

Ensemble Modeling of β -sheet Proteins

by

Charles William O'Donnell

Bachelor of Science, Columbia University, May 2003

Master of Science, Massachusetts Institute of Technology, September 2005

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

© Massachusetts Institute of Technology 2011. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
May 20, 2011

Certified by
Srinivas Devadas
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by
Bonnie Berger
Professor of Applied Mathematics & Electrical Engineering and Computer Science
Thesis Supervisor

Certified by
Susan Lindquist
Professor of Biology
Thesis Supervisor

Accepted by
Professor Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Ensemble Modeling of β -sheet Proteins

by

Charles William O'Donnell

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Our ability to characterize protein structure and dynamics is vastly outpaced by the speed of modern genetic sequencing, creating a growing divide between our knowledge of biological sequence and structure. Structural modeling algorithms offer the hope to bridge this gap through computational exploration of the sequence determinants of structure diversity.

In this thesis, we introduce new algorithms that enable the efficient modeling of protein structure ensembles and their sequence variants. These statistical mechanics-based constructions enable the identification of all energetically likely sequence/structure states for a family of proteins. Beyond improved structure predictions, this approach enables a framework for thermodynamically-driven mutational and comparative analysis as well as the approximation of kinetic protein folding pathways.

We have applied these techniques to two protein types that are notoriously difficult to characterize biochemically: transmembrane β -barrel proteins and amyloid fibrils. For these we advance the state-of-the-art in structure prediction, mutational analysis, and sequence alignment. Further, we have collaborated to apply these methods to open scientific questions about amyloid fibrils and bacterial biofilms.

Thesis Supervisor: Srinivas Devadas
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics & Electrical Engineering and Computer Science

Thesis Supervisor: Susan Lindquist
Title: Professor of Biology

Acknowledgments

It has been my absolute privilege to work with a great cast of advisors in Srinu Devadas, Bonnie Berger, and Sue Lindquist. I have learned immeasurably from them, and I thank them all for being exceptional mentors both professionally and personally.

Srinu has been my advisor since the start of graduate school and provided me with an outstanding research experience, as well exceptional guidance any time it was needed. His inexhaustible energy is inspiring, and he has freely encouraged me to pursue my own interests, such as computational biology, for all of which I am thankful. Bonnie welcomingly took me under her fold as I entered this field, and her endless enthusiasm and support has much to do with any success in my research. Moreover, I thank her for the countless ideas and thoughtful observations that have guided my work, as well as her insightful advice. Sue further welcomed me into her lab as I began this thesis work, and as a computational minority made me feel at home. I am thankful for her support and education on how to frame the important biological questions, as well as her encouragement to collaborate with her amazing lab and even to investigate bench work.

I would also like to thank my close collaborator throughout this entire thesis work, Jérôme Waldspühl, who co-developed many of these ideas, shared stimulating conversations, and is a friend. Further, I could not have achieved this work without my co-authors and co-workers Mieszko Lis, Solomon Shenker, and Sebastian Will, whose computational expertise impresses, and Randal Halfmann, Ram Krishnan, and Kendra Frederick, whose biological acumen has taught me so much. Credit also belongs to Rolf Backofen, Jessica Goodman, Tim Lu, and Peter Clote who have also been instrumental in my research. I am grateful to Collin Stultz for kindly providing insight and perspective into my work as my only non-advisor committee member.

During my time at MIT I have been surrounded by amazing colleagues and friends who have made my time here a joy. I am especially thankful to my officemates over the years. Daihyun, Prabhat, Nirav, Michael P., Karen, Ryan, Luis, Jérôme, Omer, and Chris, have been endless source of help, invigorating conversation, and good humor. Arvind has always warmly included me in his research group, and Joel Emer, Sally, Patrice, Brooke, Bob, and Karen have always been there to chat. Further, I have been constantly encouraged, impressed, and entertained by my peers, in particular Ed, Blaise, Elliot, Myron, Jae, Ian, Dave, Vijay, Michael K., Michel, Michael Z., Michael G., Daniel, Steve, Abhinav, Alfred, Muri, Michael B., Michael S.-L., Nathan, Patrick, Rohit, Raghu, Po-ruh, Sebastian, Oliver, Simon, Alex, Sandro, Jijun, Pete, Sven, Dan, Walker, Marc, Lauren,

Chris, and many others.

Above all, I thank my family for their unconditional love and support throughout the years. My parents, grandparents, brothers, and extended family have all been there through good and bad, and my wonderful children Charlotte and Henry make each new day fun. Most importantly, my wife Adrienne Lavidor-Berman has made me happier than words can express, and I look forward to her company in all that will come.

Contents

1	Introduction	14
1.1	Protein structure prediction	15
1.1.1	Computational modeling	16
1.1.2	Brief overview of sequence→structure modeling approaches	17
1.1.3	Ensemble modeling	19
1.1.4	Reconfigurable ensembles	21
1.2	Biological background	22
1.2.1	Transmembrane β -barrel proteins	23
1.2.2	Amyloids	23
2	Ensemble-based protein structure prediction	26
2.1	Goals and Overview	26
2.2	Transmembrane β -barrel modeling (Bottom-up approach)	28
2.2.1	Existing computational predictors for TMBs	29
2.2.2	Representing ensemble space via attribute grammars	29
2.2.3	Computing the partition function	32
2.2.4	Boltzmann distribution sampling	36
2.2.5	Stochastic contact map and residue contact probability	37
2.2.6	Runtime optimizations	39
2.2.7	3-dimensional model generation	40
2.3	Amyloid fibril modeling (Top-down approach)	41
2.3.1	Existing computational predictors for Amyloids	43
2.3.2	Representing ensemble space via recursive primitives	44
2.3.3	Computing the partition function	46

2.3.4	Boltzmann distribution sampling	48
2.3.5	Stochastic contact map and residue pairing probabilities	49
2.3.6	Runtime optimizations	50
2.3.7	Runtime heuristics	52
2.3.8	3-dimensional model generation	53
2.4	Energetic models	54
2.4.1	Conditioned pairwise amino acid interactions	55
2.4.2	Stacking-pair amino acid interactions	56
2.5	Schema comparison and prediction normalization	57
2.5.1	Stochastic contact map normalization	57
2.5.2	Schema comparison	57
3	Evaluation of ensemble structure predictors	59
3.1	Transmembrane β -barrels	59
3.1.1	Residue/residue β -strand contact prediction	60
3.1.2	Residue flexibility prediction via contact probability profile	65
3.1.3	Whole structure prediction via Boltzmann sampling	66
3.2	Amyloid fibrils	68
3.2.1	Validation of amyloid fibril predictions	69
3.2.2	Case analyses of well characterized amyloid proteins	71
3.3	Evaluation of energetic stacking pair potentials	80
4	Mutational landscape analysis	82
4.1	Goals and overview	82
4.2	Amyloid fibril modeling	84
4.2.1	Representing mutational landscapes	84
4.2.2	Computing the partition function and Boltzmann distribution sampling	85
4.2.3	Runtime optimizations	86
5	Evaluation of mutational landscape prediction	88
5.1	Identifying conformational shifts in amyloids	89
5.1.1	A β Iowa mutant	89
5.1.2	HET-s yeast-toxic mutants	91

5.2	Validation of predicted amyloidogenicity	93
5.2.1	Mutational occupancy:	93
5.2.2	HET-s/HET-S:	94
5.2.3	A β single-point proline mutagenesis:	94
5.2.4	A β multiple-residue mutagenesis:	95
5.3	Identification of amyloidogenicity bias of <i>Asn</i> over <i>Gln</i> in HET-s	95
5.4	Investigation of <i>E. coli</i> curli and biofilm inhibition	98
6	Simultaneous alignment and folding: consensus prediction	102
6.1	Goals and Overview	102
6.2	Transmembrane β -barrel consensus modeling	104
6.2.1	Representing consensus structure ensembles	105
6.2.2	Computing the partition function tables	107
6.3	Amyloid fibril consensus modeling	113
7	Evaluation of simultaneous alignment and folding	114
7.1	Transmembrane β -barrel consensus modeling	114
7.1.1	Dataset and evaluation technique	115
7.1.2	Model parameter selection	117
7.1.3	Validation of alignment accuracy under low sequence identity	118
7.1.4	Secondary structure prediction accuracy of consensus folds	120
7.2	Amyloid fibrils consensus modeling	123
8	Ensemble prediction of folding dynamics	125
8.1	Goals and overview	125
8.2	Modeling single β -sheet protein dynamics	127
8.2.1	Representing permutable β -sheet ensembles	127
8.2.2	Computing the partition function	128
8.2.3	Boltzmann distribution sampling	129
8.2.4	Predicting dynamics using the Fokker-Planck equation	130
8.3	Evaluation of single β -sheet protein dynamics prediction	132
8.3.1	Validation of super-secondary structure prediction accuracy	133
8.3.2	Case analysis of folding pathway prediction of the B1 domain of Protein G	136

8.3.3	tFolder running time	137
9	Web-based ensemble prediction tools	138
9.1	partiFold	138
9.1.1	Input	139
9.1.2	Output	140
9.2	AmyloidMutants	140
9.2.1	Input	141
9.2.2	Output	142
9.3	partiFoldAlign	143
9.4	tFolder	143
9.4.1	Input	143
9.4.2	Output	144
10	Conclusion	145
10.1	Summary	145
10.2	Future research	146
10.2.1	Energy model improvements	147
10.2.2	Constraint-based schema design	147
10.2.3	α/β protein schemas	148
10.2.4	Autotransporters	148

List of Figures

1-1	Classical versus ensemble modeling	21
1-2	Integration of computational modeling with experimental study	22
1-3	3-dimensional rendering of TMB protein PagP	23
1-4	3-dimensional rendering of the HET-s amyloid fibril structure	24
2-1	Structure decomposition of transmembrane β -barrels	32
2-2	2-tape representation of transmembrane β -barrels	33
2-3	2-tape representation of transmembrane β -barrels: extension and shear	34
2-4	2-tape decomposition of transmembrane β -barrel	35
2-5	Transmembrane β -barrel sampling procedure	37
2-6	Illustration of 3-dimensional TMB predictions	41
2-7	Amyloid fibril schemas used for analysis	42
2-8	The “kink” schema features allows more efficient β -helical modeling	43
2-9	Simplified recursion definition for schema \mathcal{A}	47
2-10	Example M-rule extension for “slip” (schema \mathcal{A} or \mathcal{P})	48
2-11	Anti-parallel β -strand “stacking pairs”	56
2-12	Illustration of null-hypothesis contact map	58
3-1	Illustrative representations of stochastic contact predictions	61
3-2	Metrics for comparison of ensemble predictions to X-ray crystal data	62
3-3	3-dimensional graphics of known TMB structures used in validation	63
3-4	F-measure accuracy scores of partiFold compared with BETApro	65
3-5	partiFold residue flexibility prediction accuracy compared with PROFBval	67
3-6	Coverage and accuracy of clustered partiFold predictions	68
3-7	Secondary structure prediction accuracy of AmyloidMutants and others	70

3-8	AmyloidMutants A β predictions compared with experimental data	72
3-9	AmyloidMutants HET-s predictions compared with experimental data	74
3-10	AmyloidMutants FgHET-s predictions compared with experimental data	76
3-11	AmyloidMutants amylin predictions compared with experimental data	77
3-12	AmyloidMutants α -syn predictions compared with experimental data	78
3-13	AmyloidMutants tau predictions compared with experimental data	80
3-14	The effects of reduced alphabet selection on TMB accuracy	81
4-1	Mutational landscapes add sequence dimension to structure ensembles	85
4-2	Illustration of relationship between sequence/structure ensemble samples	86
4-3	Example mutational landscape state expansion, <i>M-rule</i> from schema \mathcal{A}	87
5-1	AmyloidMutants ensemble predictions of A β_{1-40} and A $\beta_{1-40}/D23N$	90
5-2	AmyloidMutants predictions of HET-s and yeast-toxic HET-s mutants	93
5-3	A β_{40} scanning mutagenesis predictions compared with experimental data	95
5-4	AmyloidMutants amyloidogenicity predictions of multiple-residue A β mutants	96
5-5	HET-s/4N \rightarrow Q is defective for amyloid assembly	97
5-6	Predicted contact maps highlight differences in WT HET-s and HET-s/4N \rightarrow Q	97
5-7	Ensemble schema design for heterogeneous CsgA/CsgB amyloid fibril	99
5-8	AmyloidMutants predicted sequence regions for CsgA/CsgB interaction	99
5-9	CsgB peptide array seeds CsgA amyloid fibrils at two sequence positions	100
5-10	Reduction in curli amyloid formation using phage-display peptide library	100
6-1	Elements of a TMB sequence/structure alignment	105
6-2	Decomposition strategy for TMB alignment	105
7-1	Alignment depends on a balance between sequential and structural information	118
7-2	partiFoldAlign alignment accuracy for 8-, 10-, and 12-stranded TMBs	120
7-3	partiFoldAlign alignment accuracy for individual TMBs (transmembrane only)	121
7-4	partiFoldAlign alignment accuracy for individual TMBs (whole protein)	122
8-1	Permutable β -sheet schemas encoding using signed permutation	128
8-2	Decomposition of recursion for permutable β -sheet schemas	129
8-3	Indices of intermediate permutable β -sheet structures	130

8-4	Illustration of permutable β -sheet schemas sampling	130
8-5	Illustration of permutable β -sheet schema compatibility	131
8-6	Illustration of tFolder Protein G ensemble predictions	133
8-7	Predicted folding pathways of B1 domain of Protein G	135
8-8	tFolder running time performance	136
9-1	Screenshot of partiFold online tool	139
9-2	Screenshot of AmyloidMutants online tool	141
9-3	Screenshot of tFolder online tool	144

List of Tables

3.1	TMB grammar constraints used for validation	63
3.2	Summary of amyloid fibril secondary-structure prediction results	69
5.1	AmyloidMutants $A\beta_{1-40}/A\beta_{1-40}/D23N$ predictions suggest conformational switch . .	90
5.2	AmyloidMutants predictions distinguish HET-s m8 mutant as unique	91
7.1	Breakdown of OPM database TMB pairwise alignments	116
7.2	TMB structural constraints used for consensus folding	117
7.3	partiFoldAlign secondary structure assignment accuracy	123
8.1	tFolder super-secondary structure contact prediction performance	134
8.2	Comparison of predictive performance of tFolder, BETApro, and SVMcon	134

Chapter 1

Introduction

Proteins form the essential machinery of life, participating in nearly all cellular processes through the physical interaction of their chemically-unique 3-dimensional structures. Although DNA encodes the genetic information necessary for cell function, genes must be transcribed into RNA and translated into proteins to carry out their purpose (a nontrivial and error-prone step). Thus the ability of proteins to adopt useful molecular structures ultimately constrains genetic variation and mutations, and guides evolutionary change. Discovering how, when, and why these structures form and interact is a fundamental question in biochemistry and molecular biology.

Unfortunately, classical experimental techniques for determining protein structure and protein interaction mechanisms can be indirect, lengthy, or plagued by inconsistency. Indeed, our ability to determine the genetic makeup of an organism through high-throughput sequencing has outpaced structural characterization techniques by orders of magnitude. As a result, our mechanistic understanding of which genes control function, and how perturbations to cellular processes can alter phenotype depend largely on genomic information alone. While this form of modeling can provide considerable insight, it overlooks potential biophysical constraints that can play subtle but important roles. Therefore one of the great problems in computational biology has remained how to construct a meaningful biophysical model of protein structure and interaction based on the ready availability of genomic and other similar kinds of data.

To tackle this problem we propose an algorithmic framework for modeling the complex relationship between genetic information and protein structure based on two principles: (1) structure predictions should describe the entire “ensemble” of potential conformational variants, as is seen *in vivo*, and (2) the predictive model must be flexible and configurable for different biological do-

mains, allowing the incorporation of existing experimental knowledge. Our framework allows the specification of different protein structure spaces and efficiently calculates the probability of observing individual states according to a thermodynamically-inspired scoring system. Using this general approach, we construct algorithms that advance the state-of-the-art in three major computational biology problems:

Chapters 2 & 3: The accurate prediction of multiple protein structure states.

Chapters 4 & 5: The identification of key sequence mutants controlling structural variation.

Chapters 6 & 7: The simultaneous alignment of sequence and prediction of structure.

In these three domains we have applied our technique to advanced our scientific understanding of two important protein families: transmembrane β -barrel (TMB) proteins and amyloids. Further, **Chapter 8** demonstrates how ensemble methods can serve as a foundation for the efficient prediction of β -sheet folding pathway kinetics. Finally, **Chapter 9** describes the implementation of these prediction algorithms as online web-based tools, and **Chapter 10** concludes. We have previously published contributions toward this thesis [133, 169, 197, 198].

In this chapter we describe prior successes and shortcomings in protein structure modeling research, the challenges faced today, and how our ensemble modeling approach can mitigate some of these problems with acceptable tradeoffs. We also provide a brief biological overview of the importance, structure, and function of the two protein families studied in this thesis: transmembrane β -barrels and amyloid fibrils.

1.1 Protein structure prediction

Proteins are comprised of one or more linear polypeptides which, after translation by a ribosome, typically fold into compact functional macromolecules. Foundational work has established that for most proteins the amino acid chain contains sufficient information to enable the proper formation of functional conformational states [6]. While the predominant functional conformation found in the cell is sometimes referred to as the native state, many proteins can still adopt distinct structural variants dependent on stress conditions, ligand binding, protein/protein interactions, chaperones, solvents, compartmentalization, and so forth. The difference between these conformational states can be a matter of just a few Ås, such as in protein breathing [22], or dramatic, such as in the hinge-like transformation of calmodulin upon calcium binding or the complete rearrangement of

structure undergone by the prion protein PrP. Furthermore, the cell is a crowded, dynamic place, and not all peptides reach a functionally useful state — either because of misfolding, aggregation, or improper trafficking. Moreover, estimates have suggested that up to 30% of newly minted proteins are rapidly degraded due to defective translation [58]. As a result, all organisms have evolved quality control machinery that mitigates such problems through protein refolding, degradation, and other mechanisms.

Due to the importance of understanding protein structure and its consequences on function, a tremendous amount of research has focused on developing accurate experimental techniques for determining protein conformation *in vivo* and *in vitro*. Of these, the most widely-used high-resolution methods are X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, which can directly observe protein structure *in vitro* with atomistic precision. Although end results of these experiments can be very useful due to their great detail, the application of X-ray crystallography and NMR can be extremely difficult (such as when studying non-soluble proteins), and their methodology can introduce chemical conditions that differ dramatically from those *in vivo*. Further, the human effort required to conduct such experiments on any given protein, protein mutant, or ligated protein can be sizable and impractical for genomic-level study. In fact, at the time of this writing only ~42,000 unique protein structures have been solved (and deposited in the PDB [14]), while ~16 million potential protein-encoding sequences have been identified in genomic data (according to UniProt [43] annotations) — an orders of magnitude difference. Thus, often protein structure and structural interaction discovery must be accomplished by independently integrating a large number of easier to execute, lower precision experiments with computational models to arrive at a structural hypothesis.

1.1.1 Computational modeling

Research into computational protein modeling has been conducted for nearly four decades, focusing on countless problems and techniques. However, historically, one of the more useful problems computational methods can be applied to is how to predict some representation of the native protein structure based only on the polypeptide’s amino acid residue sequence (the “sequence→structure” problem). This is particularly helpful since at the beginning of most experimental investigations all that is known about a protein is its amino acid sequence (inferred from genomic data). Further, this problem can often be formulated by a clean, elegant, mathematical definition with readily available inputs (amino acid sequences) and an ostensibly verifiable output (protein structures). Al-

though countless techniques for addressing this problem have been proposed, we highlight here two general choices that must be made for nearly any protein sequence→structure modeling problem: how to represent a protein, and what assumptions to include in the model.

The first and most crucial decision when designing a protein modeling algorithm is that of how protein structures will be represented. For example, given a sequence, one can describe a detailed 3-dimensional representation that assigns relative x,y,z coordinates to every atom in the polypeptide chain, such as done by tools like Rosetta [155] or RAPTOR [214]. Alternately, a 1-dimensional, secondary structure representation can be chosen that describes information like “residues 22–30 form β -strand” (as in the tool SSPro [145]), or a 2-dimensional, super-secondary structure representation that says “residues 22–30 form a parallel β -sheet interaction with residues 46–54” (as in the tool BETApro [31]) 1-, 2-, and 3-dimensional approaches can all provide valuable information to further experimental analysis, however, these representational granularities require drastically different trade-offs to computational complexity, accuracy, and generality. That is to say, a 3-dimensional output typically provides more informative answers than a 2-dimensional representation, but can require days, weeks, or months to predict. Oppositely, 2-dimensional predictions offer less atomistic insight but may provide sufficient insight are quick and efficient to compute. This thesis adopts an approach more in line with the latter.

The second major choice is that of how much, and what type of *a priori* information will be built into the model itself. This defines the assumptions that are going to be made about a biological system and can directly effect both the accuracy of predictions, and the generality of the model in other biological applications. For example, if absolutely nothing is known about a protein, a first principles approach may be appealing, relying only on the laws of physics and thermodynamics, and removing human bias from the equation. However, calculating predictions like this can be unrealistically time consuming for even average sized systems. Alternately, if considerable information is known about a protein, such as the structure of a protein’s known homolog, algorithms can integrate this data to vastly improve the accuracy and utility of a model. However, such a choice restricts the applicability of the technique, perhaps making it useless for other analyses. In this thesis we take a middle ground between these two extremes, allowing specific *a priori* knowledge to be easily integrated into a more generic thermodynamic framework.

1.1.2 Brief overview of sequence→structure modeling approaches

Here we describe a high-level categorization of approaches for sequence→structure prediction that cover many (but not all) of the countless published techniques. These mirror the three predictor evaluation categories historically employed by the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competitions [78] (although newer CASP competitions have altered their categorization approach). Common to all categories, however, is one of two assumptions (or both): the thermodynamic hypothesis that the native state is at the global free energy minimum, or the use of probabilistic inference conditioned by known biological observations. A great number of resources exist which cover these topics in much more depth [18,40,87,128,166].

Homology modeling

Homology modeling techniques are based on the evolutionary principle that proteins of similar function tend to have similar structure, and that proteins of similar structure tend to have similar sequence. Therefore, for uncharacterized proteins whose primary amino acid sequence is similar to another characterized protein structure, computational homology modeling techniques can be extremely accurate. Typically, this distinction is drawn for sequences which are roughly 30%-40% sequence similar or greater.

Naively, a homology modeling technique uses the 3-dimensional structure of a known homologous protein to serve as an atomistically-detailed template. The relative x,y,z coordinates of the polypeptide backbone atoms are kept fixed, along with the sidechains of residues shared between sequences. The atomic coordinates of residues that differ between sequences are then added using energy minimization techniques or other probabilistic approaches. In practice, more complicated methods are employed, such as by the tool MODELLER [70].

***De novo* structure prediction**

For proteins with no known homolog, or proteins which remain disordered under physiological conditions [59], *de novo* or *ab initio* prediction methods can be used to compute putative structures. For example, molecular dynamics (MD) simulations make use of energetic force fields to simulate the physical folding of polypeptide chains in atomistic detail over time [108]. Unfortunately, the heavy computational load required by these techniques can limit their application to very short sequences. Recent developments in distributed computing [190] and hardware accelerated algorithms [168] have made great progress in this domain, enhancing our ability to answer questions about topics such as ligand docking or protein/protein interactions. However, their application to *ab*

initio sequence→structure prediction is still impractically resource intensive for most cases.

Fold recognition

Fold recognition algorithms, such as those using generic “threading” techniques, offer a conceptual middle ground between homology modeling and *de novo* prediction. Based on the idea that most proteins are composed from a finite set of recurrent, substructural “folds,” these methods detect protein homology at a much finer granularity than whole protein structures.

Naively, threading methods generate profiles from known 3-dimensional *fold* information, and then “thread” a sequence of unknown structure across these profiles. Sequence/profile scores are computed, and a composite structure is constructed, resolving potential gaps or overlaps in substructure states. For specific cases, this approach has demonstrated high accuracy, such by the tool RAPTOR [214]. The approach taken in this thesis can be best compared with this category of sequence→structure predictors, as it uses abstract definitions of folds (Chapter 2), however, significant algorithmic differences and assumptions exist. Furthermore, our framework can be generalized beyond the problem of protein sequence→structure prediction.

Combined pipelines

Finally, it is important to emphasize that the most accurate sequence→structure prediction tools integrate ideas from all three categories into a unified (typically probabilistic) model. A particularly successful example of this is the Rosetta approach based on peptide fragment-assembly [4,21,150].

1.1.3 Ensemble modeling

In this thesis, we introduce a computational modeling framework, named “*ensemble modeling*,” that combines features from many of the methodologies listed above. This approach makes use of both thermodynamics and fold similarities between homologs to simultaneously model the space of *all* potential protein structures within a predefined landscape (detailed in Chapter 2).¹ Rather than focusing on single protein structures, such ensembles are designed to model the ability of a protein to adopt different conformational states *in vivo*. Further, this principle can be extended to not only model structure, but to study mutational sequence/structure space and comparative sequence alignments, enabling an exploration of the sequence determinants of structural heterogeneity (introduced

¹We note that multiple, conflicting definitions of “ensembles” have been used in the literature, such as for voting-based schemes. Throughout this thesis we refer to ensemble models in only the context described in this section.

in Chapters 4 and 6).

Early theories of protein structure envisioned bodies of rigidly packed polypeptides [42], suggesting a singular mapping between a protein's amino acid sequence and its 3-dimensional structure. Moreover, present day databases have accumulated and organized data to seemingly support this perspective (e.g., the PDB [14]). However, in reality this relationship is far more complex [11]. By the 1980's significant evidence supported the idea that the functional native state was not fixed in stone, but that multiple substate minima could exist with different functional properties [45,63,74,154]. For example, prions can have multiple distinct, phenotypically-stable states [1] and disordered proteins can lack stable tertiary interactions altogether [187]. Further, it has been shown that during the process of folding, barriers, such as those imposed by solvent [189], could prevent a polypeptide from reaching a singular free-energy minimum fold and meta-stable intermediate structures may persist [157,163].

