
12. Kendrew, J.C., et al., *A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis.* Nature, 1958. **181**(4610): p. 662-666.
13. Cullis, A.F., et al., *The Structure of Haemoglobin. IX. A Three-Dimensional Fourier Synthesis at 5.5 [angst] Resolution: Description of the Structure.* Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1962. **265**(1321): p. 161-187.
14. Linderstrøm-Lang, K., *Proteins and enzymes.* 1952, Stanford,: Stanford University Press. vi, 115 p.
15. Berman, H., *The Protein Data Bank: a historical perspective.* Acta Crystallographica Section A, 2008. **64**(1): p. 88-95.
16. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.
17. Overhauser, A.W., *Polarization of Nuclei in Metals.* Physical Review, 1953. **92**(2): p. 411.
18. Bax, A. and S. Grzesiek, *Methodological advances in protein NMR.* Accounts of Chemical Research, 2002. **26**(4): p. 131-138.
19. Richardson, J.S., *The anatomy and taxonomy of protein structure.* Adv Protein Chem, 1981. **34**: p. 167-339.

20. Cuff, A.L., et al., *The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies.* Nucl. Acids Res., 2009. **37**(suppl_1): p. D310-314.
21. Murzin, A.G., et al. *Structural Classification of Proteins: SCOP Classification Statistics.* [HTML] 2009 February 23 [cited 2009 July 8]; Available from: http://scop.mrc-lmb.cam.ac.uk/scop/count.html#scop-1.75.
22. Bryngelson, J.D. and P.G. Wolynes, *Intermediates and barrier crossing in a random energy model (with applications to protein folding).* The Journal of Physical Chemistry, 2002. **93**(19): p. 6902-6915.
23. Virchow, R.L.K., *Cellular pathology as based upon physiological and pathological histology.* 1971, New York,: Dover Publications. xxvii, 554 p.
24. Kyle, R.A., *Amyloidosis: a convoluted story.* British Journal of Haematology, 2001. **114**(3): p. 529-538.
25. Aterman, K., *'A Pretty a vista Reaction for Tissues with Amyloid Degeneration,'1875: An Important Year for Pathology.* Journal of the History of Medicine and Allied Sciences, 1976. **31**(4): p. 431.
26. Bennhold, H., *Specific staining of amyloid by Congo red.* Münchener Medizinische Wochenschrift, 1922. **69**: p. 1537–1538.
27. Kyle, R.A., *Amyloidosis: the last three centuries.* Amyloid and Amyloidosis. Bely, M, Apathy, A (Eds), 2001.
28. Grundke-Iqbal, I., et al., *Amyloid protein and neurofibrillary tangles coexist in the same neuron in Alzheimer disease.* Proc Natl Acad Sci U S A, 1989. **86**(8): p. 2853-7.
29. Cohen, A.S. and E. Calkins, *Electron Microscopic Observations on a Fibrous Component in Amyloid of Diverse Origins.* Nature, 1959. **183**(4669): p. 1202-1203.
30. Eanes, E.D. and G.G. Glenner, *X-ray diffraction studies on amyloid proteins.* J. Histochem. Cytochem., 1968. **16**(11): p. 673-677.
31. Kishimoto, A., et al., *[beta]-Helix is a likely core structure of yeast prion Sup35 amyloid fibers.* Biochemical and biophysical research communications, 2004. **315**(3): p. 739-745.
32. Kajava, A.V., J.M. Squire, and D.A. Parry, *Beta-structures in fibrous proteins.* Advances in protein chemistry, 2006. **73**: p. 1-15.
33. Cascio, M., P.A. Glazer, and B.A. Wallace, *The secondary structure of human amyloid deposits as determined by circular dichroism spectroscopy.* Biochemical and biophysical research communications, 1989. **162**(3): p. 1162-6.
34. Sunde, M. and C. Blake, *The structure of amyloid fibrils by electron microscopy and X-ray diffraction.* Advances in protein chemistry, 1997. **50**: p. 123-59.
35. Tessier, P.M. and S. Lindquist, *Unraveling infectious structures, strain variants and species barriers for the yeast prion [PSI+].* Nat Struct Mol Biol, 2009. **16**(6): p. 598-605.
36. Griffith, J.S., *Self-replication and scrapie.* Nature, 1967. **215**(5105): p. 1043-4.
37. Prusiner, S.B., *Prions.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(23): p. 13363-83.
38. Alper, T., D.A. Haig, and M.C. Clarke, *The exceptionally small size of the scrapie agent.* Biochemical and biophysical research communications, 1966. **22**(3): p. 278-284.
39. Bolton, D.C., M.P. McKinley, and S.B. Prusiner, *Identification of a protein that purifies with the scrapie prion.* Science, 1982. **218**(4579): p. 1309-1311.

40. Aguzzi, A. and M. Polymenidou, *Mammalian Prion Biology: One Century of Evolving Concepts.* 2004. **116**(2): p. 313-327.

41. Wickner, R.B., *[URE3] as an altered URE2 protein: evidence for a prion analog in Saccharomyces cerevisiae.* Science, 1994. **264**(5158): p. 566.

42. Brown, J.C.S. and Massachusetts Institute of Technology. Dept. of Biology., *[GAR\207A] : a novel type of prion involving in glucose signaling and environmental sensing in S. cerevisiae.* 2008. p. 195 p.

43. Glover, J.R., et al., *Self-seeded fibers formed by Sup35, the protein determinant of [PSI+], a heritable prion-like factor of S. cerevisiae.* Cell, 1997. **89**(5): p. 811-9.