Despite these observations, many “classical” sequence→structure computational prediction tools have adhered to a single-sequence/single-structure model. Historical reasons for this are many. For example, single-sequence/single-structure models can be algorithmically simpler and more efficient, and can be easily validated against experimentally determined conformations found in the PDB. Further, such predictions can be thought to represent biological averages that provide *sufficient* intuition for most investigations. However, the ensemble framework described in this thesis is a member of a newer generation of computational modeling tools that seek to describe a more realistic landscape of conformational variants without sacrificing efficiency or accuracy (Figure 1-1).

Our approach views protein cellular states from a statistical mechanics perspective. According to statistical mechanics theory, molecular state is in constant flux when at equilibrium, but the proportion of molecules in each specific state remains constant, allowing one to quantify the makeup of a system. Originally conceived for modeling the behavior of gas [28], the general theory has been applied to other areas of computational biology, including lattice and non-lattice models of RNA secondary structure [54,121], transcription factor binding sites [13,71,126,194], nucleosomes [33], helix bundles [34], and globular proteins [66,67,94,123,139].

To enable a statistical mechanics approach, we use coarse, high-level protein representations to allow the Boltzmann partition function to be efficiently computed over *all* potential states. The particular choice of representation is key as it has been shown that calculating the energy of all ensemble states in 3-dimensions is computationally intractable [93]. Using the partition function, we analyze the significance of all protein states in the system and the likelihood of their occurrence.

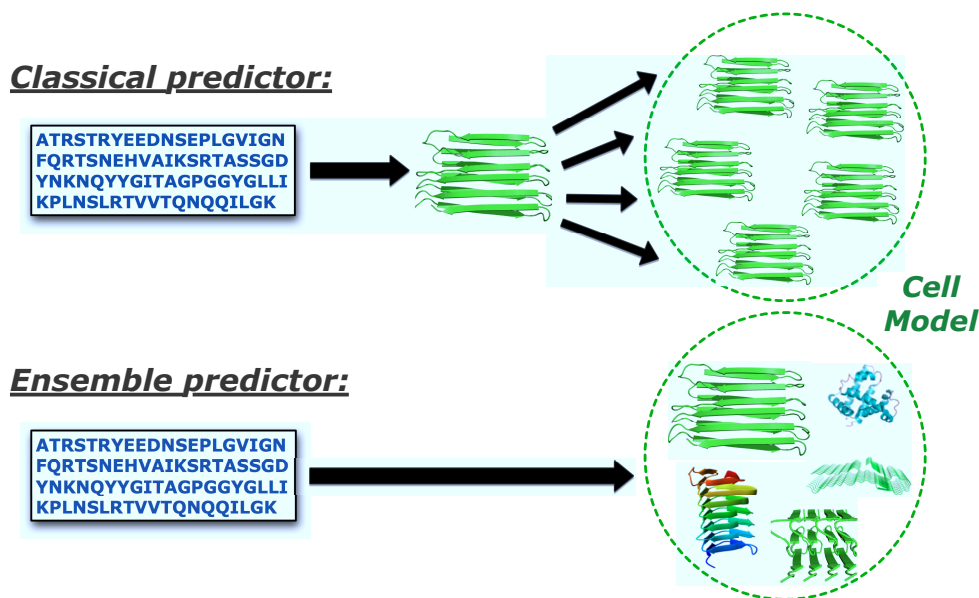


Figure 1-1: **Classical versus ensemble modeling:** Illustration depicts the difference between classical, single-structure modeling techniques and an ensemble approach. The former suggests a cell model populated by perfectly identical protein duplicates, while the latter enables a cell model populated by protein conformational variants (be they slight or dramatic differences).

Section 2.1 further details our ensemble approach.

1.1.4 Reconfigurable ensembles

The classification of prediction techniques detailed in Section 1.1.2 may be thought of as a gradient on the amount of *a priori* knowledge that is initially built into a model. Naturally, *de novo* methods (which only use protein sequence) are often less accurate than homology modeling techniques (which leverage powerful evolutionary tendencies). However, given the manner in which many biological investigations progress, targeting a computational tool for only one point along this spectrum seems inappropriate. After all, while it may be typical to begin a study with only the knowledge of protein sequence, as experimental data is gathered, more and more non-sequential information will come to light. Furthermore, published literature may often describe effects from related studies that provide very useful insight into the problem at hand.

Therefore, an ideal computational modeling tool would allow arbitrary investigator knowledge to be initially incorporated, and would grow to account for new observations — rather than serve a “one-time-use” role. Although a true ideal may be impossible, this thesis proposes an integrative approach for our ensemble framework that allows the successive incorporation of existing experi-

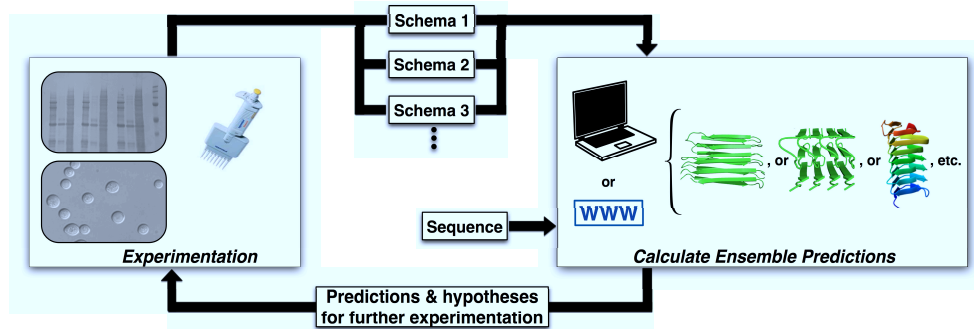


Figure 1-2: **Integration of computational modeling with experimental study:** Our proposed work flow enables integration between experimentation and predictive modeling. In an iterative fashion, experimental observations are used to modify ensemble space (via schemas, see Chapter 2), while predictions guide and suggest further experimental study.

mental (or even hypothetical) knowledge (Figure 1-2). This is implemented via constraints imposed onto an existing ensemble definition that are designed to incorporate specific experimental observations — for example, the effects of point mutations, known distance relationships from fluorescence microscopy or NMR, or structural parameters from other tools such as mass spectrometry or small-angle scattering. Each constraint refines the ensemble space being probed. With this, an ensemble predictor can be used as a rapid prototyping mechanism or hypothesis generator to iteratively refine experimental inquiry and to guide future assays or analyses. Section 2.2.2 and 2.3.2 provide further details.

We note that the integration of experimental data within a computational predictor is not novel unto itself — biological assumptions are required for any model. However, our approach is distinguished by the ease with which new information can be added to a schema (Section 2.3.2), without requiring core algorithmic re-implementation. Further, although machine learning techniques can achieve a slightly similar goal, as they can be re-trained on new data at any point, such approaches can be opaque when attempting to identify the specific contribution of various inputs, hindering hypothesis refinement.

1.2 Biological background

Our goal is to both develop computational biology algorithms for broad use and to actively advance our scientific understanding of important open problems. Therefore, we have applied our approach to further the study of two important protein families: transmembrane β -barrel proteins and amy-

loids. Both families present intriguing functions and phenotypes, and are in particular need for computational analysis as they are both relatively poorly characterized (compared with many other globular proteins). Here we introduce these families of proteins along with their relevant biological purpose and properties.

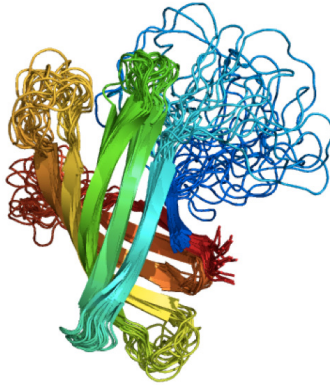


Figure 1-3: **3-dimensional rendering of TMB protein PagP:** Multiple NMR solutions [91] for the transmembrane spanning region of PagP have been overlaid and rendered using PyMOL [149].

1.2.1 Transmembrane β -barrel proteins

Transmembrane β -barrels (TMBs) constitute an important class of proteins typically found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. These proteins display a wide variety of functions and are relevant to various aspects of cell metabolism. In particular, outer-membrane proteins (OMPs) are used in active ion transport, passive nutrient intake, membrane anchors, membrane-bound enzymes, and defense against membrane-attack proteins. It has also been suggested [164] that some outer membrane proteins may be stress-response proteins, produced in abundance by bacteria in a minimum inhibitory concentration of antibiotics.

Since OMPs were discovered relatively recently and are difficult to crystallize, there are currently only about one hundred TMBs in the Protein Data Bank, and only 19 after the removal of homologous sequences. Some *in vitro* and *in vivo* mutation studies of OMPs [103, 210] have been performed, but compared with the overwhelming amount of data on globular proteins, outer membrane proteins remain a biologically important but technically difficult area of research.

1.2.2 Amyloids

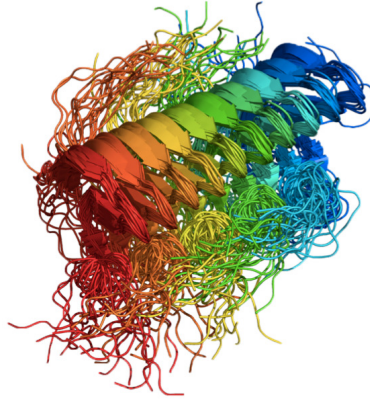


Figure 1-4: **3-dimensional rendering of the HET-s amyloid fibril structure:** Multiple NMR solutions of the HET-s amyloid fibril structure [205] have been overlaid and rendered using PyMOL [149]. Shown is a five peptide section of the fibril.

Under optimum conditions, proteins with diverse primary sequence exhibit the ability to self-assemble into structurally-varied, but highly-ordered β -sheet aggregates known as amyloid fibrils [57]. Those forming amyloid under normal physiological conditions can have profound effects on biological systems — deleterious and beneficial. On the one hand, amyloids play a role in diseases such as Alzheimer's, Parkinson's, and Huntington's, as well as systemic amyloidosis [36]. On the other, they serve vital functions in normal biology such as in human peptide hormone storage [116], biofilm formation [10], and a mechanism of protein-only inheritance by yeast prions [73]. However, the generic nature of the fold, the observation that most proteins do not form amyloid under normal conditions, and the ability of many amyloids to adopt multiple amyloid structures from the same peptide sequence [136,141] (structural strains) confounds standard sequence-specific models of protein folding. Moreover, sequences with only a small likelihood of forming amyloid can remain so given many mutations, or become abundantly amyloidogenic after only a single point change [110]. Therefore, to better understand the sequence/structure relationship of amyloid fibrils, a meaningful predictive model is required that describes the relationship between a given sequence and its mutational neighborhood.

Countless experimental studies have been performed to probe the molecular mechanism of these enigmatic structures. However, most methods (developed primarily for globular proteins) are difficult to apply to amyloids due to their large size and insolubility. Techniques such as solid-state NMR and H/D-exchange have brought us the most information about fibril structure, but only through exhaustive work and complex experimental design [114, 115, 129, 188, 205]. The high-cost

of such studies has prevented the kinds of large-scale investigations that can reveal the underlying sequence/structure relationships in functional and pathogenic amyloid folds.

Chapter 2

Ensemble-based protein structure prediction

2.1 Goals and Overview

Our work introduces a statistical mechanics-based approach for predicting “ensembles” of protein structures from sequence alone. In an ensemble predictor, each peptide sequence is presumed to fold into a complete set of millions (or billions) of unique structural states, with a single energetic value calculated for each state according to its entire conformation. From this quantified set of all possible structures, clusters of low-energy states with similar conformations can be extracted via sampling and analyzed.¹ Such an ensemble predictor differs from “classical” techniques for protein structure prediction, which typically perform an algorithmic search for an individual, lowest energy structure (Section 1.1.3).

We take a statistical mechanical approach, and model protein conformations as states within a canonical ensemble, with each state’s corresponding energetic value assigned according to a Boltzmann distribution. From this predicted ensemble of structures and their energetic scores, one can identify clusters of highly likely structural states by their relative energetics. This approach is inspired by work in the RNA modeling field that demonstrated efficient algorithms for recursively calculating the Boltzmann partition function of RNA secondary structure [121], enabling rigorous sampling of the RNA folding landscape [54]. However, these approaches cannot be directly applied to proteins as their model of RNA structure is too simplistic and restrictive for protein interactions.

¹Note, “ensemble” predictors differ from consensus predictors; the latter produces a single prediction based on the consensus of multiple authors’ algorithms.

Thus new algorithmic techniques are required to calculate the Boltzmann partition function and related quantities for proteins.

Formally, the Boltzmann partition function \mathcal{Z} can be calculated over all structural states $s = \{1 \dots n\}$ such that

$$\mathcal{Z} = \sum_{i=1}^n e^{-\frac{E_{s_i}}{RT}}.$$

To compute this value efficiently, we must decompose a protein’s energy into independent sub-structure energy scores (Section 2.4). The energy of each structural state s is defined to be $E_s = -RT \log(p_s) - RT \log(\mathcal{Z})$, and we make the assumption that E_s can be linearly decomposed into i parts such that $E_s = \sum_i -RT \log(p_{s_i}) - RT \log(\mathcal{Z})$ [40, 170]. The probability p_{s_i} thus represents the likelihood of observing a substructural state i , given the temperature T , the Boltzmann constant R , and the statistical centering constant $\log(\mathcal{Z})$. Note, however, the energy model used in this work is not parameterized by temperature, so RT can be set to 1.0 in practice. The relative likelihood of any complete structure s is thus

$$p(s_i) = \frac{e^{-E_{s_i}/RT}}{\mathcal{Z}}.$$

The definition of protein “state” greatly impacts the accuracy of an ensemble predictor: including atomic-details would result in an intractable computation, while high-level representations that work in 1-dimensional sequence space can miss important steric and energetic details. Indeed, calculating the energy of all ensemble states in any 3-dimensional representation is considered computationally intractable [93]. Further, the choice of representation must align with an algorithmic model that allows efficient calculation of ensemble quantities, such as through recursion. Our goal is to develop an efficient prediction algorithm that quickly produces accurate, physically meaningful protein structure predictions, but which is still able to calculate the partition function of a complete ensemble space.

To capture critical 3-dimensional elements while retaining efficiency, we choose to represent β -sheet protein structure according to its super-secondary structural information — each state contains a sequence and a unique set of residue/residue β -strand backbone interaction pairs. However, allowing a combinatorial number of states with arbitrary super-secondary structure interactions would result in an intractable calculation. Moreover, many of these states would be biologically infeasible. Therefore, we incorporate known biological information about the proteins in question — in our case transmembrane β -barrel proteins and amyloids — to restrict state space to a tractable number of realistic interactions. In particular, we restrict the potential shape of a predicted structure to a known

family of proteins, conceptually resembling the concept of “architectures” in CATH [140] protein structure classification. Algorithmically, we employ bottom-up and top-down recursive methods to describe state-space, as each have benefits and downsides for the particular protein family being studied.

The rest of this chapter describes the first polynomial-time, recursive algorithms to compute the Boltzmann partition function of transmembrane β -barrel proteins and amyloid fibrils. Using the partition function, we show how to rigorously sample conformations from the Boltzmann low energy ensemble, and compute the Boltzmann pair probabilities $p(i, j)$ that residues i, j form an inter- β -strand contact. Additionally, this can be used to estimate statistical mechanics parameters such as ensemble free energy, average internal energy, and heat capacity. Such results permit an insight into protein structure landscapes that cannot be gained by methods solely dedicated to the prediction of single conformation. This approach also provides a unified framework that allows us to tackle a wide variety of structural prediction problems which were previously addressed by independent algorithms.

2.2 Transmembrane β -barrel modeling (Bottom-up approach)

We first describe an algorithmic framework for modeling transmembrane β -barrel (TMBs) protein ensembles. This work has been implemented as a publically-accessible web-based tool named *partiFold*². TMBs present an interesting problem for the application of ensemble techniques because of the surprising dissimilarity between sequences that form similar β -barrel folds, and, oppositely, the conformational variety seen in known β -barrel structures (e.g., Figure 3-3). This may suggest a relatively complex conformational landscape given any sequence.

We use a bottom-up approach to recursively model this family of protein structure and enable the calculation of the Boltzmann partition function. This representation allows us to design an algorithm that most efficiently describes TMB structure, but at the considerable cost of the time needed to identify and mathematically define every substructural element’s contribution to the ensemble as a whole. This can be particularly prone to error. However, we make this choice due in part by the fact that the transmembrane β -barrel structure has been extremely well characterized and appears to be a fairly conserved structural fold [164]. Therefore the benefits are worth the effort since it would be reasonable to expect few changes to the structural model of TMBs in the future.

²Available at <http://partifold.csail.mit.edu/>

2.2.1 Existing computational predictors for TMBs

The biological importance of both transmembrane α -helix bundles and transmembrane β -barrels has motivated the creation of numerous computational modeling tools. These techniques have generally focused on either (1) identifying the membrane spanning regions of sequence, or (2) classifying a protein as transmembrane or non-transmembrane from genomic sequence. While techniques for transmembrane α -helical bundles have been developed that offer highly accurate predictions [106, 193, 201], the long range interactions and low sequence similarity found in TMBs present a significantly harder problem, resulting in lower accuracy solutions [85]. Most existing methods rely on traditional machine learning approaches such as hidden Markov models (HMMs) and neural networks (NNs) [17, 79, 111, 118, 131]), however, these algorithms do not incorporate long-range interactions that are believed important for folding [175]. Methods that account for this, such as described below, should offer a better physical model. Further, few methods aim to predict the β -strand/ β -strand interactions that form super-secondary structure, instead simply classifying whether a sequence is transmembrane or not and identifying membrane spanning regions.

However, the prediction of generic β -sheet structure (not restricted to TMBs) has a long history of useful results [7, 31, 32, 35, 38, 89, 98, 117, 145, 156, 182]. Of particular note, BETApr [31] is a stochastic secondary and super-secondary structure predictor made specifically for β -sheets. This algorithm was arguably the top performer of the CASP7 inter-residue contact predictions competition [26], and most closely resembles our intended goal to stochastically predict β -strand contacts. Although we will compare against BETApr for the case of TMBs, it should be emphasized that BETApr is not TMB specific and its graph-based approach does not support the β -barrel closure created by pairing the extremal β -strands of the β -sheet.

2.2.2 Representing ensemble space via attribute grammars

We describe a simple and unambiguous representation of transmembrane protein structure by modeling them with multi-tape context-free grammars [195, 200]. In the case of transmembrane β -barrels (TMB), this modeling explicitly separates each of the antiparallel β -strand pairs involved in the barrel. The complete structure can then be described as a sequence of individual antiparallel pairings, including the closing strand pair. While the algorithmic concepts and routines presented here can be equally described without multi-tape context-free grammars, this representation provides a more concise conceptual description that still lends itself toward an efficient computational

solution.

Grammars provide a versatile framework that can be easily adapted to match the needs of experimentalists. Indeed, experimental observations of putative residue contacts, for instance, can be used to constrain the ensemble of folds to respect some specific structural features. Obviously, many others types of constraints can be designed, as was done by Waldispühl et al. [195] where some residues known to be present in extra-cellular loops were excluded from transmembrane strands.

To accurately represent TMBs using grammars (to agree with Schulz’s summary [164]) we must describe three fundamental features of these structures: (i) the overall shape of the barrel (the number of TM β -strands and their relative arrangement), (ii) an exact description of the antiparallel β -strand pairs which explicitly lists all residue contacts and their orientation (sidechains exposed toward the membrane or toward the lumen) as well as possible strand extensions, and (iii) the inclination of TM β -strands through the membrane plane. The modeling is based on an individual schematic representation of these features which will be merged hereafter. This decomposition of the structure into elementary units is illustrated in Figure 2-1.

A TMB must be decomposed into individual blocks of antiparallel β -strands, where each β -strand is involved in two distinct pairings — an exception being the “closing” strand pair involving the first and last β -strand. To handle this distinctly non-context-free feature, we employ a representation where the sequence is duplicated on a second tape, and pairings are made from one tape to the other. Figure 2-2 illustrates this representation, which is the foundation of the modeling introduced in [195,200] and motivates the designation of the “2-tape representation.”

To facilitate our description, we introduce a notation that allows us to generalize these models to compute critical features of the folding landscape. Each block can be represented as a 4-tuple $\binom{i_1, j_1}{i_2, j_2}$, where i_1 and j_1 (s.t. $i_1 < j_1$) are the indices of the strand on the first tape and i_2 and j_2 (s.t. $i_2 < j_2$) those on the second tape. Necessarily, $i_2 < i_1 < j_2 < j_1$.

We now consider an antiparallel pairing and the corresponding 4-tuple $\binom{i_1, j_1}{i_2, j_2}$. The left strand corresponds to the subsequence $[i_2 + 1, i_1]$, the right strand corresponds to $[j_2 + 1, j_1]$, and a loop to the subsequence $[i_1, j_2]$. Additionally, we assume that the rightmost amino acid at index $i_1 - 1$ of the left strand is paired with the leftmost residue at index $j_2 + 1$ of the right strand.

Although TM β -strands are not necessarily of the same length, the length of the residues in contact is $L_c = \min(i_1 - i_2, j_1 - j_2)$ and the length of the strand extension is $L_e = |(i_1 - i_2) - (j_1 - j_2)|$. To avoid invalid configurations, only one strand from each pair can be extended. When an extension is done on the left strand, the right strand becomes shorter and the extension is then called

a *reduction*; when an extension occurs on the right strand, the latter is longer and the operation is an *extension*.

The set \mathcal{C} of residue-residue contacts involved in strand pairing can be defined as follows: $\mathcal{C} = \{(i_1 - k, j_2 + 1 + k) \mid 0 \leq k < L_c\}$. The sidechain orientation alternates strictly around the strand backbone and can be labeled: *outwards*, that is facing toward the membrane, or *inwards*, that is facing toward the inside of the barrel, or channel (which can vary from entirely aqueous to mostly filled). Thus, we distinguish the subsets of residue contacts exposed to the same environment by the definitions:

$$\begin{aligned} \mathcal{C}_0 &= \left\{ (i_1 - 2 \cdot k, j_2 + 1 + 2 \cdot k) \mid 0 \leq k < \lfloor \frac{L_c}{2} \rfloor \right\}, \text{ and} \\ \mathcal{C}_1 &= \left\{ (i_1 + 1 - 2 \cdot k, j_2 + 2 \cdot k) \mid 1 \leq k \leq \lfloor \frac{L_c}{2} \rfloor \right\}. \end{aligned}$$

Assuming the location of the closest contact is known, we can also assign the nature of the milieu (i.e., membrane or channel).

For each block $\binom{i_1, j_1}{i_2, j_2}$ representing each distinct antiparallel pairing, we integrate these features by annotating each residue appropriately. β -strand residues with sidechains oriented toward the membrane are annotated with M, while those with sidechain oriented toward the channel are annotated with C. Unpaired β -strand residues are simply annotated E. An example of this modeling is given in Figure 2-3.

The inclination of strand through the membrane is modeled using a *shear number*. This feature is implemented with the help of strand extension. Indeed, strictly alternating *reductions* and *extensions* in consecutive strand pairs allows us to obtain the desired configuration. Without loss of generality, and in conjunction with experimental observations [164], we assume that (i) the N-terminus is located on periplasmic side and that (ii) shear number is positive. It follows that the first loop (between the first and second TM strand) must be on the extra-cellular side, and a positive shear number can be maintained by alternating the application of *reductions* and *extensions*. Figure 2-3 illustrates how to proceed.

It is worth noting that, in principle, a similar 2-tape representation could be used to include other classes of β -barrel protein domains as long as their structures followed similar topological rules. TMBs are well suited to the methodology given since the cell membrane restricts the number of possible structural conformations that can arise, reducing the complexity of the representation. However, soluble β -barrel proteins can allow more flexibility in the barrel forming β -sheet, and

would thus require more complicated rules (such as consecutive strands which are out of sequence order). These changes to the representation would affect the computational speed and tractability of our later techniques.

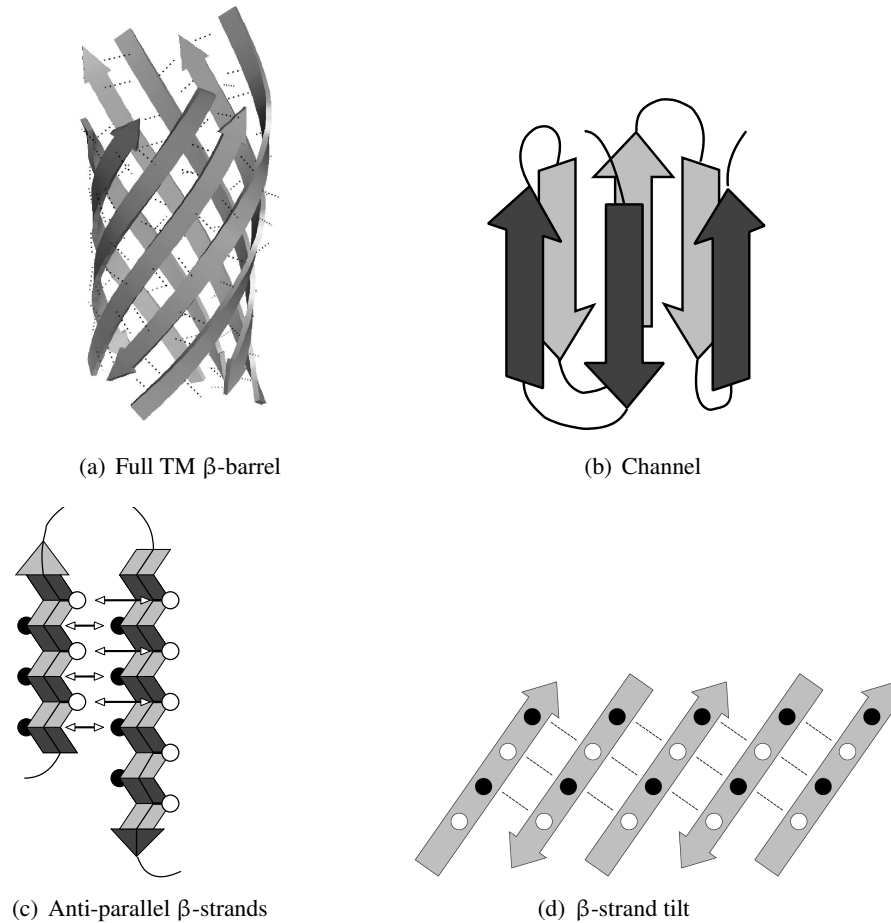


Figure 2-1: Structure decomposition of transmembrane β -barrel. **(a)** The full structure of a transmembrane β -barrel, **(b)** overall shape of the channel, **(c)** antiparallel β -strands and **(d)** inclination of TM β -strands across the membrane plane.