44. Paushkin, S.V., et al., *Propagation of the yeast prion-like [psi+] determinant is mediated by oligomerization of the SUP35-encoded polypeptide chain release factor.* The EMBO Journal, 1996. **15**(12): p. 3127.

45. Shorter, J. and S. Lindquist, *Prions as adaptive conduits of memory and inheritance.* Nature Reviews Genetics, 2005. **6**(6): p. 435-450.

46. Si, K., S. Lindquist, and E.R. Kandel, *A neuronal isoform of the aplysia CPEB has prion-like properties.* Cell, 2003. **115**(7): p. 879-91.

47. Maddelein, M.L., et al., *Amyloid aggregates of the HET-s prion protein are infectious.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(11): p. 7402-7.

48. Weissmann, C., *The state of the prion.* Nat Rev Micro, 2004. **2**(11): p. 861-871.

49. Lansbury Jr., P., *In pursuit of the molecular structure of amyloid plaque: new technology provides unexpected and critical information.* Biochemistry, 1992. **31**(30): p. 6865-6870.

50. Serpell, L.C., et al., *Fiber diffraction of synthetic alpha-synuclein filaments shows amyloid-like cross-beta conformation.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(9): p. 4897-902.

51. Nelson, R., et al., *Structure of the cross-beta spine of amyloid-like fibrils.* Nature, 2005. **435**(7043): p. 773-8.

52. Wetzel, R., *Ideas of Order for Amyloid Fibril Structure.* Structure, 2002. **10**(8): p. 1031-1036.

53. Sipe, J. and A. Cohen, *Review: History of the Amyloid Fibril.* Journal of Structural Biology, 2000. **130**(2-3): p. 88-98.

54. DePace, A.H. and J.S. Weissman, *Origins and kinetic consequences of diversity in Sup35 yeast prion fibers.* Nature structural biology, 2002. **9**(5): p. 389-96.

55. Tanaka, M., et al., *Conformational variations in an infectious protein determine prion strain differences.* Nature, 2004. **428**(6980): p. 323-8.

56. Tessier, P. and S. Lindquist, *Prion recognition elements govern nucleation, strain specificity and species barriers.* Nature, 2007. **447**(7144): p. 556-561.

57. Krishnan, R. and S. Lindquist, *Structural insights into a yeast prion illuminate nucleation and strain diversity.* Nature, 2005. **435**(7043): p. 765-772.

58. Selkoe, D., *Folding proteins in fatal ways.* Nature, 2003. **426**: p. 900-904.

59. Perfetti, V., et al., *Analysis of Vlambda -Jlambda expression in plasma cells from primary (AL) amyloidosis and normal bone marrow identifies 3r (lambda III) as a new amyloid-associated germline gene segment.* Blood, 2002. **100**(3): p. 948-953.

60. Lachmann, H.J., et al., *Natural history and outcome in systemic AA amyloidosis.* N Engl J Med, 2007. **356**(23): p. 2361-71.

61. Haass, C. and D.J. Selkoe, *Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid [beta]-peptide.* Nat Rev Mol Cell Biol, 2007. **8**(2): p. 101-112.

62. Masters, C.L., et al., *Amyloid plaque core protein in Alzheimer disease and Down syndrome.* Proceedings of the National Academy of Sciences of the United States of America, 1985. **82**(12): p. 4245-4249.

63. Walker, F.O., *Huntington's disease.* Lancet, 2007. **369**(9557): p. 218-28.

64. Dobson, C., *The structural basis of protein folding and its links with human disease.* Philosophical Transactions: Biological Sciences, 2001. **356**(1406): p. 133-145.

65. Westermark, P., et al., *Amyloid: toward terminology clarification. Report from the Nomenclature Committee of the International Society of Amyloidosis.* Amyloid, 2005. **12**(1): p. 1-4.

66. Fowler, D.M., et al., *Functional amyloid formation within mammalian tissue.* PLoS biology, 2006. **4**(1): p. e6.

67. Chapman, M.R., et al., *Role of Escherichia coli curli operons in directing amyloid fiber formation.* Science (New York, N.Y.), 2002. **295**(5556): p. 851-5.

68. Chien, P., J.S. Weissman, and A.H. DePace, *Emerging principles of conformation-based prion inheritance.* Annual review of biochemistry, 2004. **73**(1): p. 617-656.

69. Masel, J. and A. Bergman, *The evolution of the evolvability properties of the yeast prion [PSI+].* Evolution, 2003. **57**(7): p. 1498-1512.

70. True, H.L. and S.L. Lindquist, *A yeast prion provides a mechanism for genetic variation and phenotypic diversity.* Nature, 2000. **407**(6803): p. 477-483.

71. True, H.L., I. Berlin, and S.L. Lindquist, *Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits.* Nature, 2004. **431**(7005): p. 184-187.

72. Maji, S.K., et al., *Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules.* Science, 2009. **325**(5938): p. 328-332.

73. Xu, S., B. Bevis, and M.F. Arnsdorf, *The assembly of amyloidogenic yeast sup35 as assessed by scanning (atomic) force microscopy: an analogy to linear colloidal aggregation?* Biophys J, 2001. **81**(1): p. 446-54.

74. Cohen, F.E. and J.W. Kelly, *Therapeutic approaches to protein-misfolding diseases.* Nature, 2003. **426**(6968): p. 905-9.

75. Sawaya, M.R., et al., *Atomic structures of amyloid cross-beta spines reveal varied steric zippers.* Nature, 2007. **447**(7143): p. 453-7.

76. Hirota-Nakaoka, N., et al., *Dissolution of beta2-microglobulin amyloid fibrils by dimethylsulfoxide.* Journal of biochemistry, 2003. **134**(1): p. 159-64.

77. Krishnan, R. and S.L. Lindquist, *Structural insights into a yeast prion illuminate nucleation and strain diversity.* Nature, 2005. **435(7043)**: p. 765-72.

78. Serio, T.R. and S.L. Lindquist, *The yeast prion [PSI+]: molecular insights and functional consequences.* Adv Protein Chem, 2001. **59**: p. 391-412.

79. Petkova, A.T., et al., *A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16742-7.

80. Lührs, T., et al., *3D structure of Alzheimer's amyloid-beta(1-42) fibrils.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(48): p. 17342-7.

81. Ritter, C., et al., *Correlation of structural elements and infectivity of the HET-s prion.* Nature, 2005. **435**(7043): p. 844-8.

82. Wasmer, C., et al., *Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core.* Science, 2008. **319(5869)**: p. 1523-6.

83. DeLano, W.L., *The PyMOL Molecular Graphics System*. 2008, DeLano Scientific LLC.

84. Moult, J., et al., *Critical assessment of methods of protein structure prediction-Round VII.* Proteins, 2007. **69 Suppl 8**: p. 3-9.

85. Ngo, J.T. and J. Marks, *Computational complexity of a problem in molecular structure prediction.* Protein Eng, 1992. **5**(4): p. 313-21.

86. Unger, R. and J. Moult, *Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications.* Bull Math Biol, 1993. **55**(6): p. 1183-98.

87. Berger, B. and T. Leighton, *Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete.* J Comput Biol, 1998. **5**(1): p. 27-40.

88. Fraenkel, A.S., *Complexity of protein folding.* Bulletin of Mathematical Biology, 1993. **55**(6): p. 1199-1210.

89. Go, N. and H.A. Scheraga, *On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation.* Macromolecules, 1976. **9**(4): p. 535-542.

90. Li, Z. and H.A. Scheraga, *Monte Carlo-minimization approach to the multiple-minima problem in protein folding.* Proceedings of the National Academy of Sciences, 1987. **84**(19): p. 6611-6615.

91. Liwo, A., et al., *A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data.* Journal of Computational Chemistry, 1997. **18**(7): p. 849-873.

92. Bradley, P., K.M.S. Misura, and D. Baker, *Toward High-Resolution de Novo Structure Prediction for Small Proteins.* Science, 2005. **309**(5742): p. 1868-1871.

93. Singh, R. and B. Berger. *ChainTweak: sampling from the neighbourhood of a protein conformation*. 2004: World Scientific.

94. Ye, Y. and A. Godzik, *Flexible structure alignment by chaining aligned fragment pairs allowing twists*. 2003, Oxford Univ Press. p. 246-255.

95. Menke, M., B. Berger, and L. Cowen, *Matt: Local Flexibility Aids Protein Multiple Structure Alignment.* PLoS Comput Biol, 2008. **4**(1): p. e10.

96. Peng, Y. and U.H.E. Hansmann, *Helix versus sheet formation in a small peptide.* Physical Review E, 2003. **68**(4): p. 041911.

97. Prusiner, S.B., *Prion Biology and Diseases*. 2004, New York: Cold Spring Harbor Laboratory Press.

98. Chou, P.Y. and G.D. Fasman, *Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.* Biochemistry, 1974. **13**(2): p. 211-22.

99. Chen, H., F. Gu, and Z. Huang, *Improved Chou-Fasman method for protein secondary structure prediction.* BMC Bioinformatics, 2006. **7 Suppl 4**: p. S14.

100. Cuff, J.A. and G.J. Barton, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction.* Proteins, 2000. **40**(3): p. 502-11.

101. Garnier, J.O., D.J. Osguthorpe, and B. Robson, *Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.* J. Mol. Biol, 1978. **120**: p. 97-120.

102. Zimmermann, O. and U.H.E. Hansmann, *Support vector machines for prediction of dihedral angle regions.* Bioinformatics, 2006. **22**(24): p. 3009-3015.

103. Kuang, R., C.S. Leslie, and A.-S. Yang, *Protein backbone angle prediction with machine learning approaches.* Bioinformatics, 2004. **20**(10): p. 1612-1621.

104. Mount, D.W., *Bioinformatics : sequence and genome analysis*. 2nd ed. 2004, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. xii, 692 p.

105. Jones, D.T., *THREADER: Protein sequence threading by double dynamic programming.*, in *Computational methods in molecular biology*. 1998, Elsevier: Amsterdam; New York. p. xxvi, 371 p.

106. Chenna, R., et al., *Multiple sequence alignment with the Clustal series of programs.* Nucleic acids research, 2003. **31**(13): p. 3497-500.

107. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-826.

108. Rost, B., *Twilight zone of protein sequence alignments*. 1999, Oxford Univ Press. p. 85-94.

109. Abagyan, R.A. and S. Batalov, *Do aligned sequences share the same fold?* Journal of molecular biology, 1997. **273**(1): p. 355-368.

110. Edgar, R.C. and S. Batzoglou, *Multiple sequence alignment.* Curr Opin Struct Biol, 2006. **16**(3): p. 368-373.

111. Whisstock, J.C. and A.M. Lesk, *Prediction of protein function from protein sequence and structure.* Quarterly reviews of biophysics, 2004. **36**(03): p. 307-340.