2.2.3 Computing the partition function

Computing the partition function of the transmembrane β -barrel state space above is the crucial, and most computationally intensive part of our algorithmic framework. While the partition function value \mathcal{Z} itself is only a normalizing constant, this value allows us to compute the likelihood of any given conformation in the ensemble. Further, the process of computing \mathcal{Z} calculates substructural

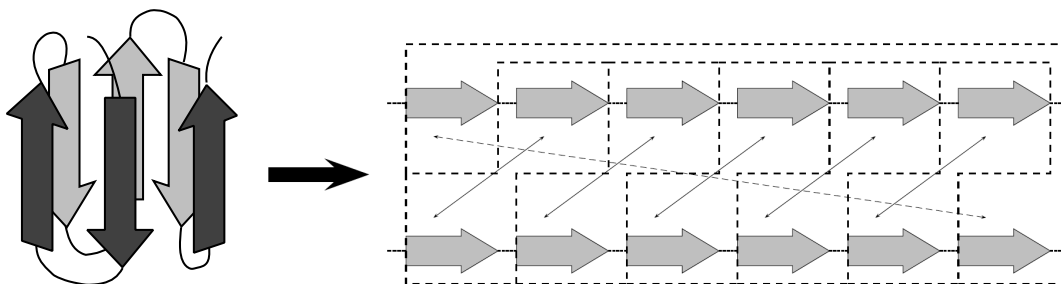


Figure 2-2: 2-tape representation of a transmembrane β -barrel. The original input tape is duplicated and pairings are only allowed from one tape to the other. All pairings are antiparallel and indicated with arrows. The closing pair connects the first and last strands and is represented by the exterior block.

scores that can also be used to efficiently sampling conformational space in a energetically-biased manner (Section 2.2.4).

A TMB structure can be represented as a sequence of antiparallel TM β -strand pairs, given any four indices i_1, i_2, j_1, j_2 and the environment x of the closing TM β -strand pair contact (i.e., “membrane” or “channel”). Given this, an energy $E(i_1, i_2, j_1, j_2, x)$ can be derived for the antiparallel β -strand pairing of $\omega_{i_1}, \dots, \omega_{i_2}$ with $\omega_{j_1}, \dots, \omega_{j_2}$. Section 2.4 provides further details on potential energetic models and the calculation of $E()$, however here we simply assume that $E(\cdot, \cdot, \cdot, \cdot, \cdot)$ returns a pseudo-energetic score. For all possible values of i_1, i_2, j_1, j_2 and x , we store the Boltzmann value $\exp(-E(i_1, i_2, j_1, j_2, x)/RT)$ in the array Q_{ap} (keeping in mind that RT may be set to 1 in practice). Since the length of TM strands, as well as those of strand extensions are bounded, the array can be filled in time $\mathcal{O}(n^2)$,³ where n represents sequence length.

$$Q_{ap}(i_1, i_2, j_1, j_2, x) = \prod_{k=1}^{L_c} \exp \left[-\frac{E(i_1 - k + 1, j_2 + k, x + k + 1 \bmod 2)}{RT} \right] \quad (2.1)$$

We necessarily assume an additive energy function (see Section 2.4), and decompose the energy of a TMB as the sum of the energy associated with each distinct antiparallel TM β -strand pair. Let n_s be the number of TM β -strands of the TMB s and let i_2^k (resp. $i_1^k - 1$) denote the index of the

³Note that this bound can be decreased to $\mathcal{O}(n)$ if we bound the length of loops. However, since we use this table to compute the contribution of the closing strand pair, this feature is not considered.

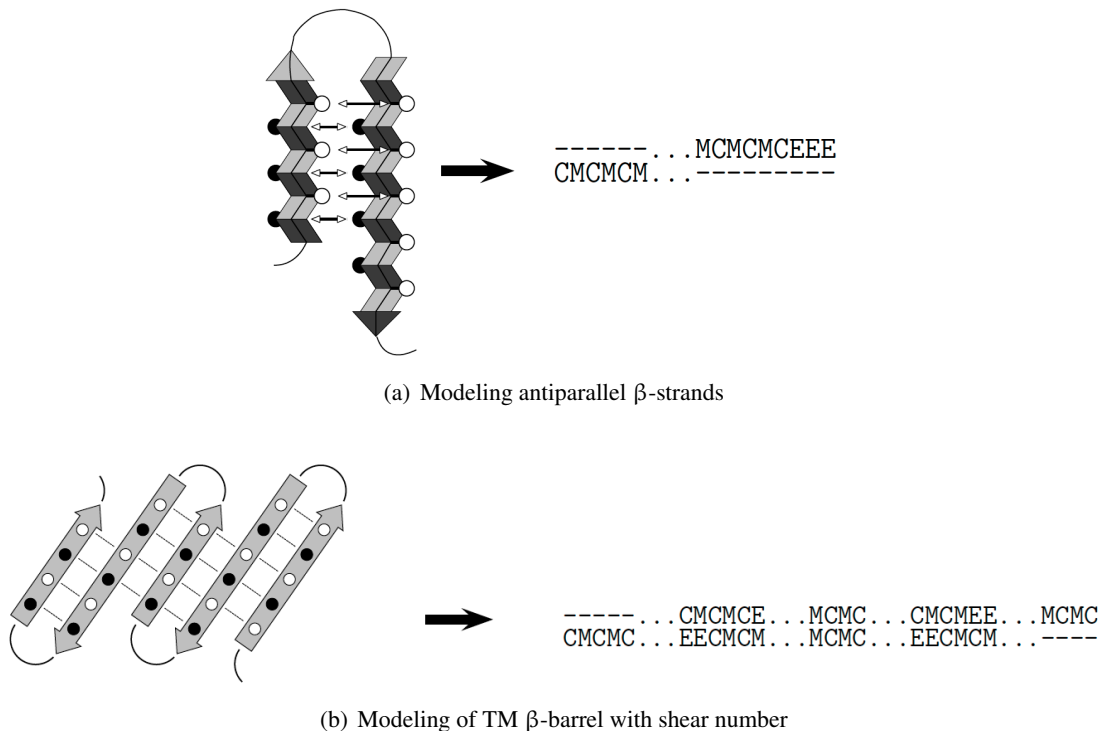


Figure 2-3: **(a)** Representation of a TM β -strand pair with extension on left strand (i.e., *extension*). Residues annotated by M [resp. C] have sidechain facing the membrane [resp. channel], while those with E are unpaired β -strand residues which extend or reduce the strand. Dots “.” represent the amino acids in the loop connecting the two strands, while dashes “-” are empty characters used to model the space available for the next pairing. **(b)** Representation of strand inclination using shear number. *Reductions* and *extensions* alternate around periplasmic loops (bottom) and extra-cellular loops in order to preserve the coherence of the orientation. The N-terminus of the protein sequence on the left diagram is at the right extremity.

leftmost (resp. rightmost) residue of the k -th strand.⁴ In order to simplify the algorithm description, in the following we will omit the parameter x used to indicate the environment of the first contact of an antiparallel TM β -strand pair. Therefore, the energy $E(s)$ of a given TMB structure s can be written as:

$$E(s) = E(i_1^{n_s}, i_2^{n_s}, i_1^1, i_2^1) + \sum_{k=1}^{n_s-1} E(i_1^k, i_2^k, i_1^{k+1}, i_2^{k+1}) \quad (2.2)$$

The Boltzmann partition function is defined as the sum $\sum_s e^{-\frac{E(s)}{RT}}$ taken over all the TMB structures s . To compute the partition function, we first introduce a dynamic table Q_{sheet} to store the partition function values for β -sheets built from concatenating antiparallel TM β -strand pairs,

⁴This notation is designed to respect the notation used for the strand pair block $\binom{i_1, j_1}{i_2, j_2}$.

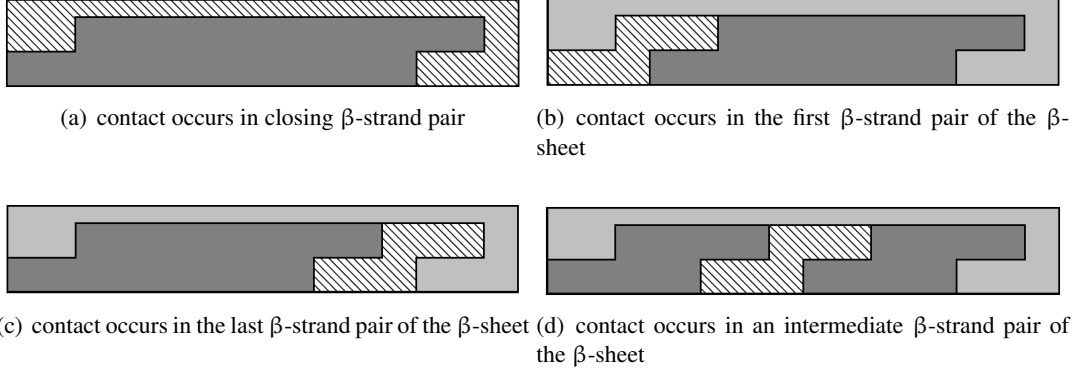


Figure 2-4: Decompositions of the transmembrane β -barrel, which allow us to isolate the antiparallel TM β -strand pair which contains the residue contact. The 2-tape block corresponding to this strand pair is crosshatched. The blocks in dark and light gray respectively represent a TM β -sheet (i.e., a sequence of antiparallel TM β -strands) and the closing strand pair (i.e., the antiparallel β -strand pairs which close the sheet and form the barrel).

i.e., TMB without closure. This table can be dynamically filled using the following recursion:

$$Q_{sheet} \begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} = \sum_{(k_1, k_2)} Q_{sheet}(i_1, i_2, k_1, k_2) \cdot Q_{ap}(k_1, k_2, j_1, j_2) \quad (2.3)$$

Once filled, we use this array to compute the partition function Q_{tmb} over all TMB. This operation consists of adding the contributions of the antiparallel β -strand pairs which close the extremities of the β -sheet. For this, we could use the values stored in Q_{ap} ; however, in practice, we use a special array that is better suited to the special rules for this last β -strand pair.⁵

$$Q_{tmb} = \sum_{(i_1, i_2)} \sum_{(j_1, j_2)} Q_{sheet}(i_1, i_2, j_1, j_2) \cdot Q_{ap}(j_1, j_2, i_1, i_2) \quad (2.4)$$

Note that in order to respect the pairwise orientation as well as strand inclination, the indices i_1, i_2 and j_1, j_2 are swapped. Finally, it should be mentioned that in computing the partition function, the dynamic programming must ensure an exhaustive and non-overlapping count of all structures; in particular, the cases treated must be mutually exclusive, as is clearly the case in our algorithm.

Using formulas from classical statistical mechanics, a number of important thermodynamic parameters can also be computed from the partition function. These parameters, including ensemble free energy, heat capacity, average internal energy, etc. (see Dill and Bromberg [52]), permit a better

⁵The rules for the closing pair, explicitly described in [195], mainly consist of relaxing some constraints, and allowing extensions on both sides of the strand.

understanding of the folding landscape. For example, as shown in [41], average internal energy of the structures $\langle E(s) \rangle$ can be computed by

$$\langle E(s) \rangle = RT^2 \cdot \frac{\partial}{\partial T} \log Q(s), \quad (2.5)$$

while the standard deviation can be computed with a similar formula. Such thermodynamic parameters provide information on the stability of folds for a given sequence.

2.2.4 Boltzmann distribution sampling

We derive a sampling technique that is used to randomly select whole structure conformations from the ensemble, weighted by their energetic score. Properties of these samples can then be calculated, such as structural clustering to identify common conformational variants. Demonstrated in Chapter 3, sampled structure clusters can often provide higher accuracy structure predictions than minimum folding energy predictions alone.

We design a rigorous sampling algorithm for TMB ensemble space inspired by an a technique for sampling RNA secondary structure according to Boltzmann distributions introduced by Ding and Lawrence [54]. Their method has been successfully applied to uncover critical features of the RNA folding landscape, as well as in biologically important applications such as *gene knock-down* experiments. For example, by analyzing Boltzmann samples of messenger RNA (mRNA), likely single-stranded regions of mRNA can identified that represent good targets for hybridization by small interfering RNAs (siRNA).

Formally, given an amino acid sequence s , we are able to randomly generate, according to the distribution of structures in the Boltzmann ensemble, low energy TMB structures for s . By sampling, we expect to be able to efficiently estimate non-trivial features concerning the ensemble of potential TMB folds.

The sampling algorithm uses the dynamic table filled during the computation of the partition function. It essentially proceeds in two steps illustrated in Figure 2-5. First, The “closing” antiparallel strand pair is sampled according to the weight of all TMBs that contain it over all possible TMB. Then, we sample each antiparallel strand pair of the TM β -sheet from left to right (or alternatively from right to left) until the last one, according to the weight of that structure over all possible TM β -sheets. The correctness of the algorithm is ensured by construction of the dynamic table in equations 2.3 and 2.4. We note that the minimum energy structure can also be computed through

similar means, by choosing minimum energy paths instead of a Boltzmann-weighted random selection.

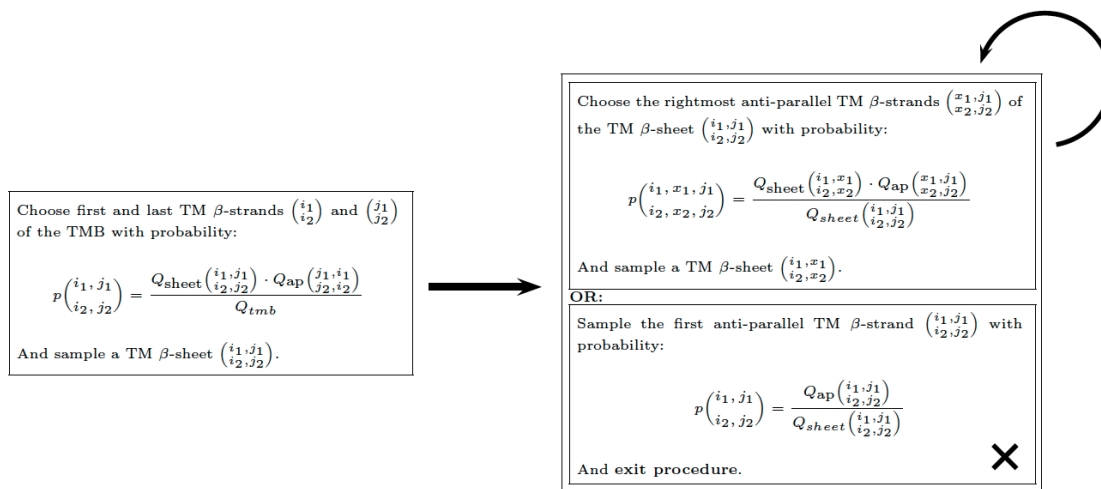


Figure 2-5: Sampling procedure: First, the first and last TM β -strands of the barrel are sampled (left box). Then, we sample the remaining TM β -sheet by iteratively sampling the rightmost antiparallel β strand of the remaining sequence, until we finally sample the first β -strand pair of the sheet.

2.2.5 Stochastic contact map and residue contact probability

Although conformational sampling serves as the primary output of our predictor, a number of other useful ensemble characteristics can be computed using the tables described in Equations 2.1 through 2.4. In particular, computing the Boltzmann pair probability that any residues i, j form an inter- β -strand contact can be highly informative — identifying sets of likely substructural components that many full-length protein conformations contain. Moreover, these values can be plotted graphically as a “stochastic contact map,” enabling easier analysis and potentially a better understanding of the conformational landscape.

Although such a property can be approximated through the analysis of a non-redundant sampling of conformational space, once our representation has been defined we can readily compute these quantities *exactly*. In this section, we present a method for computing the Boltzmann pair probabilities from the dynamic tables filled when computing the partition function value Q_{tmb} . To begin, however, we must characterize the antiparallel β -strand pairs which contain a given contact.

Property 1 *Let i and j ($i < j$) be two residues of two distinct consecutive antiparallel β -strands, and m and n (s.t. $i < m < n < j$) the two residues at the extremities of the connecting loop. Then,*

(i, j) are brought into contact if and only if $m + n = i + j$.

It follows from this proposition that (m, n) is a valid contact if and only if the antiparallel β -strands $\binom{i_1, j_1}{i_2, j_2}$ verify $m + n = i_1 + j_2 (= i_2 + j_1)$ and $i_2 < m < n < j_1$. In other terms $i_1 = m + k$ and $j_2 = n - k$ for $k \in \{0, \dots, \frac{n-m}{2}\}$

To evaluate the residue pair probability $p(i, j)$, we must compute the partition function value over all TMB $Q(i, j)$ which contain this contact. Such TMB can be decomposed into two, three, or four parts, depending on the strand pair where the contact occurs (i.e., in the the closing strand pair, the first and last pair of the sheet or in an intermediate one). All these cases are illustrated in Figure 2-4.

Let $\binom{i_1, j_1}{i_2, j_2}$ be an index of a block modeling an antiparallel TM β -strand pair. Then, we define $Q^{close} \binom{i_1, j_1}{i_2, j_2}$, $Q^{first} \binom{i_1, j_1}{i_2, j_2}$, $Q^{last} \binom{i_1, j_1}{i_2, j_2}$ and $Q^{inter} \binom{i_1, j_1}{i_2, j_2}$ to be the partition functions over all TMB structures which contain this antiparallel TM β -strand pair as, respectively, the pair closing the barrel (Figure 2-4(a)), the first pair of the TM β -sheet (Figure 2-4(b)), the last pair of the TM β -sheet (Figure 2-4(c)) or any other intermediate pair (Figure 2-4(d)). Formally:

$$Q^{close} \binom{i_1, j_1}{i_2, j_2} = Q_{sheet} \binom{i_1, j_1}{i_2, j_2} \cdot Q_{ap} \binom{j_1, i_1}{j_2, i_2} \quad (2.6)$$

$$Q^{first} \binom{i_1, j_1}{i_2, j_2} = \sum_{(y_1, y_2)} Q_{ap} \binom{i_1, j_1}{i_2, j_2} \cdot Q_{sheet} \binom{j_1, y_1}{j_2, y_2} \cdot Q_{ap} \binom{y_1, i_1}{y_2, i_2} \quad (2.7)$$

$$Q^{last} \binom{i_1, j_1}{i_2, j_2} = \sum_{(x_1, x_2)} Q_{sheet} \binom{x_1, i_1}{x_2, i_2} \cdot Q_{ap} \binom{i_1, j_1}{i_2, j_2} \cdot Q_{ap} \binom{j_1, x_1}{j_2, x_2} \quad (2.8)$$

$$Q^{inter} \binom{i_1, j_1}{i_2, j_2} = \sum_{\substack{(x_1, x_2) \\ (y_1, y_2)}} Q_{sheet} \binom{x_1, i_1}{x_2, i_2} \cdot Q_{ap} \binom{i_1, j_1}{i_2, j_2} \cdot Q_{sheet} \binom{j_1, y_1}{j_2, y_2} \cdot Q_{ap} \binom{y_1, x_1}{y_2, x_2} \quad (2.9)$$

Finally, using these functions, the partition function $Q(i, j) = \sum_S e^{-\frac{E(S)}{RT}}$, where the sum is over all TMB which contain the residue contact (i, j) , is computed as follows:

$$Q(i, j) = \sum_{\substack{i+j=i_2+j_1 \\ (i_1, i_2) \\ (j_1, j_2)}} \left(Q^{close} \begin{pmatrix} j_1, i_1 \\ j_2, i_2 \end{pmatrix} + Q^{first} \begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} + Q^{last} \begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} + Q^{inter} \begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} \right) \quad (2.10)$$

Finally, the Boltzmann probability $p(i, j)$ of a contact between the residues at indices i and j can be obtained by computing the value $p(i, j) = \frac{Q(i, j)}{Q_{tmb}}$. The contact map of a TMB can be immediately derived from this equation. However, we note that an extra field counting the number of strands in Q^{sheet} is required to ensure that the minimal number of strands in a TMB is not violated.

Assuming the length of TM β -strands and loops, as well as the number shear number values are bounded, the time complexity is $\mathcal{O}(n^3)$, where n is the length of the input sequence. When the maximal length of loop is in $\mathcal{O}(n)$, this complexity should approach $\mathcal{O}(n^4)$. Similarly, the complexity in space can be bounded by $\mathcal{O}(n^2)$.

2.2.6 Runtime optimizations

Unfortunately, the time requirements of a brute force approach to calculating Equation 2.9 are formidable. Indeed, naively applying this equation to the $\mathcal{O}(n^2)$ possible residue pairs results in an overall time complexity of $\mathcal{O}(n^5)$. In this section, we present a simple strategy using additional dynamic tables, which allows us to reduce the time complexity by a factor of $\mathcal{O}(n^2)$.

Two basic observations lead to an improvement over a brute force algorithm. First, when the TM β -strand pair which contains the residue contact is not involved, the product of the partition function of two substructures is realized over all possible configurations (i.e., $Q_x \begin{pmatrix} i_1, k_1 \\ i_2, k_2 \end{pmatrix} \cdot Q_y \begin{pmatrix} k_1, j_1 \\ k_2, j_2 \end{pmatrix}$) is computed over all possible pairs of indices (k_1, k_2)). In equation 2.9, the pairs of indices (x_1, x_2) and (j_1, j_2) are used for different residue contacts since the pair (i_1, i_2) varies. Thus we can pre-compute the values of $Q_{sheet} \begin{pmatrix} y_1, j_1 \\ y_2, j_2 \end{pmatrix} \cdot Q_{ap} \begin{pmatrix} j_1, i_1 \\ j_2, i_2 \end{pmatrix}$ over all possible (y_1, y_2) and store them in a dynamic table for later retrieval. Given (i_1, i_2) and (j_1, j_2) , let Q_{tail} be the array storing the values $\sum_{(k_1, k_2)} Q_x \begin{pmatrix} i_1, k_1 \\ i_2, k_2 \end{pmatrix} \cdot Q_y \begin{pmatrix} k_1, j_1 \\ k_2, j_2 \end{pmatrix}$. This table can be filled in time $\mathcal{O}(n^3)$. Then, in place of equation 2.9, we now have equation 2.11.

$$Q^{inter} \begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} = \sum_{(i_1, i_2)} Q_{sheet} \begin{pmatrix} x_1, i_1 \\ x_2, i_2 \end{pmatrix} \cdot Q_{ap} \begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} \cdot Q_{tail} \begin{pmatrix} j_1, x_1 \\ j_2, x_2 \end{pmatrix}. \quad (2.11)$$

Equations 2.7 and 2.8 cannot be improved in this manner since there is no redundancy in those cases. The time complexity for computing the entire contact map $p(i, j)$ is now $\mathcal{O}(n^4)$. However,

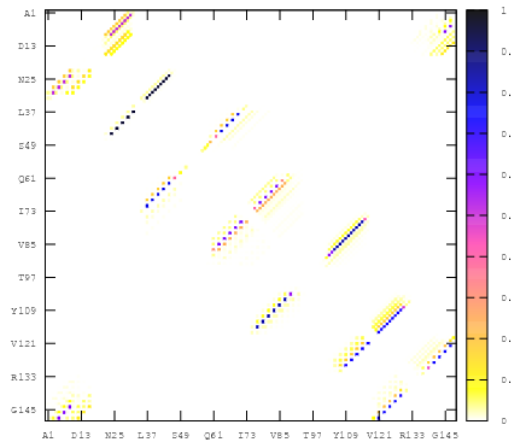
an additional observation allows us to save an additional factor n in time complexity: when a TMB structure is considered in one of the equations 2.6, 2.7, 2.8 or 2.9, the TM β -strand pair which contains the contact (i, j) also involves many other contacts. Hence, instead of using these equations to compute the values $Q(i, j)$ (and $p(i, j)$) separately, we consider each possible β -strand pair and immediately add its contribution to the partition function. From these improvements, we now have an algorithm to compute the contact map of a TMB, which runs in time $\mathcal{O}(n^3)$.

Although not explicitly mentioned thus far, we should emphasize that we can also compute the contact probability $p_x(i, j)$ for a specific environment x — i.e., membrane or channel environment. To do so, we simply need to duplicate the dynamic tables in order to take into account the sidechain orientation for extremal TM β -strand pairs.

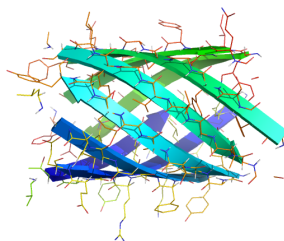
2.2.7 3-dimensional model generation

Although our TMB structural representation can provide a high level of insight through its definition of super-secondary structural elements, it is sometimes useful to be able to manipulate a true 3-dimensional atomistic model. Given that our model focuses solely on transmembrane β -barrel proteins only, it is possible to use super-secondary structure predictions to derive an atomistic structure for every Boltzmann sampled low-energy conformation. Further, this can be done without the need of any known homolog template whatsoever, allowing our technique to model TMB structures that do not directly correspond to the few existing TMB PDB conformations.

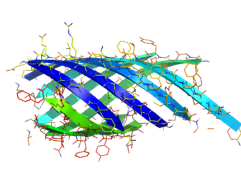
We first designed an algorithm that creates a 3-dimensional polypeptide backbone scaffold based on the number of membrane β -strands, the length of each strand, and the strand inclination. This technique uses basic geometric assumptions and a library of known backbone β -strand distances to model the β -sheet region of a TMB, and is derived from similar earlier work on transmembrane α -helices [29]. On this model we overlay the sequence specific amino acid residues, and assign sterically and energetically favorable sidechain orientations using an existing tool specialized for this purpose, SCWRL [24]. The loops between β -strand segments, having no super-secondary structure contact information, must then be added through other means. We have used the homology modeling tool Spanner [172, 173] to determine likely coil structures in these gaps in our scaffold, although other energy minimization or molecular-dynamics-based techniques could be applied. An improved approach could include a mechanism by which scaffold generation, sidechain assignment, and loop creation are iteratively refined. Such iterative steps have generally shown improved results in recent CASP competitions [217]. Figure 2-6 illustrates 3-dimensional model predictions.



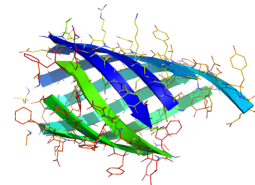
(a) 1QJ8 stochastic contact map



(b) Sampled conformation 1



(c) Sampled conformation 2



(d) Sampled conformation 3

Figure 2-6: Illustration of 3-dimensional models produced from predicted super-secondary structure information using the protein 1QJ8. (a) Predicted stochastic contact map of 1QJ8. (b)-(d) 3-dimensional rendering of a Boltzmann sampled structure (without spanning loop regions).

2.3 Amyloid fibril modeling (Top-down approach)

We now describe a different set of ensemble algorithms for modeling amyloid fibrils. This algorithmic framework, combined with methods described in Chapter 4, has also been implemented as part of a web-based tool named AmyloidMutants⁶. In this case we employ a top-down approach to recursively model such an extremely diverse family of structures, enabling an efficient calculation of the Boltzmann partition function. Unlike transmembrane β -barrel proteins, very little structural data is known about amyloid fibrils and drastically different topologies have been observed. Our goal is to design a framework that allows for the quick exploration of any one of these varied topologies. A top-down representation well suits this aim as it is generally easier to define recursively and less prone to error. However, this technique can also result in algorithmic inefficiency.

In this section we describe an approach for deriving multiple different amyloid fibril topo-

⁶Available at <http://amyloid.csail.mit.edu/>

gies, and methods for accelerating the partition function computation. We introduce “schemas” as an algorithmic construct that describe a family of millions of individual structural states that sum to represent a single fibril topology. Specifically, we have designed schemas to correspond with three largely distinct topology families: schemas \mathcal{P} , \mathcal{A} , and \mathcal{S} (Figure 2-7). These were designed to encompass the conformational variation found in most published experimental and hypothetical amyloid fibril structure models — while still allowing efficient recursive computation and excluding sterically impossible structures from the ensemble. This strategy allows maximal flexibility in the definition of a conformational landscape and at the same time minimizes the computational complexity.

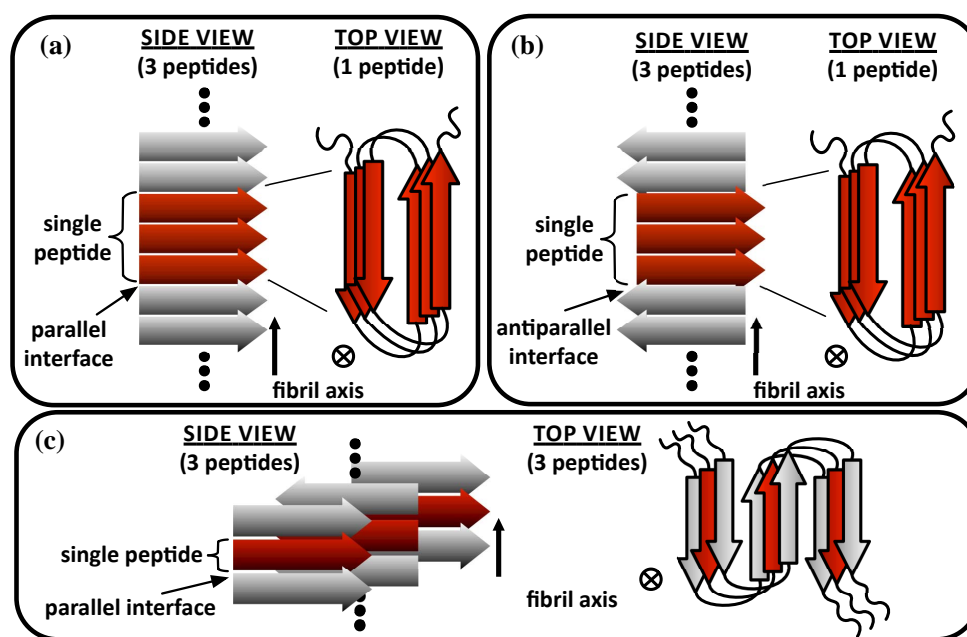


Figure 2-7: **Amyloid fibril schemas used for analysis:** Diagrammed are side and top perspectives of abstract schema. *Red* indicates a single fibril peptide flanked by two *gray* adjacent peptides along the fibril axis. (a) Schema \mathcal{P} , a 2-sheet β -solenoid with unrestricted number of rungs per peptide and parallel intra- and inter-chain interactions. Zero or one β -strand “kinks” are allowed for each β -sheet (see Figure 2-8). (b) Schema \mathcal{A} , identical to \mathcal{P} except with antiparallel inter-chain interactions. (c) Schema \mathcal{S} , a serpentine cross- β structure with unrestricted number of packed intra-chain β -sheets. All β -strand hydrogen bonds formed inter-chain.

For example, schema \mathcal{P} and \mathcal{A} describes an abstract “ β -solenoid” encompassing millions of structures with unique residue/residue interactions and varying numbers of β -strands, β -rungs, β -sheet width, coil location, residue orientation, and residue packing neighbors. Specific 2-, 3-, and 4-sheet β -helix-like structures are accounted for by the introduction of “kinks” (Figure 2-8). Sim-

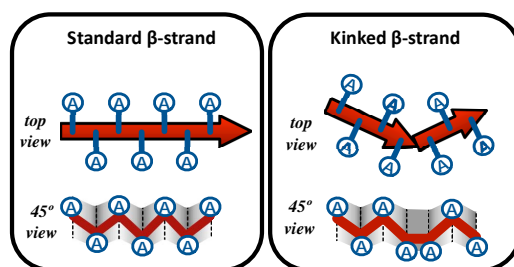


Figure 2-8: **The “kink” schema features allows more efficient β -helical modeling:** β -strand kinks are an algorithmic extension to the schemas defined in **Figure 1** that allow AmyloidMutants to model sharp β -sheet turns like those found in β -solenoids. A kink simply represents an interruption in the standard β -sheet in/out residue sidechain orientation, and separates a single β -strand into two, immediately adjacent β -strands. Modeling kinks instead of separate β -strands allows for a more precise energetic calculation, as these junctions differ from β -strands separated by extended coils, while also providing a significant computational speedup.

ilarly, schema \mathcal{S} represents millions of possible full-length peptide “serpentine” conformations, putatively containing multiple steric zipper interfaces. Conceptually, each schema “shape” can be thought again to resemble the “architecture” level of CATH [140] protein structure classification: for example, schema \mathcal{P} resembles the “2-solenoid” and “3-solenoid” classifications that make up 2 out of the 20 “mainly-beta” architectures in all proteins in CATH.

2.3.1 Existing computational predictors for Amyloids

Seminal work has shown that computational prediction of sequence amyloidogenicity can help guide and speed investigations of amyloid structure [2, 8, 69, 178, 183]. These advances enabled new possibilities for genome-wide studies, such as the discovery of 19 new functioning amyloid proteins in yeast [2].

More specialized tools [120, 181] have been further developed that detail the structure of one particular amyloid fibril conformation: “steric zippers,” a repeated, dry β -strand/ β -strand packing consisting of a few amino acids [132, 161]. However, other more elaborate amyloid conformations such as β -solenoids [205] cannot be considered by these specialized methods. Unfortunately, while these techniques can generate high-resolution structural predictions, they can only predict structural detail for regions of ~ 6 – 10 residues at a time due to the assumption of a steric zipper conformation. While such short segments may act as hot-spots for amyloid formation, a full-peptide structure prediction cannot be made which encompasses the size of amyloid sequences found in nature.

In the opposite vein, earlier tools are able to predict the amyloidogenicity of sequences of any

length, and agnostic to a particular molecular conformation, but unfortunately their structural prediction accuracy can suffer, achieving at best approximately 40% sensitivity on per-residue β -sheet location assignment and can exhibit insensitivity to sequence mutation [125]. Moreover, these predictions do not capture the finer details of β -sheet residue/residue-interactions that allow one amyloid conformation to be distinguished from another (i.e., even if they have identical β -sheet assignments).

The goal of our approach is to handle full-length amyloid sequences, improve structural prediction accuracy, and capture potential fibril structure variation, β -sheet residue/residue-interactions, and topological changes that may arise *in vivo*.

2.3.2 Representing ensemble space via recursive primitives

Schemas are defined as a recursive encoding of structure space consisting of combinations of irreducible structural subunits at the level of β -strand residue/residue and β -strand/ β -strand interactions. To represent amyloid fibril structures, which can amass thousands of peptide chains down their length, a schema formally defines only the possible conformations of a single peptide chain and its two immediate axial neighbors (see Figure 2-7). This representation models a theoretical fibril slice that is repeated indefinitely along the axis (e.g., if peptides ABCDE are adjacent in a longer fibril, then a schema defines the identical conformational landscapes of ABC, BCD, and CDE). The inclusion of axial neighbors in our model is necessary to ensure a realistic conformational symmetry between peptides — a property shown highly important in protein modeling [5]. Heterogeneous fibrils with relaxed symmetry constraints, and amyloidal interaction sites between different types of proteins can also be modeled by our schemas but are not shown in Figure 2-7.

More specifically, ensemble space is defined as putative geometric arrangement of β -sheets at the resolution of

1. intra-peptide strand-to-strand hydrogen bonding interactions along the fibril axis;
2. β -sheet-to- β -sheet packing arrangements perpendicular to the fibril axis (e.g., steric-zipper packings, etc.); and
3. Symmetry found between peptide chains, including inter-peptide strand-to-strand hydrogen bonds.

Therefore, a structural representation would indicate whether a residue is in a β -sheet or coil region, which other residue(s) it forms a hydrogen bonding pair with, which specific β -sheet it is in out

of the entire full-length protein, which specific β -sheet it faces (if applicable, such as inward-facing β -sheets in a β -solenoid), and what is the overall β -sheet architecture of the amyloid. Residue sidechain orientations are modeled by the standard β -sheet 180° reversal of each successive residue. The introduction of β -strand kinks (Figure 2-8 allow one to model a single residue deviation in the standard in/out sidechain orientation of β -strands. This scenario would physically manifest itself as two sequentially-adjacent β -strands with a sharp turn between them, as is found in many β -helices (e.g., HET-s in Figure 3-9). Our technique could also allow more complicated architectures to be constructed, such as heterogeneous-peptide fibrils, β -sheet donor-strand-exchange substructures [148,151,215], and other variants with non-symmetrical interactions. Our choice in resolution is meant to strike a practical compromise between the accuracy of energetic models, the efficiency of computation, the ease of physical interpretation, and the ability to incorporate experimental knowledge or intuition.

Note that schemas may be conceptualized as an *abstract* threading template (see Section 1.1.2). However, these should not be confused with standard threading templates used in other protein and amyloid modeling tools [181]: such tools fix a peptide backbone to a specific atomistic position and computationally score the effects of residue-specific sidechains, whereas schemas cover a much wider range of amyloid conformation and peptide backbone arrangements in 3-dimensional space using coarse representations.

Optional schema-dependent parameters can also be fixed for the three schemas defined:

1. limits on the length of β -strands or coils;
2. enabling or disabling β -sheet kinks;
3. requiring a minimum/maximum total-fibril β -sheet concentration;
4. enabling or disabling fibril twist (implemented via axially-adjacent β -strands “slipping” registration in a symmetrically consistent matter);
5. permitting N- and C-terminal coil asymmetries; and
6. allowing investigator-defined residue/residue hydrogen bond interactions to be fixed.

These parameters effect both the running time and accuracy of ensemble calculations, and allow specific point knowledge to be accounted for in the ensemble — as much or as little *a priori* knowledge as desired. This facility allows iterative tool reuse, enabling a more profitable back-and-forth

between predictions and experimentation, or can be used to make speculative predictions to help guide further experimentation.

2.3.3 Computing the partition function

Computing the partition function for schemas \mathcal{P} , \mathcal{A} , or \mathcal{S} is accomplished via a recursive definition of the energy for each amyloid fibril conformation. To account for the different types of structural interactions that can occur at the N-terminus, C-terminus, or in the middle of an individual peptide, each schema describes three recursive rules, an *N-rule*, a *M-rule*, and a *C-rule*. This effectively encodes the irreducible β -strand/ β -strand structural subunits mentioned above. The energy of each subsolution s_i is thus calculated in a depth-first manner by

$$E_{s_{i+1}} = \begin{cases} E_{s_i} + \begin{cases} E(N\text{-rule}) \\ E(M\text{-rule}) \end{cases} & \text{when applicable} \\ E(C\text{-rule}) & \text{initial } i = 0 \end{cases} .$$

Since these subunits are reused across many amyloid fibril conformations, the energetic result of every recursive call is stored in a dynamic programming table indexed by the parameters of the call. Subsequent recursive calls with the same parameters therefore perform a memoization table lookup rather than recompute the entire recursive tree.

The primary motivation of such a top-down approach is in the ease of programmatically encoding new schema definitions. This can help enable the rapid exploration of structural hypotheses as new experimental data comes to light. This separation also allows the *AmyloidMutants* tool to separate user rule encodings, which can be written quickly, from an algorithmic back-end that has been highly tuned to sample from the Boltzmann ensembles as quickly as possible. The present tool has been implemented in C++ via object-oriented templates and supports multithreading.

As an illustration, we show a simplified view of the recursions used in schema \mathcal{A} (Figure 2-9), omitting constraint checks found in each rule that ensure realistic fibril structures, allow potential *beta*-strand “slipping” (see below), or that include optional ensemble restrictions such as a limited β -sheet concentration range or specific residue/residue pairings. In this recursion, the *C-rule* is invoked first to select sequence indices $j1$ and $j2$ (anywhere along the length of the sequence), and β -strand lengths $l1$ and $l2$ (within the predefined range). Since schema \mathcal{A} describes a solenoid with antiparallel inter-peptide β -strand interactions, the C-terminals of each peptide must form a symmetric interface, where the energy of the interface between, e.g., index $j1+l1$ and $j1$ or $j1+l1$

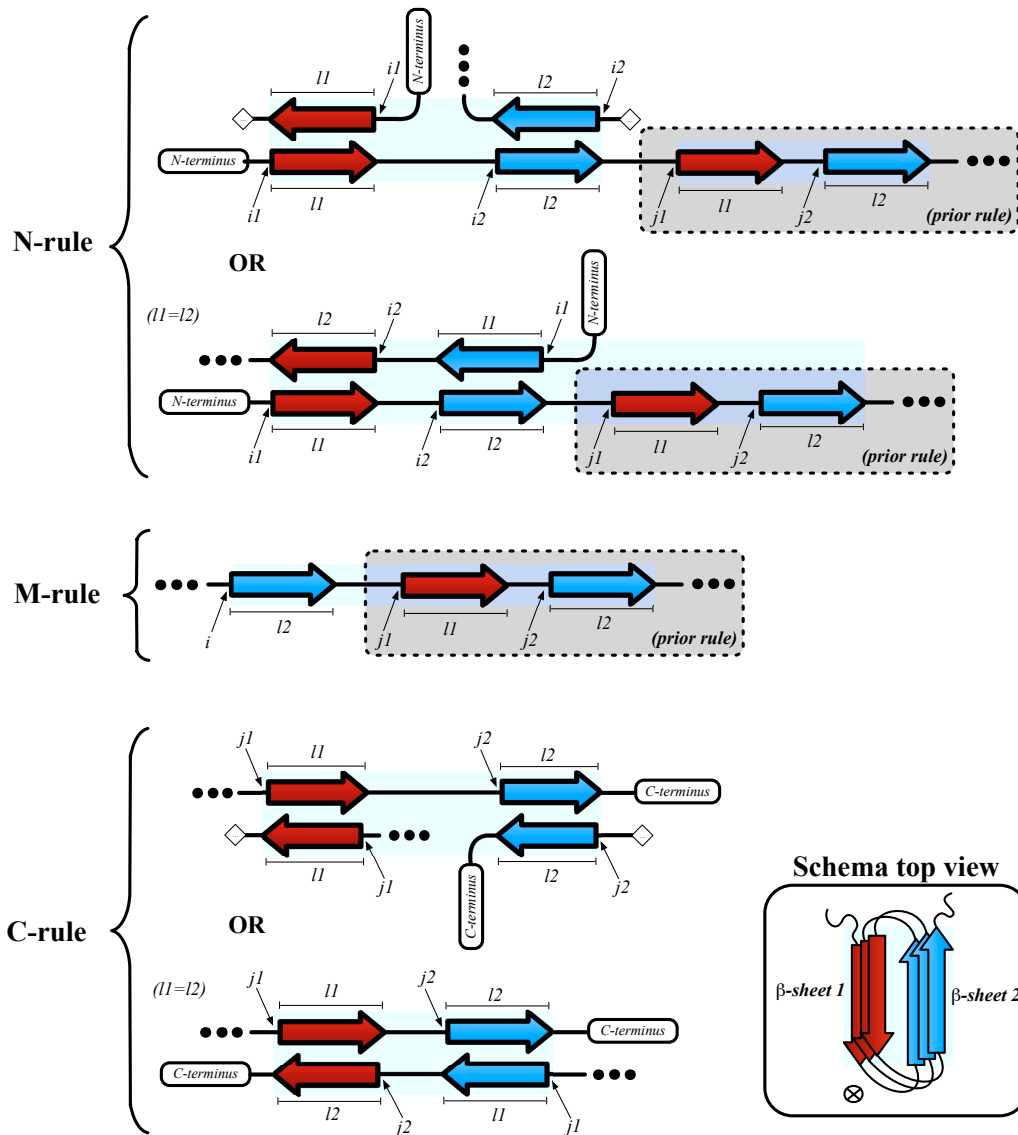


Figure 2-9: **Simplified recursion definition for schema A:** Every recursive step applies either the *N-rule*, *M-rule*, or *C-rule* to the current subsolution. The *N-rule* and *C-rule* can be applied using either of two symmetry rules for inter-peptide interactions. Adjacent β -strand/ β -strand interactions are indicated by alternating β -sheet color, while diamonds signify contiguous sequence points. The invocation of each rule selects variables $j1, j2, l1, l2$ (*C-rule*), i (*M-rule*), or $i1, i2$ (*N-rule*) according to sequence and symmetric constraints.

and $j2$ is also included. Given a sufficiently large value of $j1$ and $j2$, the *M-rule* and *N-rule* can then be called, although this is not required (permitting the case of a “single-rung” solenoid). For example, the *M-rule* selects an index i which initiates a parallel β -strand interface with another β -strand beginning at index $j2$, but only if $j1 > i + l2 + minc$ where $minc$ is the minimum length of a coil adjoining two β -strands (and only if the *M-rule* is called an even number of times, a choice effecting

N-terminal symmetry). If packing interactions are not considered (Section 2.4), only the energetic contribution of the ij_2 β -strand interaction is calculated (along with the coil region between $i+l_2$ and j_1). The *N-rule* is similarly called to select indices i_1 and i_2 which bind to j_1 and j_2 while maintaining N-terminal symmetry.

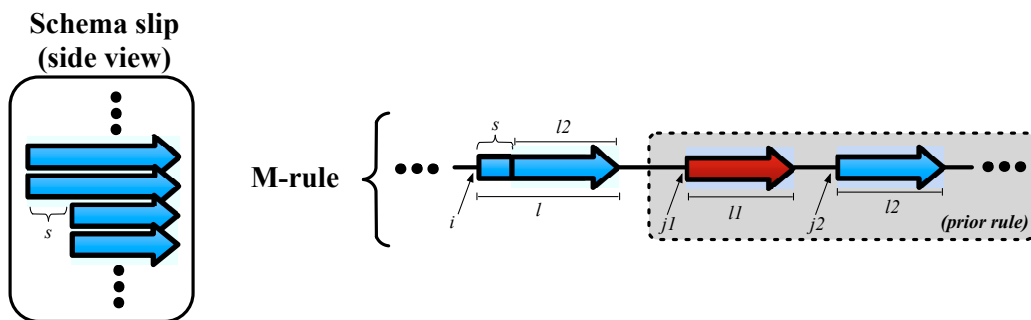


Figure 2-10: **Example *M-rule* extension for “slip” (schema \mathcal{A} or \mathcal{P}):** The use of slip allows individual β -sheet strands to vary in length, implemented in the *M-rule* via the use of variables i , l , and s . s can be positive or negative, although only a negative s is shown. The *N-rule* and *C-rule* slip extensions involve more complicated symmetry constraints than in Figure 2-9, but allow for more realistic interfaces.

One of the more important details omitted from Figure 2-9 is that of β -strand “slip.” This models β -sheets more realistically by allowing β -strand lengths to vary across the sheet. In the case of schema \mathcal{A} , the *M-rule* implements this feature by selecting an index i as well as a length l and slip value s that acts as a negative or positive offset to indicate what residue is in registry with position j_2 (Figure 2-10). The use of slip in the *N-rule* and *C-rule* is more complicated, greatly expanding the number of permissible symmetric inter-peptide interactions and removing limitations such as the requirement of $l_1=l_2$ in Figure 2-9.

Finally, we note that as in TMBs, the knowledge of the partition function \mathcal{Z} of a schema can enable the prediction of a number of other useful protein modeling properties, such as per-residue peptide flexibility estimate (akin to X-ray crystallography B-values), and thermodynamic variables such as entropy ($S = \partial/\partial T(RT \ln \mathcal{Z})$) and heat capacity ($C = 1/RT^2(\partial^2 \mathcal{Z}/\partial \beta^2)$).

2.3.4 Boltzmann distribution sampling

Similar to the case of TMBs, the principal output of our approach for amyloid fibrils is a sampling of structural states that are statistically representative of the full Boltzmann ensemble. Samples are generated by a stochastic backtracking procedure over the table of substructure conformation scores

that were memoized when computing \mathcal{Z} . A proper distribution is again maintained by biasing random backtracking selections according to substructure energy scores [54]. For each recursive step a random value is drawn between zero and the total energy of that step's recursive subtree (the sum of the score of all possible next recursions). Summing over all potential substructures until the total sum is greater than the random value thus selects the next substructure (although in some cases a table lookup can be performed faster). For convenience, the algorithm can be set to only sample unique structures during the backtracking steps (maintaining the same distribution). The implementation of more complex sampling methodologies is a topic of future work [146, 199]. Note, the minimum energy structure can similarly be computed by backtracking using a minimum energy paths instead of a Boltzmann-weighted random selection.

Individual samples can be informative in their own right, however, to gain a better picture of overall ensemble characteristics we cluster results. Sampled conformations are clustered by partitioning around mediods⁷ (a method similar to k-means), with a fixed number of clusters. To determine how many clusters to fix, we iteratively run the clustering algorithm with an increasing number of clusters until no new major conformational populations are qualitatively differentiated. Since we represent protein structure at the granularity of β -sheet residue/residue hydrogen bond pairs, the distance metric used within the clustering algorithm evaluates two structures' separation according to a sum of:

1. the per-residue secondary structure assignment overlap (i.e., 1-D Hamming distance),
2. the intersection of hydrogen-bonding-pair assignments (an F-measure, see Section 3.1.1, permitting 'one-off' $(i,j) \leftrightarrow (i,j+1)$ matches),
3. the F-measure statistic of the overlap of full β -strands (where a match occurs even if there is a shift between predicted structures, e.g., a β -strand at positions 60-67 can match another at positions 62-68), and
4. a similar F-measure statistic for the overlap of coil regions.

We have empirically found that the combination of these four criteria separates realistic structural populations better than any other individual metrics or combinations.

2.3.5 Stochastic contact map and residue pairing probabilities

⁷the sampled structure closest to the centroid cluster center

As with TMBs, we construct a stochastic contact map describing the Boltzmann-weighted likelihood $p_{i,j}$ (normalized by \mathcal{Z}) that any two residues i and j will form a β -sheet hydrogen bond, given all of the conformations in the schema ensemble. This can help identify small β -strand interaction motifs within the ensemble that may be otherwise hard to discern from full-conformation sampling.

Direct calculation of the exact likelihoods $p_{i,j}$ can be achieved by simply expanding the dynamic programming memoization table to include residue/residue pairing scores for every i, j pair at each sub-solution. However, this can significantly increase the program memory footprint required. In practice, we often estimate pairwise probabilities by performing a large conformational sampling, identifying pair frequencies in the sampled set, and normalizing $p_{i,j}^{est}$ according to each conformation's energetic score and the known partition function value \mathcal{Z} . The empirical convergence of $p_{i,j}^{est}$ to $p_{i,j}$ can vary widely according to schema, structural constraints, and the underlying sequence. However, in practice this can often be estimated within 1% error given 1,000 unique samples or less.