112. Baldwin, R.L., *In search of the energetic role of peptide hydrogen bonds.* J Biol Chem, 2003. **278**(20): p. 17581-8.

113. Berger, B., *Algorithms for protein structural motif recognition.* J Comput Biol, 1995. **2**(1): p. 125-38.

114. McDonnell, A.V., et al., *Paircoil2: improved prediction of coiled coils from sequence.* Bioinformatics, 2006. **22**(3): p. 356-8.

115. Perutz, M., et al., *Amyloid fibers are water-filled nanotubes.* Proceedings of the National Academy of Sciences, 2002. **99**(8): p. 5591.

116. Bradley, P., et al., *BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(26): p. 14819-24.

117. McDonnell, A.V., et al., *Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins.* Proteins, 2006. **63**(4): p. 976-85.

118. Menke, M., et al., *Wrap-and-Pack: a new paradigm for beta structural motif recognition with application to recognizing beta trefoils.* J Comput Biol, 2005. **12**(6): p. 777-95.

119. Waldispuhl, J., et al., *Modeling ensembles of transmembrane beta-barrel proteins.* Proteins, 2008. **71**(3): p. 1097-112.

120. Yoder, M.D., N.T. Keen, and F. Jurnak, *New domain motif: the structure of pectate lyase C, a secreted plant virulence factor.* Science, 1993. **260(5113)**: p. 1503-7.

121. Dobson, C., *Protein folding and misfolding.* Nature, 2003. **426**(6968): p. 884-890.

122. Bucciantini, M., et al., *Inherent cytotoxicity of aggregates implies a common origin for protein misfolding diseases.* Nature, 2002. **416**: p. 507-511.

123. Wickner, R.B., et al., *Prions of fungi: inherited structures and biological roles.* Nat Rev Microbiol, 2007. **5(8)**: p. 611-8.

124. Uptain, S.M. and S. Lindquist, *Prions as protein-based genetic elements.* Annual review of microbiology, 2002. **56**: p. 703-41.

125. Soto, C. and E.M. Castaäno, *The conformation of Alzheimer's beta peptide determines the rate of amyloid formation and its resistance to proteolysis.* The Biochemical journal, 1996. **314**: p. 701-7.

126. Petkova, A.T., et al., *Solid state NMR reveals a pH-dependent antiparallel beta-sheet registry in fibrils formed by a beta-amyloid peptide.* Journal of molecular biology, 2004. **335**(1): p. 247-60.

127. Wickner, R., F. Dyda, and R. Tycko, *Amyloid of Rnq1p, the basis of the [PIN+] prion, has a parallel in-register {beta}-sheet structure.* Proceedings of the National Academy of Sciences, 2008. **105**(7): p. 2403.

128. Serpell, L., *Alzheimer's amyloid fibrils: structure and assembly.* BBA-Molecular Basis of Disease, 2000. **1502**(1): p. 16-30.

129. Michelitsch, M. and J. Weissman, *A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions.* Proceedings of the National Academy of Sciences, 2000. **97**(22): p. 11910.

130. Perutz, M.F., et al., *Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(8): p. 5596-600.

131. Derkatch, I.L., et al., *Effects of Q/N-rich, polyQ, and non-polyQ amyloids on the de novo formation of the [PSI+] prion in yeast and aggregation of Sup35 in vitro.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(35): p. 12934-9.

132. Goldfarb, L.G., et al., *Transmissible familial Creutzfeldt-Jakob disease associated with five, seven, and eight extra octapeptide coding repeats in the PRNP gene.* Proceedings of the National Academy of Sciences of the United States of America, 1991. **88**(23): p. 10926-30.

133. DePace, A.H., et al., *A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion.* Cell, 1998. **93**(7): p. 1241-52.

134. Rost, B., G. Yachdav, and J. Liu, *The PredictProtein server.* Nucleic acids research, 2004. **32**(Web Server issue): p. W321-6.

135. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-10.

136. Wille, H., et al., *Structural studies of the scrapie prion protein by electron crystallography.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(6): p. 3563-3568.

137. Cheng, J. and P. Baldi, *Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms.* Bioinformatics (Oxford, England), 2005. **21**: p. i75-84.

138. Fernandez-Escamilla, A.M., et al., *Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins.* Nat Biotechnol, 2004. **22**(10): p. 1302-6.

139. Tartaglia, G.G., et al., *Prediction of aggregation-prone regions in structured proteins.* J Mol Biol, 2008. **380**(2): p. 425-36.

140. Chiti, F., et al., *Rationalization of the effects of mutations on peptide and protein aggregation rates.* Nature, 2003. **424**(6950): p. 805-8.

141. Zibaee, S., et al., *A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone.* Protein science : a publication of the Protein Society, 2007. **16**(5): p. 906-18.

142. Trovato, A., F. Seno, and S.C. Tosatto, *The PASTA server for protein aggregation prediction.* Protein Eng Des Sel, 2007. **20**(10): p. 521-3.

143. Trovato, A., et al., *Insight into the structure of amyloid fibrils from the analysis of globular proteins.* PLoS computational biology, 2006. **2**(12): p. e170.