Finally, in some cases it is also helpful to analyze a modified stochastic contact map which presents the probability of any residue i and j forming a β -strand hydrogen bond with respect to the null-hypothesis stochastic contact map of that schema. This can further aid interpretation by removing schema bias and is discussed in Section 2.5.

2.3.6 Runtime optimizations

The choice to describe amyloid fibril schemas in a recursive top-down manner allows rapid prototyping of new schema types, but can decrease computational efficiency and restrict the use of optimizations like those presented for TMBs (Section 2.2.6). While more specialized optimizations can be implemented for each unique schema, we describe here methods that improve the runtime of any schema.

To do so we make judicious use of precomputation, operation ordering, and parallelization, and have designed new data structures that are optimized for our dynamic programming regime, including an optimized parallel hash table implementation. Further, in our approach, one of the primary determinants of computation speed is how well memory accesses patterns can be aligned to data caching and locality hierarchies found in modern computer systems. We have ordered our calculations, sometimes advancing or delaying results, to best optimize memory access. However, the recursive nature of our amyloid fibril schema implementation imposes some overheads over the case of TMBs.

Precomputation and parallelization

Throughout recursive steps, the energetic score of a subsolution is precomputed whenever possible. For example, before initiating the recursive *C-rule* calls (Section 2.3.3), the energy score of all potential β -strand/ β -strand rungs is calculated iteratively and grouped to attempt to minimize cache misses. This step also permits energetic threshold (see below).

We also take advantage of modern symmetric multiprocessing (SMP) multi-core systems to compute the partition function of a schema. Specifically, each recursive call to a *C-rule* is independently assigned to a one of many threads within a pool, accessing only a single shared resource, the memoization hash table. This assignment can also be grouped and ordered to take advantage of cache locality between cores. The high “fan-out” of this procedure permits thousands of concurrent threads. A parallelized version of our sampling procedure can be implemented in the same fashion. However, since sampling threads are read-only procedures, conformational sampling can be performed across multiple SMP cores, or even multiple computer systems via an option to serialize the memoization tables to disk.

Hash table implementation

Accessing memoized energy scores from recursive calls is the primary bottleneck in many dynamic programming approaches, including ours. Therefore, optimizing this data structure can yield considerable performance gains.

The most efficient implementation of a memoization table is a simple direct-mapped block of reserved memory where each schema sub-solution indices one and only one address in an atomic (lock-free) manner. However, such an implementation does not scale for more complicated schema definitions, as it can be extremely memory wasteful if the function mapping sub-solution indices to a physical address is sparse. Methods for perfect hashing [51,75] and minimal perfect hashing [19,30] can help determine efficient addressing functions, but are impractical for the size of most schema definitions. Therefore, we have implemented our own parallel hash map structures which has been optimized based on the following observations of our algorithm:

1. Objects are only ever inserted or read from the hash map, and are never deleted
2. Each map entry can only ever contain one value. Multiple inserts (e.g., from different threads) can either be ignored or can overwrite the value.

With these in mind we designed two separate concurrent hash table implementations that are either optimized for runtime speed or for minimal memory usage. The first implementation, optimized for speed, is based on standard quadratic probing, but uses a cpu-specific “compare and swap” (CAS) mechanism to ensure that table inserts are performed atomically. This avoids the costly use of operating-system level memory locks, and prevents data structure corruption caused by multiple threads attempting to insert at the same time. It also ensures that all read operations are valid. However, this method does not very good use of memory, and performance can drop significantly when more than 50% of the table has been filled [12].

To construct a hash table that makes better use of memory we implemented a concurrent version of a cuckoo hashing data structure [137]. This approach uses multiple hash functions to improve memory, but sacrifices computation speed. Our specific implementation uses 5 hash functions and 2 cells-per-bucket, which allows for efficient inserts and reads even when the table is over 90% full [64]. However, a single CAS operation can no longer guarantee data integrity during insertions so a more complicated locking mechanism is used along with its associated computational overhead. Further improvements could be made to these data structures, such as with the use of hopscotch hashing [84].

2.3.7 Runtime heuristics

The use of schemas in our framework is the essential component that enables a tractable partition function calculation over what would otherwise be an intractable number of states. However, by their nature, schema definitions can be encoded permissively, and thus some recursions can simply be too complex to compute using available amounts of memory or time. Memory size in particular can be the major factor deciding whether a computation can be done since more complex schemas generally require much larger memoization tables. Thus, for cases where memory is limited, we have implemented a set of heuristics that attempt to approximate the partition function through an intentional reduction of the number of states in ensemble space.

Energetic thresholding

Our first heuristic approach is based on the assumption that the majority of interesting low-energy conformations are composed of low-energy substructures. Thus we approximate the energetic landscape of an ensemble by filtering out high-energy β -strand/ β -strand interactions. This filter is applied before any recursive rules are called, during the precomputation of potential β -strand/ β -strand

rung energies.

In *AmyloidMutants*, we permit (but do not require) an algorithmic parameter that limits the ensemble analysis to only the lowest $N\%$ of β -strand/ β -strand interactions, as defined by their substructural energetic score. During recursive calls to the *C-rule*, *M-rule*, or *N-rule*, indices are not selected if the resulting β -strand/ β -strand interaction energy does not meet the threshold. This is implemented as a modification to the precomputation data structure above to limit the computational overhead of the decision. Such a thresholding approach has been applied successfully in similar RNA [86] and protein [198] structure analyses, and has the benefit of dramatically reducing memory usage and improving runtime speed, while maintaining a similar distribution of low-energy states.

Randomized space reduction

Our second approach to reduce the overall space and time complexity of the ensemble relies on the assumption that randomized pruning of the recursive call tree may result in a similar distribution. This is obviously a much stronger assumption and the resulting ensemble predictions can differ significantly from exact calculations depending on the specific schema. *AmyloidMutants* implements a randomized procedure for truncating the recursive descent through substructural states, with a non-uniform probability of truncation depending on call-tree depth. Importantly, this is achieved through a non-trivial hash table index filtering scheme which is calculated before the recursion begins. This is necessary since the specific random truncations performed during the initial computation of the Boltzmann partition function must be identically reproduced during conformational sampling.

2.3.8 3-dimensional model generation

A similar 3-dimensional prediction pipeline can be constructed for amyloid fibrils as describe for transmembrane β -barrels (see Section 2.2.7). However the algorithm for creating a polypeptide backbone scaffold must be tailored to the geometry of each and every schema, requiring considerable effort. With this in place, SCWRL [24] and Spanner [172, 173] remain decent choices for computing sidechain orientations and loop assignments. The integration of this feature into the tool *AmyloidMutants* is ongoing work.

2.4 Energetic models

Essential to the accuracy of our ensemble techniques is a potential-energy scoring function derived from frequency observations of specific residue/residue interactions in (non-sequence-homologous) PDB [14] protein structures. A predicted structure’s energy is then related to the sum of potentials for all residue/residue interactions (see below). Historically, many protein and RNA modeling tools have used similar ideas of knowledge-based potentials or potentials of mean force to accurately predict structure [20, 104, 170, 183, 197, 219] (although the biophysical interpretation of such potentials can vary). Key to success has often been the use of residue/residue interactions (or base-pairs in RNA), which can capture many of the important, energetically stabilizing features of 3-dimensional structure without requiring a high level of molecular detail (which complicates efficient algorithm design). Further, constructing an energetic scoring function from known PDB structures has the added benefit that no *a priori* expert information is required, and that as new structures are solved, the scoring function becomes more refined. However, relying on PDB structures also limits the ability to incorporate environmental conditions such as pH into the energy scoring function. This may be possible through *a priori* expert manipulation, but has not been explored thus far. Ultimately, the specific accuracy of any of these scoring functions depends on detailed choices on how to derive interaction frequencies from 3-dimensional structures.

As a reminder, we compute the Boltzmann partition function by making the assumption that the energy E_s of any protein state s can be linearly decomposed. Therefore the energy $E_s = -RT \log(p_s) - RT \log(\mathcal{Z})$ can be defined by $E_s = \sum_i -RT \log(p_{s_i}) - RT \log(\mathcal{Z})$ [40, 170], with the probability p_{s_k} represents the likelihood of observing a substructural state k , such as the propensity for two residues to pair within a β -sheet. In this way structure states are predicted according to steady-state conditions and do not directly reflect folding kinetics.

We introduce two novel energy functions inspired by classical statistical potential models of residue/residue interactions [20, 89, 124, 195]: conditioned pairwise interactions and residue stacking pairs. Both energetic models can be applied to either transmembrane β -barrel or amyloid fibril prediction algorithms. However, a key feature of our algorithms is the ability to include a wide range of statistical potential scoring metrics such as quasi-chemical interaction propensities [124] or even a combination with position-specific scoring matrices [120].

2.4.1 Conditioned pairwise amino acid interactions

Our energy scoring function is based on the statistical potential that two residues pair within a β -sheet [20,195], uniquely conditioned by the 3-dimensional environment found in the PDB structure, such as amphipathicity and solvent accessibility, β -strand edge proximity, residue-stacking ladders, β -sheet edges, and β -sheet twist (e.g., $p(i|j, env)$). This increases the dimensionality of information captured by simple pairwise contacts. Since our ensemble representations conceptually represent a coarse 3-dimensional topology, an appropriate set of energies can be chosen at each step of the search through structure space — in other words, each structural state not only defines a set of contact pairs, but also an associated environment for each pair (e.g., residues/residue pairs facing toward the center of the β -solenoid in amyloid fibril schemas \mathcal{P} and \mathcal{A} would be considered to have an environment that is solvent inaccessible). Note, importantly, that these environments are not assigned to sequence positions, but are associated only with each of the millions of ensemble states. We combine these β -sheet statistical potentials with similar potentials for consecutive residues forming coil ($p(i, j)$). Further, an optional hydrophobic packing score can be added, describing the propensity for two residues to pack between two β -sheet faces [107]. The relative influence of each of these terms can be scaled independently so one can investigate multiple facets of structural interactions.

To obtain statistical potentials for all possible amino acid pairs we compute the probability of observing pairs in solved β -sheet structures, conditioned on each of the environments. Similar to prior methods [20, 49, 99, 195], we analyze the 50% non-redundant set of PDB [14] protein sequences (PDB50), regardless of model resolution or whether the structures were derived using X-ray crystallography or NMR. The tool STRIDE [76] is then used to identify secondary structure features, solvent accessibility, and hydrogen bonds. When validating the accuracy of our predictors (Chapter 3), all solved structures of TMBs and amyloid fibrils are removed. Frequencies of amino acid pairs for each environmental condition are extracted using specific rules. For example, the identification of residues belonging to an amphipathic regions is determined by the alternation of at least 4 buried and exposed residues, where buried residues are required to have less than 4% the solvent accessible area as when that residue is in an extended G-X-G tripeptide [37], and an exposed residue is required to have an area greater than 15%. For consistency and to avoid parameter fitting, specific statistical *bonuses* have not been included in these potentials (e.g., special treatment of proline residues).

2.4.2 Stacking-pair amino acid interactions

We now define a novel energy function inspired from the classical Turner model for RNAs [213] and using for the first time in β -sheet structures. To describe this, we first introduce the notion of a *stacking pair* in a pair of β -strands. Intuitively, this consists of the stacking of two spatially adjacent pairs of hydrogen-bonding residues that have the same sidechain orientation. Figure 2-11 depicts such an arrangement. More specifically, consider an antiparallel β -strand pair and two residues, indexed i and j , such that i corresponds to an amino acid in the first strand and j to an amino acid in the second one. Then, assuming both pairs are hydrogen-bonded, the 2-tuple $((i, j), (i + 2, j - 2))$ is said to be a *stacking pair* of β -strand residues. The choice of the pair $(i + 2, j - 2)$ (as opposed to $(i + 1, j - 1)$) ensures that residues on the same face of the β -sheet are grouped since these are much closer in physical space and more likely to interact with one and another.

Similar to the methods for calculating pairwise statistical potentials, *stacking pairs* frequencies are tabulated, and we estimate the conditional probability $P((X, Y) | (U, V))$ of observing the amino acid pair (X, Y) given an adjacent stacking pair (U, V) . Thus we define $E(i, j, x | i + 2, j - 2)$ as the energy of the contact between residues ω_i and ω_j , with the environment x , given the adjacent stacking pair ω_{i+2} and ω_{j-2} . However, since a table of amino acid specific stacking pair potentials would require 20^4 entries, the only way to extract meaningful information from the PDB50 is to determine potentials based on a reduced residue alphabet. We investigated a number of reduced alphabet sets and decided upon the Wang & Wang 5-letter reduced alphabet [202]. Section 3.3 studies further the rationale behind this choice.

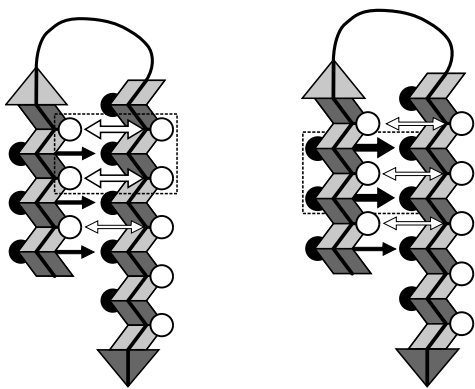


Figure 2-11: Anti-parallel β -strand “stacking pairs.”

2.5 Schema comparison and prediction normalization

To enable the efficient computation of the partition function \mathcal{Z} , our ensemble predictions are computed under the assumed condition that the input sequence is a member of a particular protein family — a transmembrane β -barrel protein or an amyloid. These assumptions impose specific structural constraints, allowing the tractable calculation of an exponential problem, but also remove any non-TMB or non-amyloid solutions from the output space. This approach is chosen to allow the use of our tools in the frequent biological circumstances where a protein has already been characterized or assumptions have already been made from other evidence. The *de novo* prediction of whether an arbitrary sequence will fold into any TMB structure or amyloid fibril state is a much different problem not directly addressed by our approach. None-the-less, given that we allow ensemble predictions of various, incompatible schema spaces, it is important to address any inherent biases in our results.

2.5.1 Stochastic contact map normalization

Schemas define a particular set of millions of possible structures, and therefore the likelihood that two residues i and j within this set form a β -strand hydrogen bond is not uniform across all (i,j) pairs. Although this non-uniformity may only introduce a small bias in pairwise probabilities, it can sometimes be informative to view predicted stochastic contact maps (Section 2.2.5 and Section 2.3.5) which have been normalized. An illustration of this variance can be seen by calculating the “null hypothesis” probability of any (i,j) pair under the assumption of a constant energy for all interactions within an ensemble (a “null hypothesis contact map”). This is most easily computed by fixing the Boltzmann constant $RT = \infty$ in our energy model. Subtracting null-hypothesis (i,j) pairs from any predicted $p_{i,j}$ highlights predicted specificity. Figure 2-12 visualizes this point.

2.5.2 Schema comparison

The comparison of predicted protein conformations from different schemas (for example, structures sampled from schema \mathcal{P} and schema \mathcal{A}) is a complicated problem and not a primary goal in our modeling framework. Such comparisons require additional assumptions and are not the same as more informative comparisons within a single calculated ensemble. Specifically, our energy model is based on statistical potentials, which serve as a crude analog for free energy and are based on strong assumptions [104, 170]. However, these potentials do not separately account for enthalpic and entropic energetic contributions. When analyzing predictions within a single schema, struc-

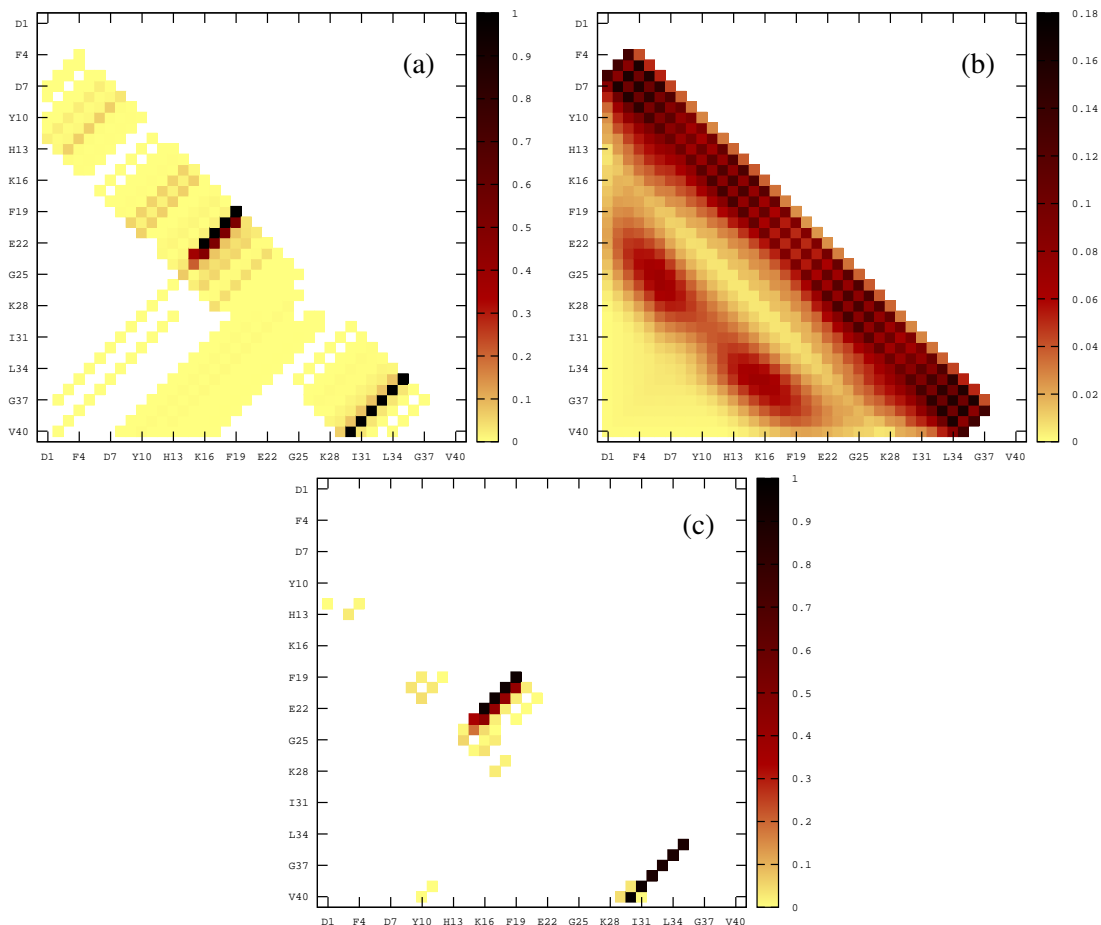


Figure 2-12: **Illustration of null-hypothesis contact map:** AmyloidMutants predicted contact map of $A\beta_{1-40}$ (using schema \mathcal{A} with allowable β -strand lengths between 6 and 12 residues) **(a)**, and the null hypothesis contact map of a sequence of the same length using the same schema and parameters **(b)** (note the change in density scales). The positive difference between **(a)** and **(b)** is depicted in **(c)**.

ture pseudo-energies can be compared without specific regard to entropy, since the entropy of all predictions within the ensemble is the same and fixed by the schema. However, comparisons between schemas involve a different number of potential (dissimilar) states, requiring a less ambiguous model of entropy. Such a model could involve a simple count of non-overlapping states, used to scale pseudo-energy scores, or a more thorough analysis of null-hypothesis contact distributions; in any case, requiring additional energetic assumptions. This question is a subject of ongoing research.

Chapter 3

Evaluation of ensemble structure predictors

In this chapter we validate the accuracy and utility of our ensemble techniques by comparing predicted results against known structural information of transmembrane β -barrels and amyloid fibrils. We further evaluate the impact of our new energetic models (introduced in Section 2.4) on predictive accuracy.

3.1 Transmembrane β -barrels

The partiFold framework described in Section 2.2 calculates the Boltzmann partition function to predict the ensemble of structural conformations a TMB may assume. From this, protein conformations can be sampled, and experimentally testable TMB properties can be calculated that further describe the energetic folding landscape. In the following section we demonstrate the flexibility, reliability, and potential of the approach by evaluating three different prediction problems:

1. Residue/residue β -strand contact predictions can be performed with state-of-the-art accuracy using stochastic contact maps.
2. X-ray crystal per-residue B-values and residue flexibility can be predicted with accuracy rivaling that of leading algorithms using contact probability profiles.
3. Whole structure prediction accuracy can be improved over minimum folding energy approaches by performing Boltzmann distributed structure sampling.

3.1.1 Residue/residue β -strand contact prediction

Residue/residue contact prediction has the goal of assigning a statistical likelihood of interaction to every potential residue/residue interaction pair (i.e., n^2). In our model and others describing super-secondary β -sheet structure, this interaction represents the pairwise adjacency of two β -strand amino acids that share hydrogen-bonds across their backbones. Beyond providing a 2-dimensional interpretation of protein structure, these predictions remain an important concern when reconstructing 3-dimensional models [26, 78, 147].

How to evaluate ensemble predictions

The class of predictions enabled by partiFold embody whole-ensemble properties of a protein, and not simply residue/residue β -strand contact predictions. Therefore, we first describe means for interpreting these contact probabilities from ensemble stochastic contact maps. The stochastic contact map reflects the likelihood of two β -strand amino acids pairing in the defined Boltzmann distribution of conformations, and not the residue/residue contacts involved in any one single structure. Figure 3-1 depicts two ways to view information from a stochastic contact map. On the left, a the full contact map of 1P4T is shown, identifying the probability of contact for all possible pairs of residues across all conformations in the Boltzmann distribution. On the right, a single structure is chosen (in this case the X-ray structure of 1QJ8, but it could be any sampled conformation), and displayed as an unrolled 2-dimensional representation of the β -barrel strands and their adjacent residue contacts. Using the stochastic contact map, residue contact pairs are then colored to indicate a high (red) or a low (cyan) probability in the Boltzmann distributed ensemble. From this, substructures may be identified by their relative likelihood of pairing.

For our residue/residue contact prediction comparisons, we define a set of single contact predictions by selecting all pairwise contacts that have a probability greater than a given threshold p_t in the stochastic contact map. Other approaches could be used (such as sampling and clustering of contacts), however this metric provides a stochastic ensemble-wide view of the folding landscape and can help identify signal from noise through the parameter p_t . Further, validation of our results are limited by the availability of a *single* solved X-ray crystal structure for each test protein. Therefore, we focus validation on the task of single contact prediction of X-ray crystal structures even though much more information can be obtained from our results about the nature of the folding landscape, suggesting future experimental directions. In our tests it is this set that is compared

against the corresponding contacts found in X-ray crystal structures as annotated by STRIDE [76]. To evaluate our contact predictions we rely on three standard measures: the coverage (i.e., sensitivity), where $\text{coverage} = \frac{\text{number of correctly predicted contacts}}{\text{number of observed contacts}}$, the accuracy (i.e., positive predictive value), where $\text{accuracy} = \frac{\text{number of correctly predicted contacts}}{\text{number of predicted contacts}}$, and the F-measure, where $\text{F-measure} = \frac{2 \cdot \text{Coverage} \cdot \text{Accuracy}}{\text{Coverage} + \text{Accuracy}}$.

To demonstrate how these metrics apply to stochastic residue/residue contact prediction, we refer to Figure 3-2 depicting the accuracy of contact prediction for 1QJ8 as a function of the size of the predicted set (e.g., p_i). On the left one finds a high predictive accuracy ($\approx 60\text{--}70\%$) when the number of contact predictions made is roughly the number of contacts in the X-ray crystal structure ($\approx 100\text{--}120$ pairs). The flatness of the curves further indicates a good separation between accurate, highly probable contacts, and background predictive noise. This type of result could suggest a good scaffold of likely contacts when constructing a 3-dimensional model of an unknown structure.

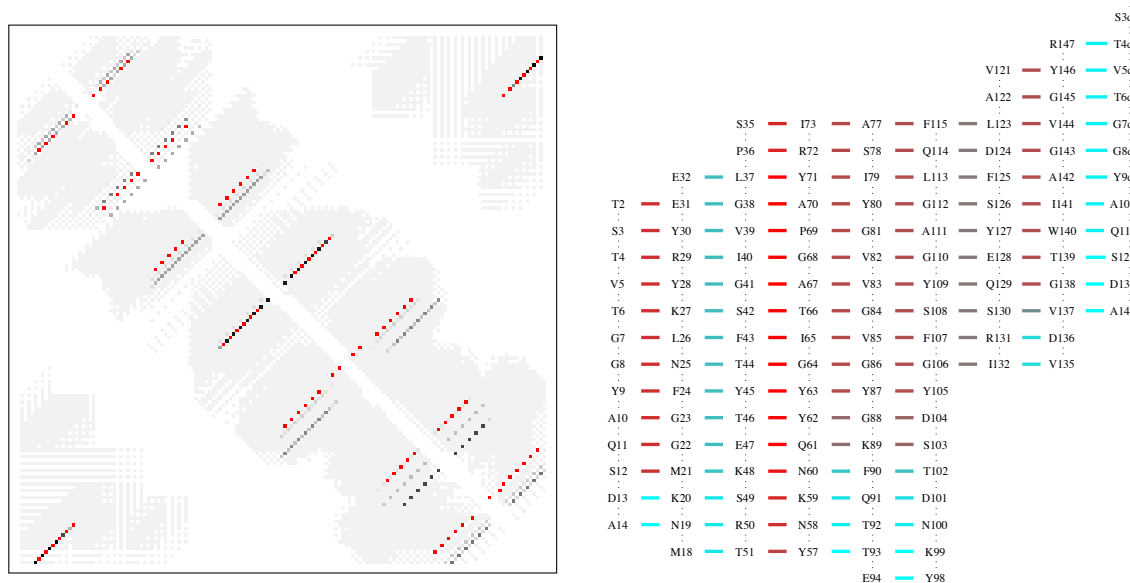


Figure 3-1: **Illustrative representations of stochastic contact predictions: Left:** Stochastic contact map for 1P4T. Horizontal and vertical axes represent residue indices in sequence (indices 1 to 155 from left to right and top to bottom), and points on the map at location (i, j) represent the probability of contact between residues i and j (where darker gray implies a higher probability). The X-ray crystal structure contacts of 1P4T are shown in red. **Right:** 2-dimensional representation (un-rolled β -barrel) of 1QJ8 X-ray crystal structure showing only those residues involved in β -strands (shown vertically and successively numbered) and their associated, in-register H-bonding partners. Computed contact probabilities are indicated by color hue (highly probable in red, low probability in cyan). The leftmost β -strand is repeated on the right to allow the barrel to close, labeled *dup*.

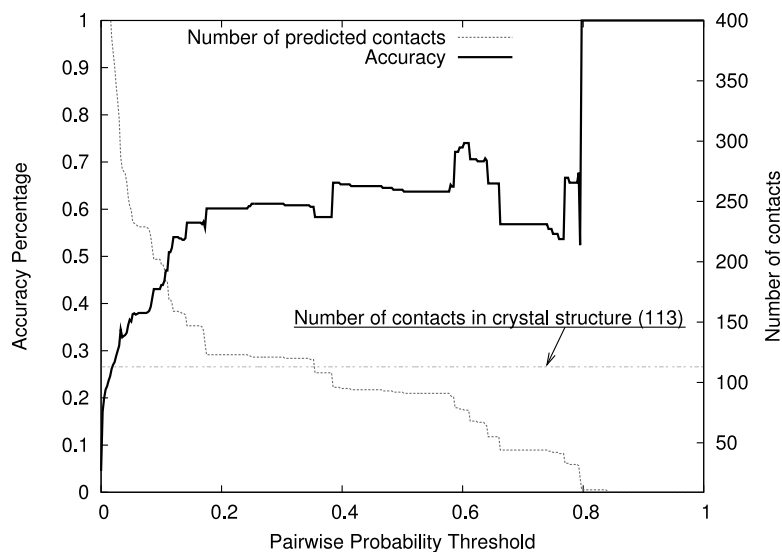


Figure 3-2: **Metrics for comparison of ensemble predictions to X-ray crystal data:** Ensemble predictions produce only residue-pair contact *probabilities*, preventing direct comparison to X-ray crystal data. For example, one approach is to choose a threshold to select which residue-pairs exist. Shown is a plot of prediction accuracy as a function of the size of the predicted contact set for the protein 1QJ8.

TMB structures used for comparison

Very few TMB have experimentally-derived structures deposited in the PDB. After removing homologous sequences, and focusing on monomeric, amphipathic TMBs without any plugging domains, we find in our test set 8 proteins with known X-ray crystal structures (PDB codes: 1QJ8, 1P4T, 1QJP, 1THQ, 1K24, 1QD6, 1TLY, and 1I78 — Figure 3-3). Larger OMPs such as porins have been excluded since they typically exist in trimer, and can contain short α -helical loops which are critical for stabilization. Similarly, a number of TMBs are found to have large plug domains within the barrel itself, likely stabilizing the structure in an irregular, possibly dynamic fashion. Given *a priori* knowledge of such configurations, it may be possible to adjust our model to provide accurate predictions, however, the current energy function has not been formulated with this goal.

This paucity of experimental structural information is in fact recurrent in TMB structure prediction research. For instance, only 8 structures were used to train and evaluate PROFtmb, a state-of-the-art genomic-level TMB existence predictor [17].

We further divide our test set of 8 to distinguish *short* (<200aa: 1QJ8, 1P4T, 1QJP, 1THQ) and *long* (>200aa: 1TLY, 1K24, 1I78, 1QD6) proteins, and apply two different choices of grammatical constraints (Table 3.1, adopted from Waldspühl et al. [195]). This matches an observed link

between the length of the peptide sequence and the length, number, and shear of the strands that make up the barrel. Moving forward, we believe that the effectiveness of our techniques could be enhanced by a well-formulated machine learning approach to parameter optimization as has been applied to the case of RNA [55,56].

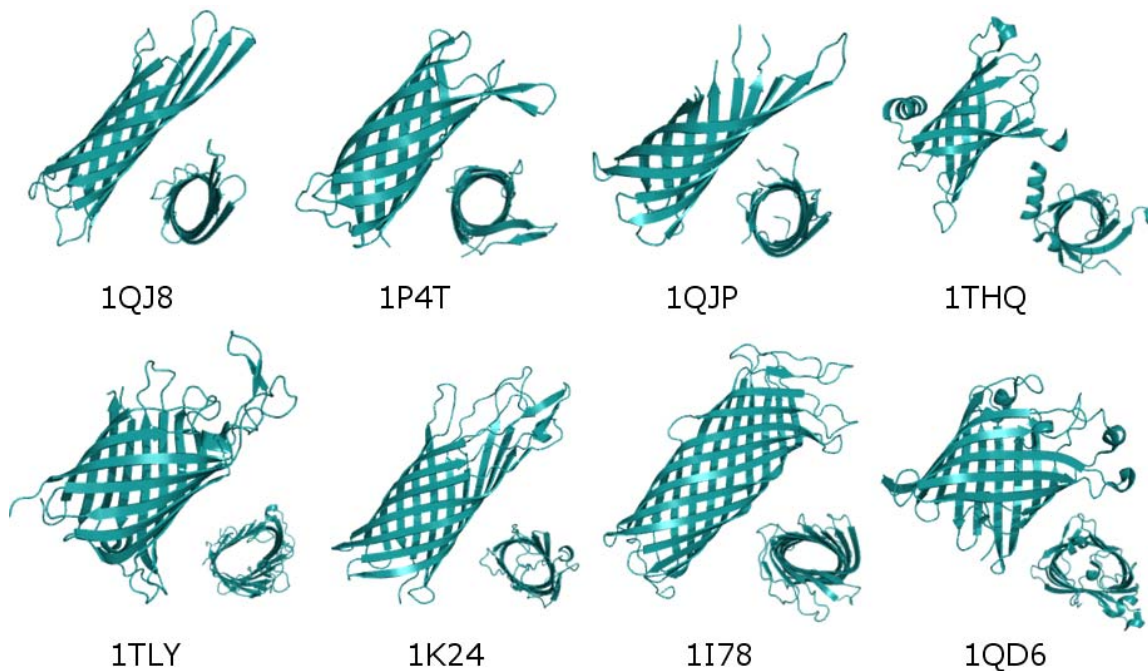


Figure 3-3: **3-dimensional graphics of known TMB structures used in validation:** Presented are the 3-dimensional structures of the 8 known monomeric, amphipathic, plug-domain-free TMBs – shown from both the side and top.

Constraint	<i>short</i> TMBs	<i>long</i> TMBs
Min/max number of β -strands	8-8	10-12
Min/max β -strand length	11-21	10-25
Min/max shear value	0-3	0-4
Min/max periplasmic loop length	2-15	2-8
Min/max extra-cellular loop length	2-25	2-35
Maximum length variation between β -strands	6	10

Table 3.1: **TMB grammar constraints used for validation:** Constrains chosen for either *short* TMB proteins (<200aa: 1QJ8, 1P4T, 1QJP, 1THQ) or *long* TMB proteins (>200aa: 1TLY, 1K24, 1I78, 1QD6) based on observed characteristic differences.

Evaluation of contact prediction

We test the accuracy of our algorithm using the described approach and display the results in Figure 3-4. We also compare partiFold’s residue contact prediction accuracy with the abilities of BETApro [31], another state-of-the-art residue/residue contact prediction technique. It should be noted that while BETApro does provide a stochastic contact map of β -strand interactions, its interaction probabilities are not related to a Boltzmann distribution of conformations, but rather based on a sophisticated neural network and graph algorithm that aims to predict a single structure. Its energy model also appears to not be common across all proteins, and, unlike our model, incorporates secondary structure and solvent accessibility profiles of the target amino acid sequence. Finally, BETApro was designed for, and trained on, globular proteins, and it does not support important aspects of β -barrel architectures such as circular β -sheets. Thus, during comparison, one must keep in mind the BETApro was not designed specifically for TMBs.

A comparison of F-measure scores (see Section 3.1.1) is plotted in Figure 3-4. The range of peak scores shown varies from 0.16–0.66, which indicates good coverage and accuracy when considered against F-measure scores reported for CASP7 inter-residue contact predictions of 0.02–0.09 [26,32]. For all but two proteins tested, our predictor strictly improves upon the results of BETApro, with a median peak score of 0.33 versus 0.19. More importantly, partiFold provides more consistent results across all proteins, and maintains flattened curves, indicative of good separation between high probability contacts and noise.

The performance of 1K24 and 1QD6 can be directly attributed to their inclusion of extra-cellular structural components outside of the barrel (see Figure 3-3). Since our current model focuses only on the barrel fold of a TMB, extra β -sheets and α -helices can be missed, as in 1K24 and 1THQ, degrading performance (the latter much more strikingly due to its already short sequence). In 1QD6, a large number of 3_{10} and α -helical structures cap the β -barrel and partially interact with the β -sheet walls, creating a small environment inconsistent with our energy model and interfering with predictions. The results of a modified set of constraints that are meant to partially compensate for α -helices results in an improved peak score (0.30-0.40), but such intervention would require *a priori* evidence of such a configuration. We note that BETApro uses a more complicated (and less transparent model) that incorporate secondary structure annotations to identify these kinds of regions. Inclusion of this parameter is a subject of future research.

Worth final mention, consistent with prior predictors (including BETApro), our algorithm does

not yet model bulges in β -sheets, and suffers slightly in performance where bulges exist. However, of the proteins tested, only 8 of 76 β -strand pairs contained bulges (type C or W [27]). Further, across β -sheets in general, only 14% of paired strands have bulges, and of those, 90% have only a single bulge [31]. Therefore, the impact of bulges on the results in Figure 3-4 should be minimal. However, the possibility that our approach can aid in bulge discovery is also subject of ongoing research.

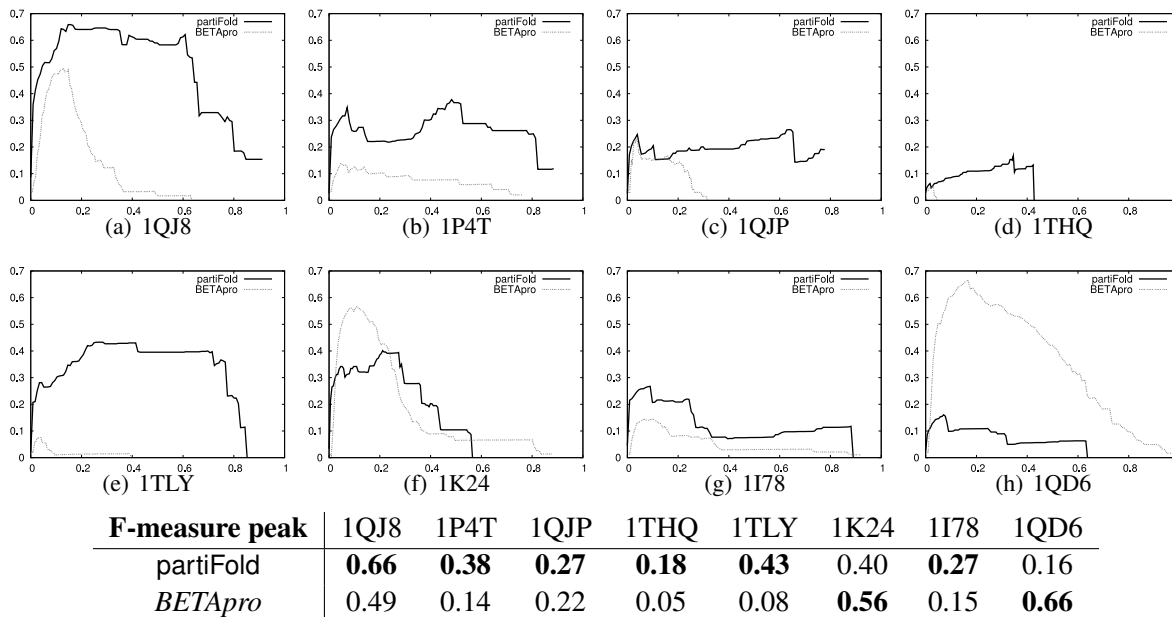


Figure 3-4: **F-measure accuracy scores of partiFold compared with BETApr**: Residue pairing contact F-measure scores (y-axis) comparing partiFold (black) and BETApr (gray) as a function of number of contacts predicted (i.e., all contacts with contact probability greater than any threshold p_t along x-axis). Bold entries in table show higher performance.

3.1.2 Residue flexibility prediction via contact probability profile

It is also possible to use ensemble predictions, such as stochastic contact maps, to predict per-residue flexibility and entropy. Interestingly, to a first approximation, this flexibility can correlate with the *Debye-Waller factor* (a.k.a. the B-value) found in X-ray crystal structures [152]. This demonstrates an important purpose of computing the Boltzmann partition function: to provide a biologically-relevant grounding for the prediction of experimentally testable macroscopic and microscopic properties. Predicting residue B-values is important because it roughly approximates the local mobility of flexible regions, which might be associated with various biological processes, such

as molecular recognition or catalytic activity [162]. In our context, flexible regions are strong candidates for loop regions connecting antiparallel TM β -strands that extend either into the extracellular or intracellular milieu.

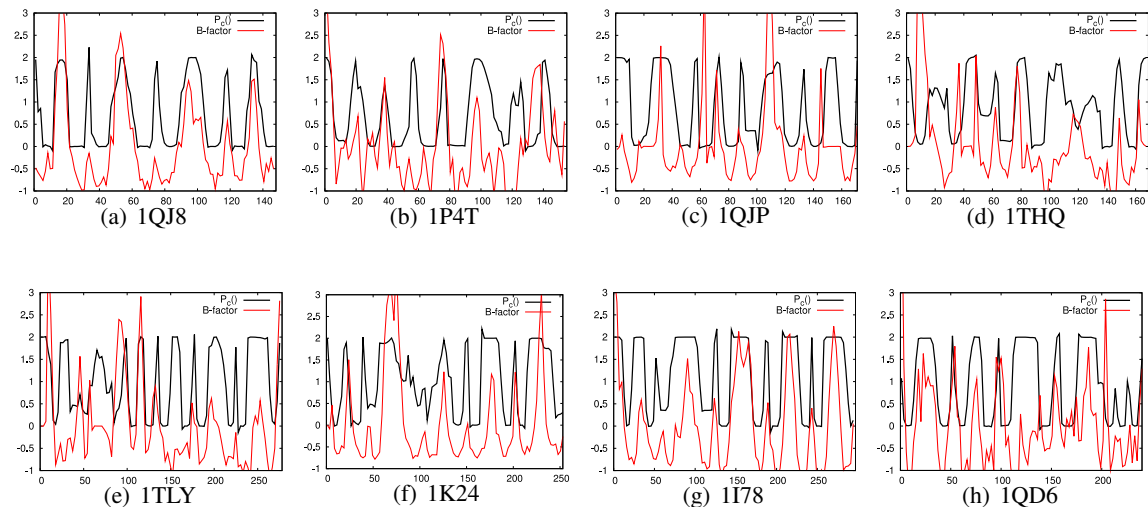
We demonstrate our per-residue flexibility predictions on the 8 TMB proteins tested in Figure 3-5. We define the *contact probability profile* of every amino acid index i in a TM β -barrel to be $P_c(i) = 2 - \sum_{j=1}^n p_{i,j}$, and in compare this against the normalized B-value, $B_{norm} = \frac{B - \langle B \rangle}{\sigma}$, a ratio commonly used for such a comparison [162]. Since a residue may be involved in two contacts in a β -sheet the value of $P_c(i)$ can range between 0 and 2 where higher values indicate a greater chance for flexibility. Similarly, residues with a positive B-value are considered flexible or disordered while others are considered rigid.

Computing the cross-correlation coefficient between the P_c and B-value of our test proteins, we find that partiFold compares well against PROFBval [162], a leading algorithm tuned specifically for B-value prediction. In fact, the more generally applicable partiFold method improves upon or matches 4 of the 8 TMBs. We have computed the per-residue contact entropy (defined as $S_i = \sum_{j=1}^n -p_{i,j} \log(p_{i,j})$) for the same test proteins and found similar results.

3.1.3 Whole structure prediction via Boltzmann sampling

Finally, we demonstrate that ensemble approaches to structure prediction can often characterize protein structure better than classical minimum folding energy techniques. To do so we perform stochastic conformational sampling (see Section 2.2.4) to explore the structural landscape defined by the Boltzmann partition function, By clustering a large set of full TMB structure predictions, a small distinguishable collection of unique conformations are exposed. In this set of clusters, we show in Figure 3-6 that some individual clusters tend to match the X-ray crystal structure better than the single minimum folding energy (m.f.e.) structure alone. This supports evidence [218] that the native state may be best described through sampling.

In this examination we sample 1,000 TMB structures and group them into 10 clusters according to hierarchical clustering. Similar to earlier methods [54], for each cluster we designate a centroid representative conformation that is chosen as the structure with the minimal total distance to all other structures in the set. To facilitate this clustering, we introduce a metric of *contact distance*: $d_c(S_1, S_2) = |\mathcal{C}_1| + |\mathcal{C}_2| - 2 \cdot |\{\mathcal{C}_1 \cap \mathcal{C}_2\}|$, where \mathcal{C}_1 and \mathcal{C}_2 are the sets of contact in S_1 and S_2 (which represents the minimal number of contacts to be removed and added to pass from S_1 to S_2 or vice versa).



Correlation coefficient	1QJ8	1P4T	1QJP	1THQ	1TLY	1K24	1I78	1QD6
partiFold	0.76	0.26	0.34	0.05	0.25	0.49	0.64	0.27
PROFBval	0.62	0.43	0.54	0.18	0.25	0.40	0.56	0.57

Figure 3-5: **partiFold** residue flexibility prediction accuracy compared with **PROFBval**: Contact probability profile (black, y-axis) and normalized B-value curve (red, y-axis) for partiFold as a function of residue index from left to right (x-axis). Due to the simple shape of most TMBs, experimental B-values tend to oscillate from high to low. Regions of B-value curves which are flat at 0 represent residues missing from the X-ray crystal structure (e.g., 1QJP residues 146-159, 1THQ residues 38-47, etc.). Bold entries in table indicate higher performance.

Figure 3-6 reports the coverage and accuracy of contact predictions for the largest cluster produced and for the cluster who's centroid structure best matches the X-ray crystal structure contacts (minimizing $d_c()$, labeled “best”), ignoring clusters with fewer than 15 samples. Both centroid scores and scores for the highest coverage and accuracy sample (“top sample”) within that cluster is listed. Comparing coverage and accuracy scores, surprisingly the centroid structures of both the largest *and* “best” cluster often outperform the scores obtained by the minimum folding energy structure. This is despite the fact that in five of the cases the “best” cluster is *not* the largest cluster produced (note 1THQ and 1I78). From this we see that the minimum folding energy structure does not always best describe the structure found by X-ray crystallography. This might even suggest that alternate conformations might be found in the Boltzmann distribution with high probability, although a more sophisticated energy model, including, for instance, an explicit term for the entire connecting loops, would be required to understand this result. In future work we intend to improve upon these simple clustering techniques and further explore these implications on TMB folding landscapes.

Protein	largest cluster				size	m.f.e.	
	centroid		top sample			cov.	acc.
	cov.	acc.	cov.	acc.			
1QJ8	0.65	0.67	0.78	0.82	375	0.65	0.58
1QJP	0.38	0.34	0.33	0.33	422	0.19	0.21
1P4T	0.20	0.18	0.41	0.39	309	0.13	0.14
1THQ	0.11	0.09	0.11	0.09	358	0.08	0.11
1TLY	0.32	0.33	0.32	0.34	303	0.36	0.40
1K24	0.15	0.17	0.53	0.58	428	0.09	0.08
1I78	0.17	0.24	0.23	0.32	373	0.17	0.13
1QD6	0.14	0.14	0.22	0.22	568	0.05	0.06

Protein	"best" cluster				size	m.f.e.	
	centroid		top sample			cov.	acc.
	cov.	acc.	cov.	acc.			
1QJ8	0.65	0.67	0.78	0.82	375	0.65	0.58
1QJP	0.38	0.34	0.33	0.33	422	0.19	0.21
1P4T	0.43	0.38	0.42	0.39	109	0.13	0.14
1THQ	0.20	0.16	0.20	0.16	40	0.08	0.11
1TLY	0.37	0.38	0.37	0.39	15	0.36	0.40
1K24	0.31	0.34	0.31	0.35	24	0.09	0.08
1I78	0.22	0.31	0.27	0.36	53	0.17	0.13
1QD6	0.14	0.14	0.22	0.22	568	0.05	0.06

Figure 3-6: **Coverage and accuracy of clustered partiFold predictions:** Contact predictions are compared against X-ray crystal structure. Centroid representative structure scores are given as well as the top performing sample in that given cluster. Bold numbers show the trend of improvement in the centroid structure’s coverage and accuracy over that of the m.f.e. structure. **Above:** Largest cluster produced when sampling 1,000 TMB structures. **Below:** “Best” cluster produced, as defined by the cluster containing the centroid conformation with the minimal $d_c()$, but no fewer than 15 samples.

3.2 Amyloid fibrils

Using the AmyloidMutants framework described in Section 2.3, we again validate the accuracy of our structural predictions by comparing against known amyloid fibril structural data. Unfortunately, very few amyloids have been structurally characterized, and for those that have been studied, the information obtained can be coarse and sparse (for example, H/D-exchange data). Further, contradictory results even exist from some amyloid fibril experimental studies. Therefore, rather than computing F-measure scores of residue/residue contact predictions, etc., as we did for TMBs, we describe our ensemble predictions for each amyloid fibril protein in detail on a case-by-case basis. By pursuing this detail we highlight the power of an ensemble predictor in revealing multiple

high-likelihood structures, and the ability to help guide further experimentation.

	A β	HET-s	Amylin	α -syn	Tau
sequence length	42	73	37	140	441
correct β -regions	2 of 2	4 of 4	3 of 3	5 of 5	7 of 8
false-pos. β -regions	0	0	0	2	2
percent sens./spec.	100/100	95/95	70/91	81/95	68/95
SOV measure	100	90	97	62	62

Table 3.2: **Summary of amyloid fibril secondary-structure prediction results:** For the five proteins tested we list whether β -strands were correctly predicted, per-residue sensitivity and specificity, and the SOV accuracy measure [216].

3.2.1 Validation of amyloid fibril predictions

We demonstrate below that AmyloidMutants’s ensemble structure predictions offer state-of-the-art secondary-structure prediction accuracy while further describing interesting super-secondary structure and ensemble characteristics. We have chosen to evaluate our predictions on experimental data for five of the best studied WT amyloid proteins: A β [115, 143] (39-42aa), HET-s [205] (73aa), amylin [96, 114] (37aa), α -synuclein [80, 188] (140aa), and tau [129, 192] (htau40, 441aa). This set covers pathogenic and functional amyloids found in nature for which there are a number of published structural experiments, including NMR secondary structure chemical shift and H/D exchange data. The ability to accurately predict the structure of such peptides could potentially help elucidate how native amyloid-related processes (such as biofilm formation) impact cellular function, and allow for targeted experimentation or therapeutics.

Table 3.2 provides a quick summary of our secondary structure prediction accuracy. For the five proteins tested, AmyloidMutants predicts β -sheet regions that agree with published, experimentally-derived structure models in 21 of 22 cases — for a per-residue secondary-structure classification sensitivity/specificity of of 82%/95% and an average SOV score of 82 [216]. For comparison, the best of the available full-length amyloid prediction tools [8, 69, 120, 178, 183] produced a maximum per-residue classification sensitivity/specificity of 42%/90% (Zygggregator) on this same comparison. Figure 3-7 illustrates the specific differences in secondary structure prediction accuracy across all proteins and predictors tested (including the HET-s homolog FgHET-s, see below). We note that while other amyloid predictors compared may provide additional non-structural output, such as toxicity, our focus is on each tool’s simple sequence-to-secondary-structure predictive abilities. Furthermore, β -sheet per-residue secondary-structure classification is used to compare tools since

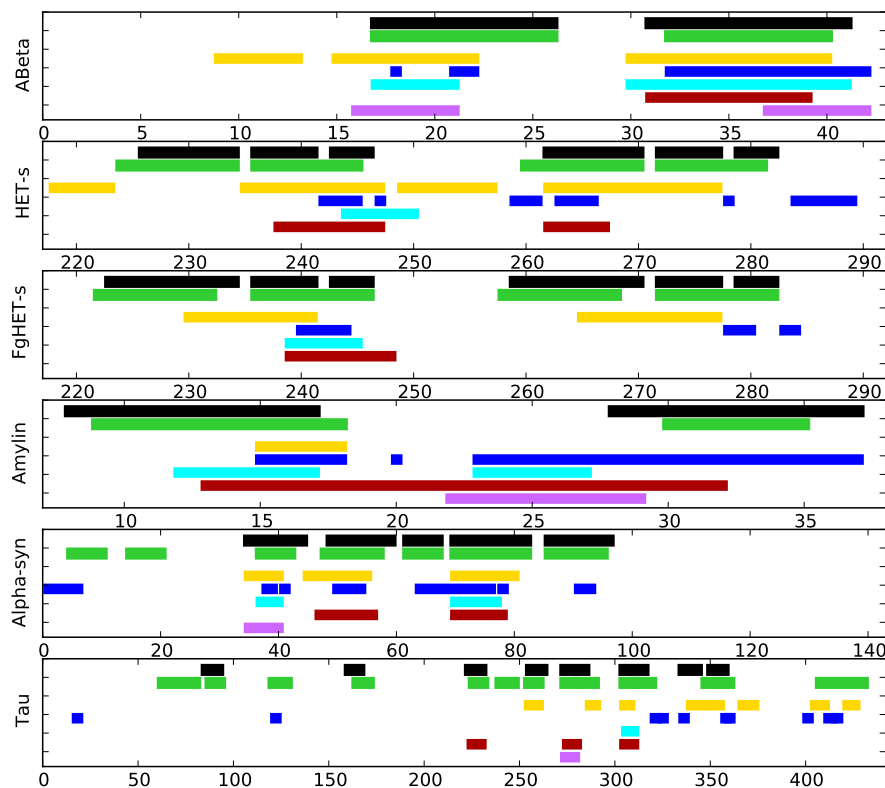


Figure 3-7: **Secondary structure prediction accuracy of AmyloidMutants and others:** AmyloidMutants per-residue β -strand assignments indicate amyloid core regions, comparable with existing per-residue amyloidogenicity predictors. AmyloidMutants predictions (*green*) outperform those tools available for testing when using their default settings and thresholds. BETASCAN (*gold*) [8], Zyggregator (*blue*) [178], TANGO (*cyan*) [69], PASTA (*red*) [183], and Waltz (*purple*) [120], when compared against experimental structure models supported by NMR, H/D-exchange, and mutational analysis (*black*) [114, 115, 129, 188, 205, 206]. Note, the BETASCAN, Zyggregator, TANGO, and PASTA tools most closely match our approach’s ability to predict full-length per-residue amyloidogenicity, whereas Waltz aims to predict short hotspots that could specifically adopt a steric zipper.

this is a common output that all other tools report; however our approach can provide a more rich prediction including residue/residue interactions.