144. Heise, H., et al., *Molecular-level secondary structure, polymorphism, and dynamics of full-length alpha-synuclein fibrils studied by solid-state NMR.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(44): p. 15871-6.
145. von Bergen, M., et al., *Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(10): p. 5129-34.
146. Kajava, A.V., U. Aebi, and A.C. Steven, *The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin.* Journal of molecular biology, 2005. **348**(2): p. 247-52.
147. Sachse, C., M. Fandrich, and N. Grigorieff, *Paired beta-sheet structure of an Abeta(1-40) amyloid fibril revealed by electron microscopy.* Proc Natl Acad Sci U S A, 2008. **105**(21): p. 7462-6.
148. Jenkins, J. and R. Pickersgill, *The architecture of parallel ß-helices and related folds.* Progress in Biophysics and Molecular Biology, 2001. **77**(2): p. 111-175.
149. Esposito, G., et al., *The solution structure of the C-terminal segment of tau protein.* Journal of peptide science : an official publication of the European Peptide Society, 2000. **6**(11): p. 550-9.
150. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
151. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-9.
152. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment.* Proteins, 1995. **23**(4): p. 566-79.
153. Penel, S., et al., *Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings.* Protein engineering, 2003. **16**(12): p. 957-61.
154. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
155. Kajava, A.V., et al., *A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure.* Proc Natl Acad Sci U S A, 2004. **101**(21): p. 7885-90.
156. Shewmaker, F., R.B. Wickner, and R. Tycko, *Amyloid of the prion domain of Sup35p has an in-register parallel beta-sheet structure.* Proc Natl Acad Sci U S A, 2006. **103**(52): p. 19754-9.
157. Alberti, S., et al., *A systematic survey identifies prions and illuminates sequence features of prionogenic proteins.* Cell, 2009. **137**(1): p. 146-58.
158. Petkova, A., et al., *Self-Propagating, Molecular-Level Polymorphism in Alzheimer's ß-Amyloid Fibrils*, in *Science*. 2005, American Association for the Advancement of Science. p. 262-265.
159. Sondheimer, N., et al., *The role of Sis1 in the maintenance of the [RNQ+] prion.* EMBO J, 2001. **20**(10): p. 2435-42.
160. Wickner, R.B., et al., *Prions beget prions: the [PIN+] mystery!* Trends Biochem Sci, 2001. **26**(12): p. 697-9.
161. Bryan, A.W., Jr., et al., *BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis.* PLoS Comput Biol, 2009. **5**(3): p. e1000333.
162. Zhang, C., J.L. Cornette, and C. Delisi, *Consistency in structural energetics of protein folding and peptide recognition.* Protein Sci, 1997. **6**(5): p. 1057-64.
163. Street, A.G. and S.L. Mayo, *Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 9074-6.

164. Avbelj, F., P. Luo, and R.L. Baldwin, *Energetics of the interaction between water and the helical peptide group and its role in determining helix propensities.* Proc Natl Acad Sci U S A, 2000. **97**(20): p. 10786-91.

165. Pal, D. and P. Chakrabarti, *beta-sheet propensity and its correlation with parameters based on conformation.* Acta Crystallogr D Biol Crystallogr, 2000. **56**(Pt 5): p. 589-94.

166. Minor, D.L., Jr. and P.S. Kim, *Context is a major determinant of beta-sheet propensity.* Nature, 1994. **371**(6494): p. 264-7.

167. Gazit, E., *A possible role for pi-stacking in the self-assembly of amyloid fibrils.* FASEB J, 2002. **16**(1): p. 77-83.

168. Jacobson, H. and W.H. Stockmayer, *Intramolecular Reaction in Polycondensations. I. The Theory of Linear Systems.* The Journal of Chemical Physics, 1950. **18**(12): p. 1600-1606.

169. Pace, C.N., et al., *Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds.* J Biol Chem, 1988. **263**(24): p. 11820-5.

170. Yoder, M., S. Lietzke, and F. Jurnak, *Unusual structural features in the parallel beta-helix in pectate lyases.* Structure, 1993. **1**(4): p. 241-51.

171. Liou, Y.C., et al., *Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein.* Nature, 2000. **406(6793)**: p. 322-4.

172. Beaman, T., M. Sugantino, and S. Roderick, *Structure of the hexapeptide xenobiotic acetyltransferase from Pseudomonas aeruginosa.* Biochemistry(Washington), 1998. **37**(19): p. 6689-6696.

173. Yoder, M.D., et al., *New domain motif: the structure of pectate lyase C, a secreted plant virulence factor.* Science, 1993. **260(5113)**: p. 1503-7.

174. Steinbacher, S., et al., *Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer.* Science, 1994. **265(5170)**: p. 383-6.

175. Chothia, C. and A.G. Murzin, *New folds for all-beta proteins.* Structure, 1993. **1(4)**: p. 217-22.

176. Raetz, C.R. and S.L. Roderick, *A left-handed parallel beta helix in the structure of UDP-N-acetylglucosamine acyltransferase.* Science, 1995. **270(5238)**: p. 997-1000.

177. Baumann, U., et al., *Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with a calcium binding parallel beta roll motif.* Embo J, 1993. **12(9)**: p. 3357-64.

178. Richardson, J.S. and D.C. Richardson, *Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation.* Proc Natl Acad Sci U S A, 2002. **99(5)**: p. 2754-9.

179. Wang, W. and M.H. Hecht, *Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins.* Proc Natl Acad Sci U S A, 2002. **99(5)**: p. 2760-5.

180. Perutz, M.F., et al., *Amyloid fibers are water-filled nanotubes.* Proc Natl Acad Sci U S A, 2002. **99(8)**: p. 5591-5.