For each protein mentioned, AmyloidMutants was run on each sequence for all three schemas \mathcal{P} , \mathcal{A} , and \mathcal{S} , with the schema that agreed best presented. Predicted ensemble results are derived from a stochastic sampling of whole structures out of the Boltzmann statistical mechanical ensemble. Populations of similar structures are identified and separated via PAM clustering, which takes as input the number of clusters, and relies on a distant metric that combines secondary structure, energy score, hydrogen bond registration, coil location, and β -strand overlap. For each cluster a

medioid¹ is selected to represent that population. Although rough computational tests can be applied to evaluate the schema fitness (see Section 2.5), in a typical real-world scenario (and as has been applied thus far), an uncharacterized amyloid sequence is predicted using all schemas, and results are compared against the body of existing experimental data or used to guide further disambiguating experimentation. No restrictions are placed on the location or size of structural elements with the exception of individual β -strand lengths, which is fixed to a range of 6 to 12 residues for efficiency purposes, and can vary within a single structure (except when otherwise noted).

3.2.2 Case analyses of well characterized amyloid proteins

Although per-residue β -sheet secondary-structure classification is used to compare the accuracy of our approach against existing tool, ensemble models can provide a much more rich prediction including residue/residue interactions. A detailed analysis is provided for each protein describing these added benefits, along with a demonstration of how ensemble predictions can help identify alternate fibril conformations (which align with published experimental data).

Amyloid Beta (A β):

Our modeling tool accurately predicts two distinct structural populations that recapitulate the findings of multiple NMR models of the A β peptide, a 39 to 42 residue product of human APP cleavage associated with Alzheimer's disease [92]. An ensemble analysis of A β is particularly poignant as it has many known isoforms (A β ₁₋₄₀, A β ₁₋₄₂, A β _{1-40/D23N}, A β _{1-40/E22Q}, etc.) and subsequences (A β ₁₆₋₂₂, A β ₁₁₋₂₅, etc.) that have been reported to form a diverse range of fibril structures, including strain polymorphisms within the same sequence [144]. Our tool predicts the experimentally observed structure of two possible A β conformations, recapitulating two distinct experimental models of the peptide based on NMR, H/D-exchange, and mutational analysis [115, 143]. After clustering, the highest likelihood medioid structure nearly identically matches the latter of these two models [115] (Figure 3-8(a)), including β -strand positions, interior/exterior sidechain orientation, and the inter-peptide parallel hydrogen bonding registration. This cluster accounts for 55% of the ensemble. Interestingly, the second highest likelihood medioid exhibits a clear shift in one of the β -strand regions and aligns very closely with the earlier NMR model [143]. This cluster is more heterogeneous, including many other structural arrangements, and accounts for 39% of the ensemble.

¹the sampled structure closest to the centroid cluster center

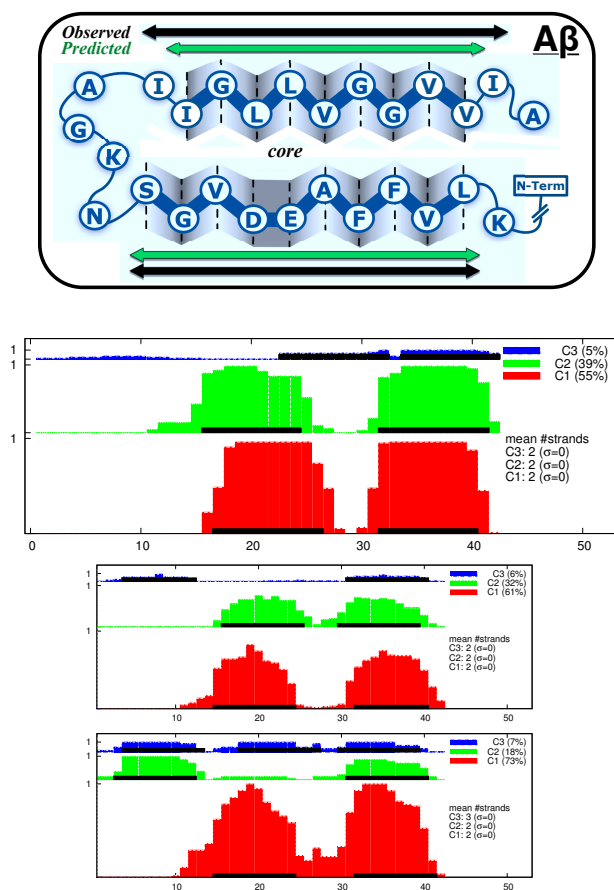


Figure 3-8: **AmyloidMutants A β predictions compared with experimental data:** (a) Diagram depicts β_{1-42} β -strand in gray, residues in blue (with in/out orientation), and β -sheet/ β -sheet packing as one β -strand above another, packed residues facing center. Predicted structure (*green arrows*) mirrors NMR structure [115] (*black arrows*), including most packing orientations. Predicted kink likely occurs because schema does not account for known D23/K28 salt-bridge. (b) Schema \mathcal{P} ensemble predictions of A β_{1-42} , clustered into three populations. For all such graphs, structure cluster populations are separated vertically into colors red, green, and blue, where each cluster's vertical size reflects its energetic weight within the ensemble (given as a percentage in the key). The x-axis indicates residue sequence position while the y-axis indicates the energetically weighted frequency of β -structure at that position within each specific cluster. For example, full bar heights indicate that all structures within the cluster contain a β -strand at that residue position, while a half bar height suggests that half of the energetic weight of that cluster contains structures with a β -strand in that position (i.e., the sum of the energies of the conformations with a β -strand in that position totals half the sum of all conformation energies). The single medioid structure for each cluster has its β -strand regions indicated by a black bar. The average number of β -strands within all structures in each cluster is also indicated. Results were attained by sampling conformations with a fixed β -strand length between 9–10. (c) Schema \mathcal{A} ensemble predictions of A β_{1-42} using the same parameters. (d) Schema \mathcal{S} ensemble predictions of A β_{1-42} using the same parameters.

Furthermore, recent experimental studies of A β conformational variation have shown that fibrils formed under quiescence and agitation differ, for instance, in the assignment of position 15 to

β -strand [144]. The predicted clusters also make this rough distinction: the larger cluster does not contain a β -strand at position 15, while the smaller does. Moreover, brain-seeded fibrils have exhibited spatial proximity between residues F19/I31, whereas unseeded *in vitro* fibrils do not [138]. AmyloidMutants also exhibits such a divergence in the predicted ensemble. Predicted β -strands place F19 on the interior (buried) side of the fibril in 89% of the ensemble, while a smaller, but still significant 19% of the ensemble contains a buried I31 — the combined case where both are buried makes up 16% of the ensemble. The predicted assignment of both F19/I31 as buried would allow the kind of spatial proximity observed experimentally.

Ensemble prediction of $A\beta_{1-42}$ were made using schema \mathcal{P} . We note that both schemas \mathcal{P} and \mathcal{S} (but not schema \mathcal{A}) are capable of predicting the exact published single-rung β -solenoid structure of $A\beta_{1-42}$ due to an intersection in the conformational space defined by each model. However, schema \mathcal{P} defines a much larger space of possible structures, and therefore was chosen to highlight the discriminative power of our scoring function.

To predict an ensemble of $A\beta_{1-42}$ structures with β -strands of length 6 to 12 using schema \mathcal{P} , we first calculate seven sets of sampled structures, fixing a different β -strand length to each, and determine cluster populations across all of those structures combined. This is done to help ensure a well-distributed coverage of β -strand lengths with fewer samples — predictions that allow the length to vary between 6 and 12 within a single execution explore variations in kink location more often than variations in length. This is due to the two-fold dimensionality increase caused by kinks, and the fact that changes in kink location (and the resulting change in β -strand residue orientations) often can induce a smaller energy difference than changes in β -strand length. We note, however, that single predictions allowing the strand length to vary between 6–12 do qualitatively agree with our iterative approach. From inspection of this data we find the most energetically favorable cluster predominantly contain β -strands of length 9 to 10. Predicting an ensemble of structures based on this length results in two major structural clusters, shown in Figure 3-8(b)), with the largest cluster's mediod structure containing β -strands at positions 17-26 and 32-39, and the smaller at positions 12-24 and 30-40. Although the larger cluster fills 55% of the ensemble and the small 39%, we note that the close similarity between these two clusters of structures may introduce error in calculating a specific percentage value.

For illustrative purposes we include predictions of $A\beta_{1-42}$ using schemas \mathcal{A} and \mathcal{S} (Figure 3-8(c) and Figure 3-8(d)) In these cases similar β -sheet interaction regions present themselves, although the specific β -strand residue/residue pairing can be quite different. This highlights a difficulty in

directly comparing schemas against one another.

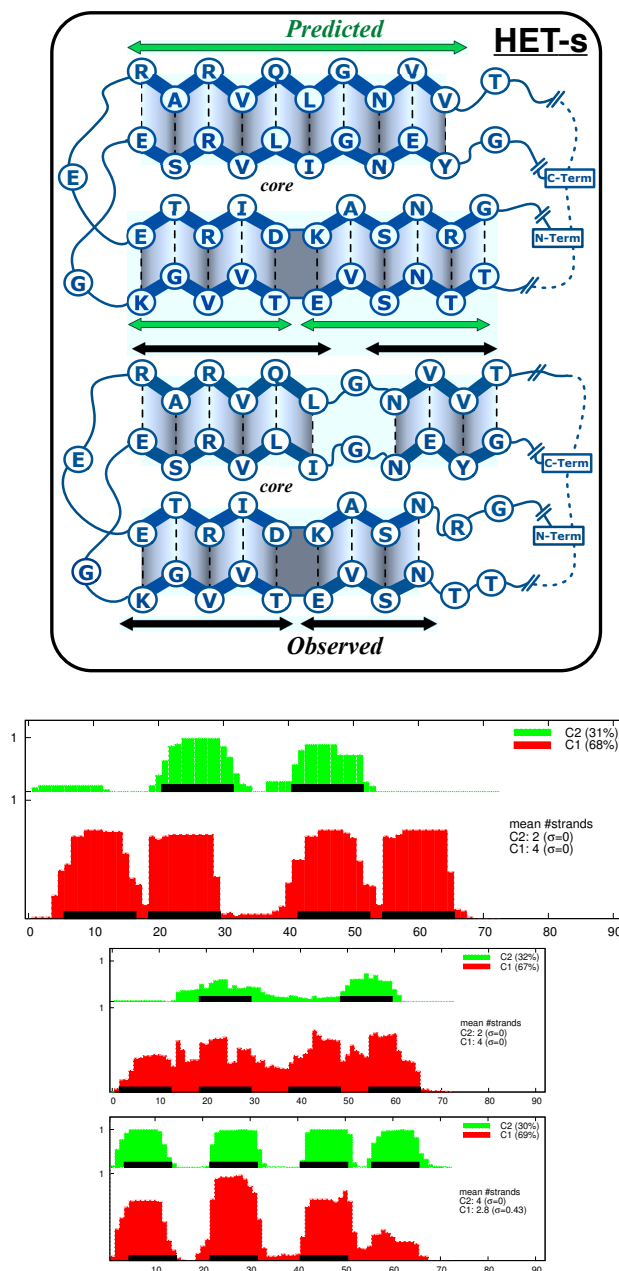


Figure 3-9: **AmyloidMutants HET-s predictions compared with experimental data:** (a) Amyloid-Mutants predictions of HET-s (*top, green arrows*) compared to NMR model [205] (*bottom, black arrows*) show near identical match, including residue orientations and kink location. Depicted similarly to Figure 3-8. (b) Schema \mathcal{P} ensemble predictions of HET-s, clustered into two populations, similar to above. As explained in Section 1, results were attained by sampling conformations with a fixed β -strand length between 10–11. (c) Schema \mathcal{A} ensemble predictions of HET-s using the same parameters. (d) Schema \mathcal{S} ensemble predictions of HET-s using the same parameters.

HET-s:

On the largest solved amyloid 3-dimensional structures to date, our ensemble predictions almost exactly reproduce the NMR model of the *Podospora anserina* HET-s prion [205]. In its prion form, this protein plays a role in heterokaryon incompatibility, and presents a much more complex fibril structure than that of A β , with 73 amino acids forming a well-ordered β -helix with two rungs per chain and four β -strands per rung. Of these four β -sheets, two are sequentially adjacent, differentiated only by a kink in the standard 'in/out' β -sheet sidechain orientation, while the other two are separated by a single glycine residue (black arrows in Figure 3-9(a)). This architecture is consistent with our 2-sheet kinked solenoid schema \mathcal{P} .

AmyloidMutants strongly predicts two possible structures, the most likely of which forms a two-rung β -solenoid that almost exactly mirrors the NMR model, including hydrogen bond registration, sidechain orientation, and kink location (green arrows in Figure 3-9(a), accounting for 68% of the ensemble). The lower likelihood conformation incorporates only a single rung, also matching a rung in the NMR structure. This strong predictive bias toward only two possible structures may relate to the observed conformational homogeneity of HET-s fibrils [46]. Achieving such high accuracy on this difficult β -structural topology supports our tool's use for analysis on a broad range of fibril types.

Similar to the procedure used for A β_{1-42} , to predict an ensemble of HET-s structures with β -strands of length 6 to 12 using schema \mathcal{P} , we first sample and cluster conformations with fixed β -strand length 6, 7, 8, 9, 10, 11, and 12. Inspection reveals that the most energetically favorable cluster of structures predominantly contains β -strands of length 10 to 11. Figure 3-9(b) shows ensemble predictions using this length parameter and the two clear structural populations: the larger "two-rung-per-chain β -solenoid/ β -helix" and the smaller a "one-rung-per-chain" β -solenoid/ β -helix."

Again, for illustrative purposes we include predictions of HET-s using schemas \mathcal{A} and \mathcal{S} (Figure 3-9(c) and Figure 3-9(d)). Here we see marked shifts in β -sheet interaction regions between schemas, along with changes in population distribution and β -strand residue/residue pairing.

FgHET-s:

Interestingly, recent experimental studies have partially characterized a distant homolog to *Podospora anserina* HET-s found in *Fusarium graminearum* [206]. Although FgHET-s exhibits only 38% sequence similarity, solid-state NMR and H/D-exchange data suggests an extremely similar β -

solenoidal structure as in PaHET-s. Despite the large difference in sequence, AmyloidMutants predictions agree very well with the FgHET-s structural model, including β -strand location, hydrogen-bond registration, sidechain orientation, and kink location. This is shown in Figure 3-10(a). Similar to HET-s, predictions were made using β -strand lengths of 10 to 11, with cluster results shown in Figure 3-10(b).

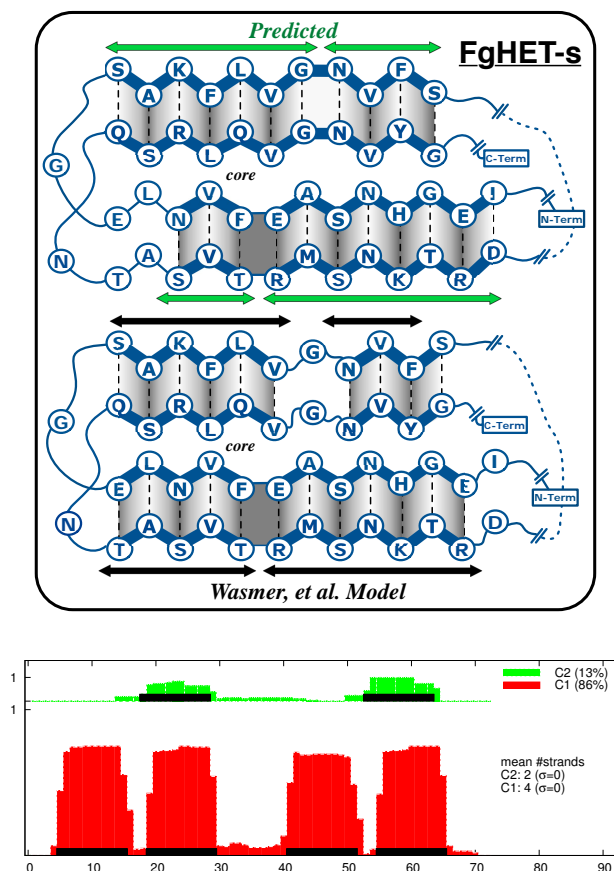


Figure 3-10: **AmyloidMutants FgHET-s predictions compared with experimental data:** (a) AmyloidMutants predictions of FgHET-s (*top, green arrows*) closely match a recent partial characterization by solid-state NMR and H/D-exchange [206] (*bottom, black arrows*). Depicted similarly to Figure 3-8, arrows highlight match, including residue orientations and the location of β -strand kinks. (b) Schema \mathcal{P} ensemble predictions of FgHET-s. Results clustered into two populations allowing β -strand length to range between 10–11.

Amylin:

In the case of human amylin, a 37-residue peptide hormone associated with type 2 diabetes, ensemble predictions accurately reveal two structural populations that each agree with competing experimental models. AmyloidMutants predictions indicate a 2-sheet β -solenoid conformation forming

80% of the ensemble, in agreement with recent NMR and microscopy results [114] (Figure 3-11(a)); and a much less likely 3-sheet serpentine model that aligns almost perfectly with an older model of amylin structure [96]. Interestingly the NMR and microscopy results identify an inter-protofibril interaction between Phe23 and Tyr37 — something beyond the scope of our schema. However, our β -solenoid predictions clearly separate into two distinct populations, one incorporating Phe23 into a β -sheet and one that does not. This highlights a benefit of an ensemble analysis: the existence of high-likelihood alternate structures may draw attention to an otherwise overlooked putative structural interaction.

The Boltzmann ensemble for Amylin was computed using schema \mathcal{S} because of this schema's ability to incorporate both 2-sheet β -helical structures as well as 3-sheet serpentine structures within the same conformational space. This allows a comparison between the energetic favorability of 2-sheet versus 3-sheet structures. Figure 3-11(b) presents the predicted ensemble when allowing β -strand length to vary between 6 and 12 residues long (since schema \mathcal{S} does not include kinks). Since residues 1-7 have been shown non-critical to fibril formation due to a disulfide bond between Cys2 and Cys7, we explicitly fix positions 2 and 7 as non- β -sheet-forming, effective throughout all computed structures within the ensemble (such point-wise constraints can be similarly be applied

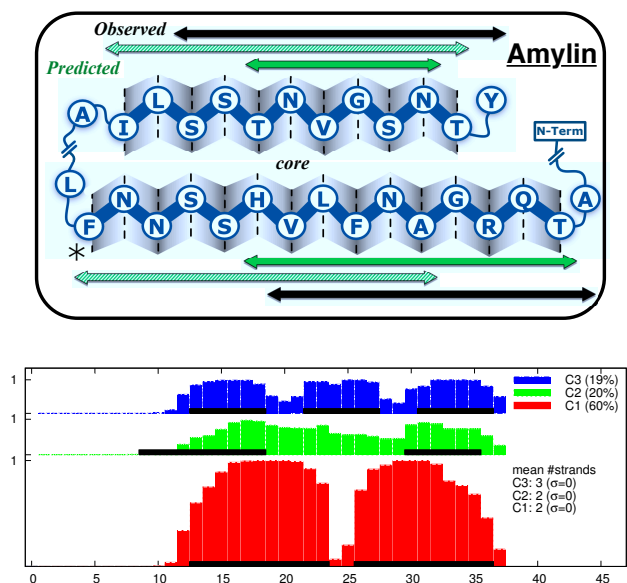


Figure 3-11: **(a)** The top two Amylin predictions (*solid, striped green arrows*) align well to NMR model [114] (*black arrows*). Predictions differ only by their inclusion of Phe23 (*) within β -sheet, a residue experimentally shown to form non- β -sheet inter-peptide interactions not considered by schema. **(b)** Schema \mathcal{S} ensemble predictions of Amylin. Results clustered into three populations allowing β -strand lengths to range between 6–12.

to schemas \mathcal{P} and \mathcal{A}). Clustered into three populations, 80% of the ensemble consists of 2-sheet β -solenoid structures with the remaining minority containing 3 sheets.

α -synuclein:

Substantial effort has also been undertaken to elucidate the amyloid structure of the 140-residue protein α -synuclein, whose amyloid deposits have been associated with Parkinson's disease [44]. The best current structural data [80, 188] suggests a solvent-protected fibril core between residues 30-110 containing roughly five β -sheet structures. AmyloidMutants ensemble predictions agree extremely well with these results, aligning all five β -sheet regions, and identifying other important experimental observations such as a β -sheet break around residues 67-68 (see Figure 3-12(a)).

Ensemble predictions of α -synuclein were performed using schema \mathcal{S} since this schema permits many β -sheets of differing lengths to pack together without the need for intra-peptide hydrogen bonding interactions. Structures were sampled, allowing β -strands to range in length from 6 to 12, and were clustered into two populations, shown in Figure 3-12(b). Although β -structure regions predicted within the fibril forming region of 30-110 show excellent agreement with experimental observations, two false-positive β -strand structures are apparent around positions 5-10 and 15-20.

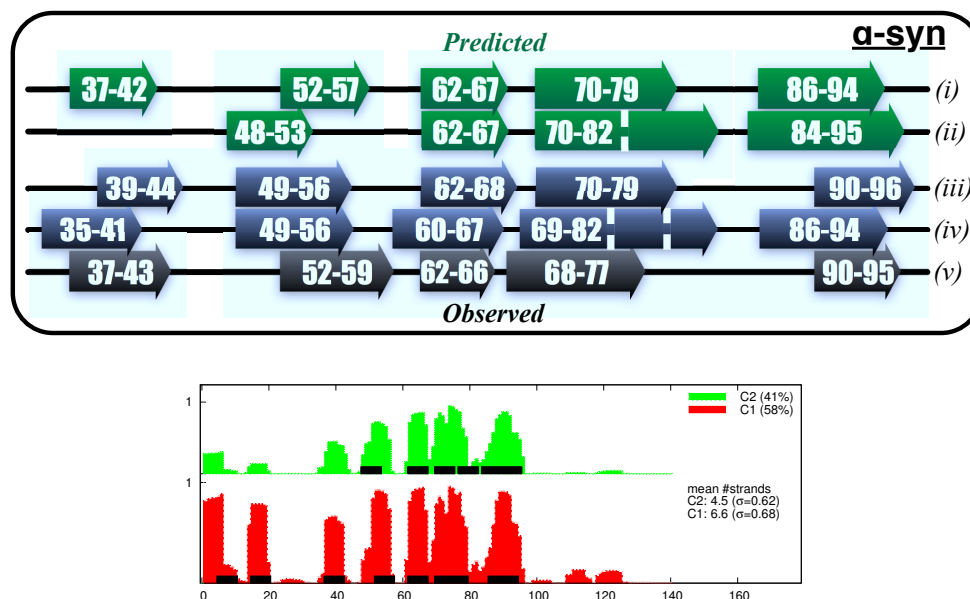


Figure 3-12: **AmyloidMutants α -syn predictions compared with experimental data:** (a) The top two α -synuclein predictions (*i,ii*) agree very well with H/D exchange data (*iii,iv*) and the NMR model (*v*) [80, 188]. (b) Schema \mathcal{S} ensemble predictions of α -synuclein. Results clustered into two populations allowing β -strand lengths to range between 6–12.

The likely reason for the prediction of amphipathic β -strands in this region is because this disordered N-terminal is believed to favor a lipid-binding amphipathic α -helix structure [191].

Tau (τ):

Human tau protein (htau40) is a natively unfolded microtubule-associated protein that can aggregate into tangled fibrils in Alzheimer's disease. NMR studies have shown this 441 amino acid long amyloid to form a mixture of up to 8 transient β -sheet regions [129], with two specific β -strands necessary for fibril assembly [192] (positions 306-311 and 275-280).

Predicted β -sheets align very closely with these observed regions in 7 of 8 cases, as shown in Figure 3-13(a). Moreover, AmyloidMutants identifies the two hexapeptides experimentally observed necessary for assembly [192] by predicting their β -strand interactions as having the strongest score.

Ensemble predictions of the 441 residue long Tau were performed using schema \mathcal{S} for similar reasons as for α -synuclein. Structures were sampled and clustered into two populations, shown in Figure 3-13(b) (again, permitting β -strands to range from 6 to 12). The number of clusters were fixed to 2, although note that both clusters appear quite similar, suggesting only small β -strand registration variations across the ensemble, and an especially strong consensus on the large regions do not form fibril. Despite the high accuracy in predicting experimentally observed β -strand regions, β -sheet structure is incorrectly predicted around positions 121-128 and 408-430, which overlaps with observed α -helices (which the schema does not incorporate) similar to α -synuclein. However, overall, the sensitivity and specificity of our predictions over such a long sequence is striking when compared to existing methods.