181. Kisker, C., et al., *A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon Methanosarcina thermophila.* The EMBO Journal, 1996. **15**(10): p. 2323.

182. Cordell, S., R. Anderson, and J. Löwe, *Crystal structure of the bacterial cell division inhibitor MinC.* The EMBO Journal, 2001. **20**: p. 2454-2461.

183. Sugantino, M. and S. Roderick, *Crystal structure of Vat (D): an acetyltransferase that inactivates streptogramin group A antibiotics.* Biochemistry, 2002. **41**(7): p. 2209-16.

184. van Santen, Y., et al., *1.68-A Crystal Structure of Endopolygalacturonase II from Aspergillus niger and Identification of Active Site Residues by Site-directed Mutagenesis.* Journal of Biological Chemistry, 1999. **274**(43): p. 30474-30480.

185. Kostrewa, D., et al., *Crystal Structures of Streptococcus pneumoniaeN-Acetylglucosamine-1-phosphate Uridyltransferase, GlmU, in Apo Form at 2.33 Å Resolution and in Complex with UDP-N-Acetylglucosamine and Mg2+ at 1.96 Å Resolution.* Journal of Molecular Biology, 2001. **305**(2): p. 279-289.

186. Simkovsky, R. and J. King, *An elongated spine of buried core residues necessary for in vivo folding of the parallel beta-helix of P22 tailspike adhesin.* Proc Natl Acad Sci U S A, 2006. **103(10)**: p. 3575-80.

187. Wang, X., L. Olsen, and S. Roderick, *Structure of the lac Operon Galactoside Acetyltransferase.* Structure, 2002. **10**(4): p. 581-588.

188. Mayans, O., et al., *Two crystal structures of pectin lyase A from Aspergillus reveal a pH driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases.* Structure, 1997. **5**(5): p. 677-689.

189. Huang, W., et al., *Crystal structure of chondroitinase B from Flavobacterium heparinum and its complex with a disaccharide product at 1.7 Å resolution.* Journal of Molecular Biology, 1999. **294**(5): p. 1257-1269.

190. Eddy, S., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-763.

191. Schuster-Bockler, B., J. Schultz, and S. Rahmann, *HMM Logos for visualization of protein families.* BMC Bioinformatics, 2004. **5**: p. 7.

192. Altschul, S., et al., *Basic local alignment search tool.* J. Mol. Biol, 1990. **215**(3): p. 403-410.

193. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2008. **36**(Database issue): p. D25-30.

194. Pickersgill, R., et al., *The structure of Bacillus subtilis pectate lyase in complex with calcium.* Nature Structural Biology, 1994. **1**(10): p. 717-723.

195. Emsley, P., et al., *Structure of Bordetella pertussis virulence factor P. 69 pertactin.* Nature, 1996. **381**(6577): p. 90-92.

196. Akita, M., et al., *Crystallization and preliminary X-ray analysis of high-alkaline pectate lyase.* Acta Crystallogr D Biol Crystallogr, 2000. **56**(Pt 6): p. 749-50.

197. Johansson, K., et al., *Crystal structure of plant pectin methylesterase.* FEBS Letters, 2002. **514**(2-3): p. 243-249.

198. Thomas, L., et al., *Structure of pectate lyase A: comparison to other isoforms.* logo. **58**(2 Part 6): p. 1008-1015.

199. Michel, G., et al., *The Structural Bases of the Processive Degradation of ?-Carrageenan, a Main Cell Wall Polysaccharide of Red Algae.* Journal of Molecular Biology, 2003. **334**(3): p. 421-433.

200. Vitali, J., et al., *The three-dimensional structure of Aspergillus niger pectin lyase B at 1.7-Å resolution.* Plant Physiol, 1998. **116**: p. 69-80.

201. Jenkins, J., et al., *Three-dimensional structure of Erwinia chrysanthemi pectin methylesterase reveals a novel esterase active site.* Journal of Molecular Biology, 2001. **305**(4): p. 951-960.

202. Graether, S., et al., *Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect.* Nature, 2000. **406**(6793): p. 249-251.

203. Olsen, L. and S. Roderick, *Structure of the Escherichia coli GlmU Pyrophosphorylase and Acetyltransferase Active Sites.* Biochemistry(Washington), 2001. **40**(7): p. 1913-1921.

204. Leinala, E., et al., *A ß-Helical Antifreeze Protein Isoform with Increased Activity: structural and functional insights.* Journal of Biological Chemistry, 2002. **277**(36): p. 33349-33352.

205. Lo Leggio, L., et al., *The structure and specificity of Escherichia coli maltose acetyltransferase give new insight into the LacA family of acyltransferases.* Biochemistry, 2003. **42**(18): p. 5225-35.

206. Beaman, T., et al., *Three-dimensional structure of tetrahydrodipicolinate N-succinyltransferase.* Biochemistry(Washington), 1997. **36**(3): p. 489-494.

207. Binda, C., et al., *Cross-Talk and Ammonia Channeling between Active Centers in the Unexpected Domain Arrangement of Glutamate Synthase.* Structure, 2000. **8**(12): p. 1299-1308.

208. Dodatko, T., et al., *Crystal structure of the actin binding domain of the cyclase-associated protein.* Biochemistry, 2004. **43**(33): p. 10628-10641.

209. van den Heuvel, R., et al., *Structural Studies on the Synchronization of Catalytic Centers in Glutamate Synthase.* Journal of Biological Chemistry, 2002. **277**(27): p. 24579-24583.

210. Kanamaru, S., et al., *Structure of the cell-puncturing device of bacteriophage T4.* Nature, 2002. **415**(6871): p. 553-7.

211. Mizuno, M., et al., *Crystal structure of Aspergillus niger isopullulanase, a member of glycoside hydrolase family 49.* J Mol Biol, 2008. **376**(1): p. 210-20.

212. Soisson, S., et al., *Structural Basis for Ligand-Regulated Oligomerization of AraC.* Science, 1997. **276**(5311): p. 421.

213. Holm, L. and C. Sander, *Alignment of three-dimensional protein structures: network server for database searching.* Methods Enzymol, 1996. **266**: p. 653-62.

214. Rotkiewicz, P., *iMol, a free Mac OS X molecular visualization tool from PIRX.* Found on the WWW at http://www. pirx. com/iMol, 2003.

215. Schwede, T., et al., *SWISS-MODEL: an automated protein homology-modeling server.* Nucleic acids research, 2003. **31**(13): p. 3381.

Supplementary Tables

Supplemental Table 3-S1, β-helices used in BETASCAN statistical analyses.

PDB id	Protein	Species
1air	Pectate Lyase E	Erwinia chrysanthemi
1bhe	Polygalacturonase	Erwinia carotovora ssp. Carotovora
1bn8	Pectate Lyase	Bacillus subtilis
1czf	Endo-Polygalacturonase II	Aspergillus niger
1dab	Virulence Factor P.69 Pertactin	Bordetella pertussis
1dbg	Chondroitinase B	Flavobacterium heparinum
1ee6	Pectate Lyase	Bacillus sp. Strain KSM-p15.
1h80	Iota-Carrageenase Of	Alteromonas fortis
1hg8	Endopolygalacturonase	Fusarium moniliforme
1ib4	Polygalacturonase	Aspergillus aculeatus
1idj	Pectin Lyase A	Aspergillus sp.
1jrg	Pectate Lyase A	Erwinia chrysanthemi
1jta	Pectate Lyase A (C2 Form)	Erwinia chrysanthemi
1k5c	Endopolygalacturonase I	Stereum purpureum
1ktw	Iota-Carrageenase	Alteromonas sp. Atcc43554
1nhc	Endopolygalacturonase I	Aspergillus niger
1ogm	Dex49a From	Penicillium minioluteum
1qcx	Pectin Lyase B	Aspergillus niger
1qjv	Pectin Methylesterase PEMA	Erwinia chrysanthemi
1rmg	Rhamnogalacturonase A	Aspergillus aculeatus

1rwr	Filamentous Hemagglutinin Secretion Domain	Bordetella pertussis
1tsp	Tailspike Protein	Bacteriophage P22
2pec	Pectate Lyase C	Erwinia chrysanthemi

Supplemental Table 3-S2. Nonredundant set of sequences from aggregative proteins, derived from [138].

1Cro	MQTLSERLKKRRIALKY
2Cro	YKMTQTELATKAGVK
3Cro	YKQQSIQLIEAGVTKR
4Cro	TKRPRFLYEIAMALNSD
5Cro	AMALNCDPVWLQYGTKRGKA
Acyl-phosphatase17-Jan	STAQSLKSVDYEVFGRV
Acyl-phosphatase18-33	QGVSFRMYTEDEARKI
Acyl-phosphatase34-53	GVVGWVKNTSKGTVTGQVQG
Acyl-phosphatase54-68	PEDKVNSMKSWLSKV
Acyl-phosphatase69-85	GSPSSRIDRTNFSNEKT
Acyl-phosphatase86-98	ISKLEYSNFSVRY
Ada-2HH1-Wt	VPSNEEQIKNLLQLEAQEHLQY
Ada-2HH2-WT	FVNVQAVKVFLESQGIAY
Alpha-synucleinNAC1-18	EQVTNVGGAVVTGVTAVA
Alpha-synucleinNAC1-18s	TVNGVGEVTATAVQGVAV
Alpha-synucleinNAC6-14	VGGAVVTGV
Amyloid-betaAB3	HQKLVFFAE
Amyloid-betaHABP5	KKPVFFAED

Amyloid-betaWhole	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
AraAra1	AVGKSNLLSRYARNEFSA
AraAra2	RFRAVTSAYYRGAVG
AraAra3	TRRTTFESVGRWLDELKIHSD
AraAra4	AVSVEEGKALAEEEGLF
AraAra5	STNVKTAFEMVILDIYNNV
Beta-microglobulinA	IQRTPKIQVYSRHPAE
Beta-microglobulinB	NGKSNFLNCYVSG
Beta-microglobulinC	FHPSDIEVDLLK
Beta-microglobulinD	NGERIEKVEHSDLSFSKD
Beta-microglobulinE1	DWSFYLLYYTEFTPTGKDEYA
Beta-microglobulinF	PTGKDEYACRVNHVT
BPTIP1-15	RPDFSLEPPYTGPSK
BPTIP16-28	ARIIRYFYNAKAG
BPTIP29-44	LSQTFVYGGSRAKRNN
BPTIP45-58	0FKSAEDSMRTSGGA
CheYCheY1	DFSTMRRIVRNLLKELGYN
CheYCheY2	EDGVDALNKLQAGGY
CheYCheY3	MDGLELLKTIRADSAY
CheYCheY4	AKKENIIAAAQAGASGY
CheYCheY5	PFTAATLEEKLNKIFEKLGMY
ComComA1	DHPAVMEGTKTILETDSNLS
ComComA2	EPSEQFIKQHDFSSY
ComComA3	VNGMELSKQILQENPH
ComComA4	EVEDYFEEAIRAGLH
ComComA5	TESKEKITQYIYHVLNGEIL

FlaxodoxinFXN1	GTGNTEKMAELIAKGIIESGKDY
FlaxodoxinFXN3	EESEFEPFIEEISTKISY
FlaxodoxinFXN4	GDGKWMRDFEQRMNGYGSV
FlaxodoxinFXN5	EPDEAEQDSIEFGKKIANIY
GlutexAlpha-4	DQKEAALVDMVNDGVEDLRCKYATLIYT
GlutexAlpha-5	YEAGKEKYVKELPEHLKPFETLLSQ
GlutexAlpha-6	QISFADYNLLDLLRIHQVLN
GlutexAlpha-7	PLLSAYVARLSA
GlutexAlpha-8	PKIKAFLA
myoglobinAB-Domain	VLSEGEWQLVLHVWAKVEA
myoglobinA-Helix	EGEWQLVLHVWAKVEADVAGHGQDILIRLFK
myoglobinBC-Turn	KSHPET
myoglobinB-Helix	DVAGHGQDILIRLFKS
myoglobinCCD-Domai	HPETLEKFDRFKHLK
myoglobinD-Helix	TEAEMKA
myoglobinEF-Turn	SEDLKKHGVTVLTALGAILK
myoglobinE-Helix	KKGHHEAE
myoglobinFG-Turn	ATKHKIP
myoglobinF-helix	ELKPLAQSHA
MyohemerithinAB-loop	Am-YEQLDEEHKKIFKGIFDCIRD
MyohemerithinA-helix	RDNSA
MyohemerithinBC-loop	DAAKYSEV
MyohemerithinB-helix	SAPNLATLVKVTTNHFTHEEAMMD
MyohemerithinCD-loop	GLSAPVD
MyohemerithinC-helix	EVVPHKKMHKDFLEKIGGL
MyohemerithinC-terminal	GTDFKYKGKL

MyohemerithinD-helix	AKNVDYCKEWLVNHIK
MyohemerithinN-terminal	GWEIPEPYVWDESFRVFY
PlastocyaninPc-1	LEVLLGSG
PlastocyaninPc-10	IPAGVDAVKISM
PlastocyaninPc-10a	EIPAGV
PlastocyaninPc-10b	DAVKIS
PlastocyaninPc-11	MPEEELL
PlastocyaninPc-12	MPEEELLNAPGETYVVTL
PlastocyaninPc-13b	APGET
PlastocyaninPc-14	GETYVVTL
PlastocyaninPc-14a	ETYVVT
PlastocyaninPc-15	VTLDTKGTY
PlastocyaninPc-16	GTYSFYT
PlastocyaninPc-16a	TYSFYC
PlastocyaninPc-18	MVGKVTVN
PlastocyaninPc-19	GTVSFVTSPHQGAGMVGKVTVN
PlastocyaninPc-2	LEVLLGSGDGSLVFV
PlastocyaninPc-3	SLVFVPSEFS
PlastocyaninPc-5	SEFSVPSGEK
PlastocyaninPc-6	KIVFKNNA
PlastocyaninPc-6a	GEKIVFKNNAGFPHNVVFDE
PlastocyaninPc-8	KNNAGFPHNV
PL-B1PL-B1-114-138-pH-2.4	KGTFEKATSEAYAYADTLKKDNGEY
PL-B1PL-B1-136-155D-pH-6.1	GEYTVDVADKGYTLNIKFAGD
PL-B1PL-B1-95-114-pH-4.1	VTIKANLIFANGFTQTAEFKG

ProteinGProteinG21-40	TYKLINGKTLKGETTTEA
ProteinGProteinG2-19	GDAATAEKVFKQYANDNGVD
ProteinGProteinG41-56	GEWTYDDATKTFTVTE
RasP21A	GVGKSALTIQLIQNHFVY
RasP21B	EYSAMRDQYMRTGEG
RasP21C	INNTKSFEDIHQYREQIKRVKDS
RasP21D	ARTVESRQAQDLARSYGIP
RasP21E	RQGVEDAFYTLVREIRQHK
Ribosome-L9Alpha-1	GYANNFLFKQG
Ribosome-L9Alpha-2	TPANLKALEAQKQKEQR
Ribosome-L9Beta-1	MKVIFLKDVKG
Ribosome-L9Beta-2	KGKKGEIKNVAD
Ribosome-L9Beta-3	LAIEATPA
Spectrin-SH3M-2	AYVKKLDSGTGKELVLAL
Spectrin-SH3M-4	YDYQEKSPREVTMKKGD
Spectrin-SH3M-68	DILTLLNSTNKDWWKVEVNDRQGFVPA
Spectrin-SH3M-C	GGKDWWKVGG
t-ProteinK19	PGGGKVQIVYKPV
t-ProteinK19d	PGGGKVYKPV
t-ProteinK19Gluc4	QTAPVPMPDLKNVKSKIGSTE
t-ProteinK19Gluc41	NLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE
t-ProteinK19Gluc42	VKSE
t-ProteinK19Gluc78	QTAPVPMPD
t-ProteinV313-K321	VDLSKVTSK
t-ProteinV335-E342	GQVEVSKE