To further analyze AmyloidMutants's identification of the two hexapeptides experimentally observed as necessary for assembly, predictions were performed with the energy model artificially biased against β -sheet formation, implemented via a simple scaling factor. This acts as a crude method to identify peptide regions which have the highest likelihood of forming β -structure. Figure 3-13(c) shows these biased clustered ensemble predictions. In this case, the median of the largest cluster predicts strands at positions 274-279 and 305-310, in good agreement with experimental evidence that these regions are crucial in initiating fibril assembly [192].

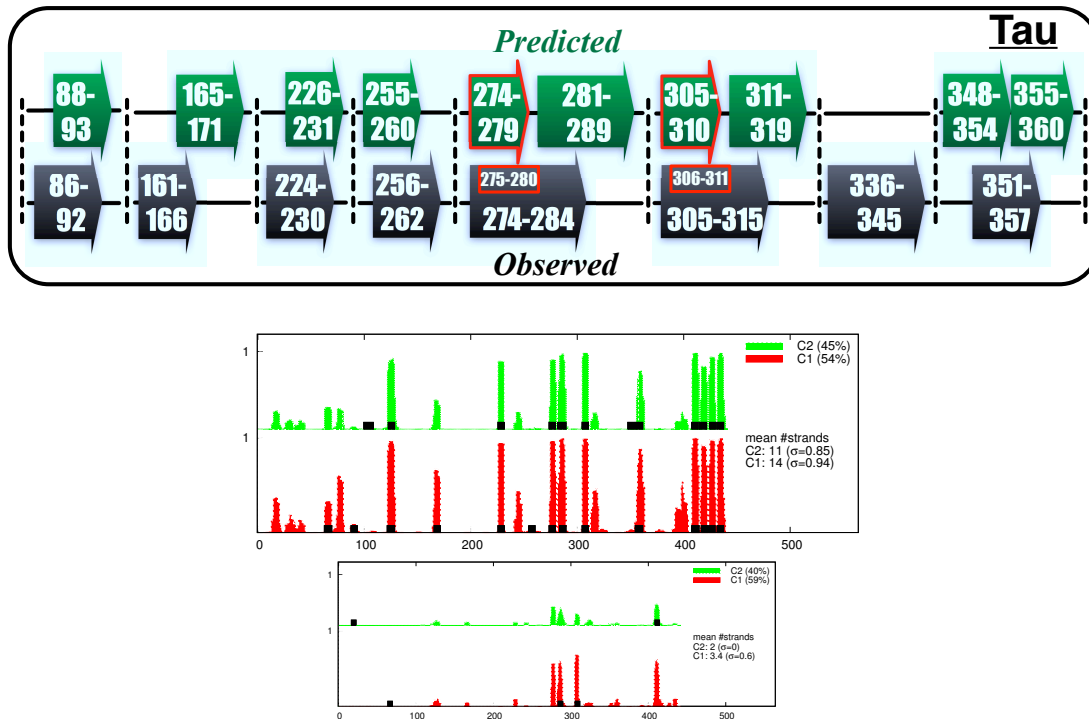


Figure 3-13: **AmyloidMutants tau predictions compared with experimental data:** (a) Tau predictions identify 7/8 β -regions observed experimentally [129]. The highest AmyloidMutants scores (red boxes) specifically identify regions 274-279 and 305-310, positions believed crucial to fibril nucleation [192]. (b) Schema S ensemble predictions of Tau. Results clustered into two populations allowing β -strand lengths to range between 6–12. (c) Sampled conformations when artificially biasing against β -strand formation within the energy model.

3.3 Evaluation of energetic stacking pair potentials

Section 2.4.2 introduced a new form of statistical potential involving *stacking pairs*. While this improves specificity by calculating energetic scores for environmentally near amino acid 4-tuples, the increase in dimensionality (i.e., a table requiring 20^4 entries) necessitates the use of a reduced residue alphabet. Here we justify our choice for using the Wang & Wang 5-letter reduced alphabet [202] for the amino acid stacking pair potentials.

Following preliminary study, five alphabets were selected to represent a broad range of residue classifications, and their predictive abilities were fully tested using partiFold on our available TMB protein structures. To illustrate this we present results for the protein 1QJ8 in Figure 3-14. Pairwise energy parameters from Waldispühl et al. [195] are also included for comparison.

These plots show that the Wang & Wang alphabet offers the highest combination of coverage and accuracy for contact prediction, though a few other alphabets offer decent accuracy for a smaller

coverage. Interestingly, for transmembrane β -barrels the majority of stacking pair potential alphabets outperformed the non-stacking pair potentials used by Waldispühl et al. [195], supporting the hypothesis that stacking pairs better describe this β -sheet energy potential. Experimentation on other TMB proteins revealed varied results, though the Wang & Wang alphabet tends to remain the best candidate. One reason for this may be the biophysically important segregation of aspartic and glutamic acids into their own residue classes, reducing stacking charge clashes.

Reduced alphabet	Group 1	Group 2	Group 3	Group 4	Group 5
Wang & Wang 5-letter (WW5)	CMFI LVWY	ATH	GP	DE	SNQ RK
Wang & Wang Variant 5-letter (WWV5)	CMFI	LVWY	ATGS	NQDE	HPRK
Chemical differentiation 4-letter (Chem4)	IVL	FYWH	KRDE	GACS	TMQNP
Li 4-letter (Li4)	CFYW	MLIV	GPA TS	NHQE DRK	-
C Murphy 4-letter (Mur4)	LVI MC	AGS TP	FYW	EDNQ KRH	-

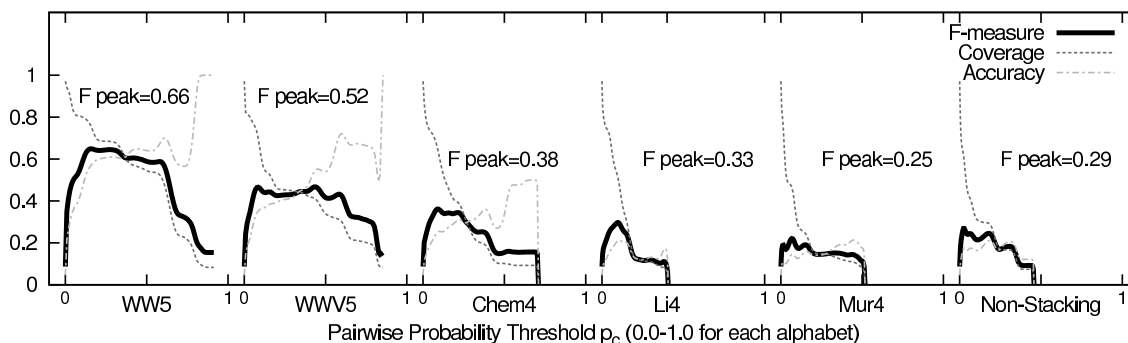


Figure 3-14: **The effects of reduced alphabet selection on TMB accuracy:** Above: Amino acid groupings for reduced alphabets selected and tested. Below: Smoothed F-measure/coverage/accuracy plots for the 1QJ8 protein across five reduced alphabets, plus the non-reduced, non-stacking energy potential previously used in [195] for comparison.

Chapter 4

Mutational landscape analysis

4.1 Goals and overview

Chapter 2 described our ensemble approach for creating informative models of protein structure landscapes. In this chapter we expand upon this concept and introduce *mutational landscapes* — constructions that identify energetically likely sequence/structure states across a defined ensemble. Thus, a mutational landscape predictor extends the dimensionality of structural ensembles to include possible sequence variation within each state. This technique enables the exploration of mutations that stabilize, reconfigure, or destabilize conformational forms, identifying sequence determinants of structural heterogeneity. Further, this allows for analyses of energetically-favorable mutational pathways in sequence space or the study of networks of phenotypically neutral variation, foundational concepts in population genetics [102].

The exploration of mutational landscapes is an important step in understanding differences between related structural states (such as amyloid schema topologies), how mutational variants arise in the wild, and to elucidate evolutionary relationships between related proteins. This capability depends on a thermodynamic characterization of all points within the sequence landscape, and is necessary for the discovery of non-additive functional relationships between sequences and conformational epistasis [135]. Further, the utility of such an approach has been demonstrated for RNA [196], where mutational landscapes have been proposed to model the secondary structures and the structures of k -neighboring sequences. This made possible the prediction of putative deleterious mutations in RNA viral pathogens. This approach inspires our work, however the differences between RNA and protein modeling require notably different techniques.

Recall, to predict the Boltzmann partition function of a structural ensemble, we compute \mathcal{Z}

according to $\forall s, \mathcal{Z} = \sum_s e^{-E_s/RT}$, given temperature T , the Boltzmann constant R , and a Boltzmann-distributed energy score E_s for every conformation s . To extend the notion to analyze sequence/structure ensembles (or mutational landscapes), we redefine the partition function \mathcal{Z} as $\forall \omega, \forall s, \mathcal{Z} = \sum_{\omega} \sum_s e^{-E_s/RT}$, given sequences ω and structures s . This encodes not only statistical variations in protein structure, but variations in protein sequence, distributed according to the energetic likelihood of that sequence’s conformations. With this one can not only predict the most energetically favorable structure and sequence assignment, but a single quantitative energetic score can be used to measure the difference between two sequences, between two structures, or between both. Note, there is no explicit cost for inserting a mutation; the mutated sequence residues simply impact the energetic score of sequence/structure states. The inclusion of sequence comparison scores within the prediction are the subject of simultaneous alignment and folding techniques, introduced in Chapter 6.

For reasons described below, we have focused our efforts on the design of an algorithmic framework for amyloid fibril mutational landscape prediction. However, many of the same principles can be easily applied to our methods for transmembrane β -barrel modeling (Section 2.2), although requiring more significant modification to our TMB representations than that of amyloid fibrils (Section 2.3).

At face value, the ability of most proteins to form a characteristic cross- β -sheet amyloid structure *in vitro* [57] seems at odds with the relatively small number of amyloid forming proteins that have been identified *in vivo*, and the apparently high sequence dependence some amyloids show when compared against sequence homologs. Moreover, the existence of both beneficial functional amyloid sequences, and putatively pathogenic “misfolded” amyloid proteins suggests a more complicated sequence/structure relationship than is found in standard protein folding models. The power to accurately predict amyloid structure from sequence, and to fully characterize the amyloidogenicity of an entire mutational landscape provides insight into this problem by identifying recurring sequence motifs, coarse 3-dimensional residue arrangements, and putative mutational pathways linking the sequences of known amyloid structures. The immediate impact of this could improve our ability to identify amyloid structures from genomic data alone, to better understand familial mutations that intensify pathogenesis in diseases such as Alzheimer’s, to predict the interaction strength of fibril regions that may be involved in nucleation, and to enable targeted peptide design to alter fibril structure or inhibit fibril formation.

4.2 Amyloid fibril modeling

In this section we describe the ensemble algorithmic framework for modeling amyloid fibril mutational landscapes. This work has been implemented as part of the publically-accessible web-based tool AmyloidMutants¹. Since our technique for calculating mutational landscapes is built on top of our earlier ensemble approach (Section 2.3), many of the algorithmic specific will not be reproduced in this section. Therefore, we only treat the specific extensions and new methods required to additionally model mutational variation.

4.2.1 Representing mutational landscapes

To represent amyloid fibril sequence/structure ensembles we extend the notion of amyloid schemas introduced in Section 2.3 to include sequence variation. Such a prediction of sequence/structure ensembles has long been considered computationally intractable since the number of states in the ensemble doubles for every point mutation introduced. However, recent work in RNA has shown that combining sequence and structure information within a single memoization table can enable the calculation of small sequence/structure ensembles [196]. This approach enumerates the energetic landscape of sequence neighbors with up to k base-pairs mutants (for small values of k since there are $\binom{N}{k} \cdot 3^k$ such sequences of length N). Unfortunately, using this approach on proteins is prohibitively expensive due to the size and complexity of protein structure, and the larger alphabet of mutant possibilities (i.e., amino acids instead of base-pairs).

To model the mutational landscape of an amyloid fibril, we again define generative schemas that restrict the exponential set of sequence/structure states. These schemas, however, are defined in two parts: (1) a recursive encoding of structure space, including the same features listed in Section 2.3 (e.g., the incorporation of *a priori* knowledge), and (2) a protocol giving a list of all allowed mutations of the input protein sequence. Therefore, this new ensemble contains the same conformational states as before, only each structure has multiple energetic scores, one for every potential sequence variant, resulting in a 2-dimensional landscape covering both sequence and structure. Figures 4-1 and 4-2 depict this concept.

Sequence space is defined as an explicitly enumerated set of allowed mutations off a base sequence, per-sequence-position, per-residue. For example, one mutation in this protocol might specify “position index 10 can either be *Ala*, *Leu*, or *Val*.” Similarly, mutational assignments can be

¹Available at <http://amyloid.csail.mit.edu/>

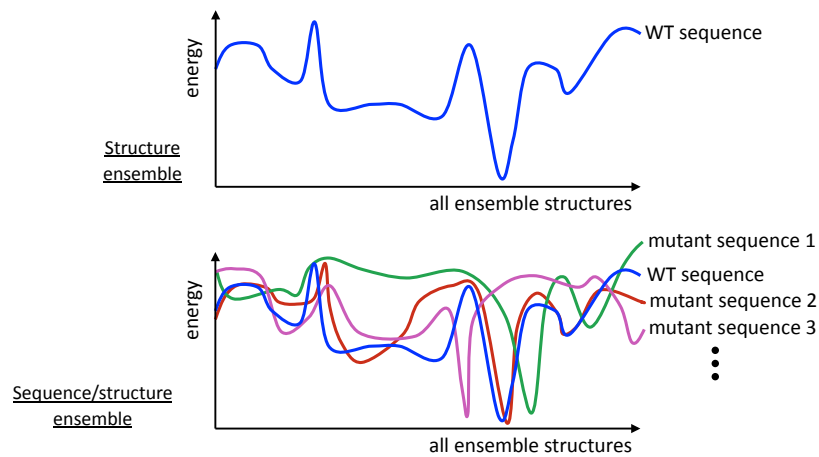


Figure 4-1: **Mutational landscapes add sequence dimension to structure ensembles:** Structural ensembles introduced in Chapter 2 calculate an energy score for all conformations within a schema given a single (“WT”) sequence. Sequence/structure ensembles calculate an energy score for all conformations within a schema as well as an energy score for all sequence neighbors permitted by the mutational protocol.

grouped to allow the sequence ensembles to include any set of sequences, for example, “position index 10 and 20 are both *Ala* or position 10 and 20 are both *Leu*, but do not consider the case where position 10 is *Ala* and position 20 is *Leu*.” Implicit mutations are not considered, and, presently, deletions and insertions remain a subject of future research since they would require significant changes to structural ensemble representations.

The explicit specification of both index and allowable mutant residues avoids an exponential computation, as a complete landscape would require 20^N residue permutations in a sequence of length N . From an implementation standpoint, this kind of schema information could be defined by a user in a textual format following a fixed convention. This is what is done in *AmyloidMutants*, where a text file is accepted at runtime. Accordingly, since the protein sequence is also known at runtime, short-hand definitions can be supported and expanded dynamically, such as “all it *Val* can be *Val* or *Ala*.”

4.2.2 Computing the partition function and Boltzmann distribution sampling

To compute the partition function of a sequence/structure schema we follow a similar recursion as shown in Section 2.3.3, and expand potential substates for each rule to include sequence variations. However, since our encoding relies on an explicit mutation protocol, we do not need to expand *all*

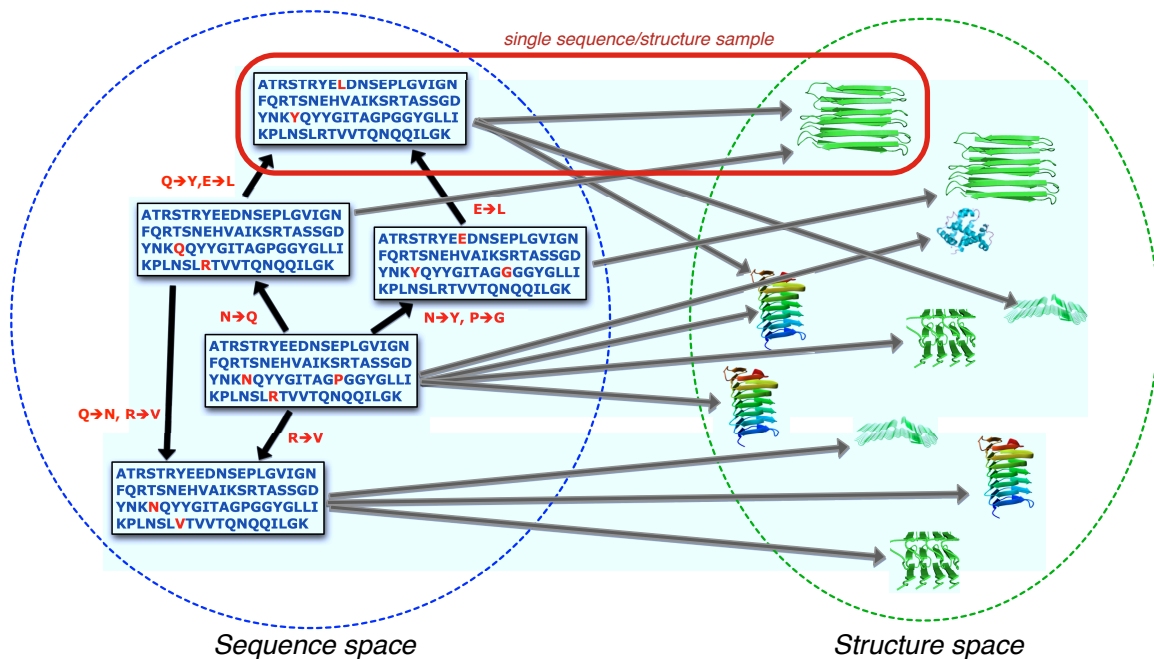


Figure 4-2: **Illustration of relationship between sequence/structure ensemble samples:** A mutational landscape can be computed efficiently by exploiting the overlap between related mutant sequences and related low-energy structures.

structural substates to account for different sequences, but only those that overlap specified mutation sites. This reduces the dimensionality of the problem considerably. Figure 4-3 illustrates the effect of this expansion on the *M-rule* from Section 2.3.3, Figure 2-9.

The definition of our ensemble recursions and dynamic programming tables are therefore modified to accept as a parameter this substate expansion mapping. For any given mutation protocol, an analysis is initially performed on the input sequence to determine which substates require sequence variation expansion. A separate bitvector records this analysis, and remaps potential sequence mutants to integer permutations that index precomputed energetic scores for all substates (as mentioned in Section 2.3.6). However, the algorithmic complexity of this new recursion still depends on the definition of the mutation protocol, and can become effectively exponential if a highly permissive protocol is chosen.

Sampling is performed similarly to that described in Section 2.3.4, expanding state space to include the predetermined sequence expansions when encountered by the recursion.

4.2.3 Runtime optimizations

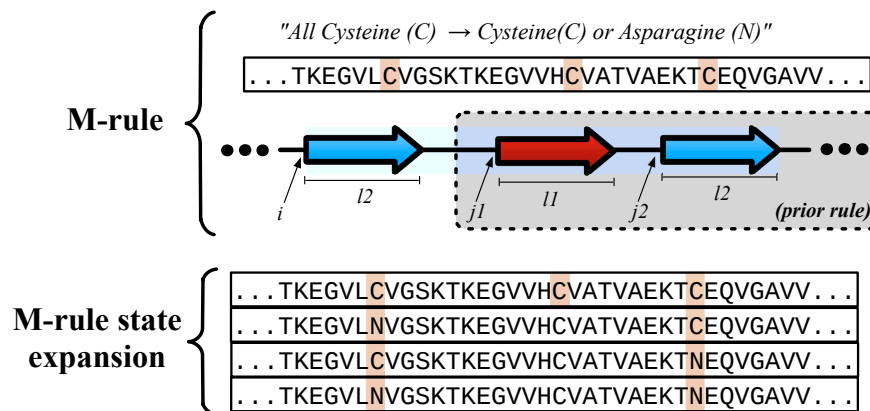


Figure 4-3: **Example mutational landscape state expansion, *M*-rule from schema \mathcal{A} :** Given the mutational protocol “All Cysteines in the sequence can be Cysteine or Asparagine,” the application of the *M*-rule to the sequence shown above expands to compute only 4 new sequence states instead of 8 (assuming no packing interaction scores, see Section 2.4). This is because the second cysteine does not contribute to the energy of these states and need not be expanded.

Techniques for implementing an efficient calculation of a sequence/structure ensemble partition function highly resemble techniques for structure-only ensembles described in Chapter 2. However, the additional algorithms required to map sequence variation to substate expansion are the ultimate performance bottleneck in mutational landscape prediction. Optimization of this mapping function is essential since it is encountered (and computationally evaluated) at every step of the recursion and is solely responsible for decreasing the effective number of states within the ensemble (reducing the memoization table size). Moreover, if the effective number of states is not reduced enough, the partition function may only be calculable via heuristics (Section 2.3.7).

As mentioned, this mapping is derived before the recursive traversal of substructure states, pre-computing which substates will require expansion and saving this in a custom data structure based on integer permutations. Such permutations allow a fast iterative search over sequence variants during evaluation of the *C*-rule, *M*-rule, or *N*-rule, enabling pooling of memory local substates temporally or in accordance with memory hierarchies, and allowing the removal of redundant computation. Further instruction and data ordering can be optimized at the level of recursion index selection. Thresholding, parallelization, and hash table optimizations listed in Section 2.3.6 can also be immediately applied, as well as the heuristics described in Section 2.3.7.

Chapter 5

Evaluation of mutational landscape prediction

In this chapter we validate the accuracy of our mutational landscape algorithms for amyloid fibrils, and demonstrate the scientific utility of this approach through its application to open biological questions in amyloid structure. While Chapter 3 demonstrated the accuracy of our ensembles structural predictions, there is perhaps a greater value in the ability to discover which mutations effect a change in amyloid fibril structure(s), what that change is, and to calculate a meaningful, quantitative comparison between mutants. Results are generated using the AmyloidMutants tool and experimental studies were carried out with collaborators at the Whitehead Institute for Biomedical Research, MIT, and Boston University.

We first validate the sensitivity of our algorithm's super-secondary structural predictions by using AmyloidMutants to distinguish shifts from one conformation to another when point mutations are made. Specifically, in agreement with experimental observations, our tool identifies separate, incompatible amyloid conformations that are preferentially induced by WT A β and the A β Iowa mutant [186], as well as similarly distinct conformations resulting from WT and yeast-toxic mutant strains of HET-s [48]. We further validate sequence-level amyloidogenicity predictions by comparing AmyloidMutants mutational occupancy scores with HET-s/HET-S studies, A β scanning-mutagenesis work, and studies of A β multiple-residue mutants.

Finally, we use AmyloidMutants to probe the amyloidogenic relationship between chemically similar residues such as *Asn* and *Gln*, revealing a specific *Asn*-sensitivity in the HET-s sequence. Moreover, we describe a study investigating the structural properties of the *E. Coli*. curli proteins

— important amyloid fibril proteins essential to biofilm formation. With the goal of creating an effective therapeutic for biofilm inhibition, we performed structural studies of these proteins and designed a targeted mechanism for disrupting amyloid fibril polymerization.

5.1 Identifying conformational shifts in amyloids

A key benefit of our approach over existing amyloid structure modeling tools is the ability to predict super-secondary structure information and higher-order topologies, allowing the identification of one amyloid β -sheet conformation from another, even when they share the same secondary structure assignments. This is important as such structural changes can have a dramatic impact on oligomerization and nucleation rates [101], disease infectivity [186], and prion propagation [2]. Here we use this ability to identify potential alternate, distinct amyloid fibril conformations that arise in the A β familial “Iowa” mutation [186] and yeast-toxic mutants of HET-s [15, 48], highlighting consistencies with published experimental data.

5.1.1 A β Iowa mutant

Recent studies [186] suggest that A $\beta_{1-40/D23N}$ may form an antiparallel β -strand fibril conformation that differs completely from the known experimental models discussed in Section 3.2.2 [115, 143]. This work suggests an antiparallel β -sheet around residues 16-22 (with unknown length), with an inter- β -strand interface such that L17 bonds to A21 (designated as having “17+k \leftrightarrow 21-k” registry [186]). Similarly, a second antiparallel β -sheet likely exists around positions 30-36, with L34 and F19 in close contact. Interestingly, this specific A β_{1-40} registry has only previously been seen in the peptide fragment A β_{16-22} , which lacks D23 [185], while the antiparallel forming fragment A β_{11-25} exhibits inverted “17+k \leftrightarrow 22-k” and “17+k \leftrightarrow 20-k” registries [142] (Table 5.1).

To analyze this point mutant, the Boltzmann partition function for both A β_{1-40} and A $\beta_{1-40/D23N}$ was computed using schema \mathcal{A} (which allows antiparallel inter-peptide interactions), restricting β -strand length between 9 and 10 residues long, and conformations were sampled and clustered from the ensembles (Figure 5-1). The composition of these ensembles in terms of β -strand registration (e.g., 17+k \leftrightarrow 21-k) were then calculated and reported in Table 5.1.

From the table, we conclude that AmyloidMutants’ A $\beta_{1-40/D23N}$ predictions strongly prefer a “17+k \leftrightarrow 21-k” registry conformation, with predicted contacts between L34/F19, and very little variation within the ensemble. This arrangement agrees with observed A β_{16-22} structures. Conversely,

	“17+k↔22-k”	“17+k↔21-k”	Other
A β_{1-40} registry	QKLVFFAE X V V X EAF F V L KQ	QKLVFFAE X X EAF F V L KQ	
Predicted A β_{1-40}	69%	6%	25%
Predicted A $\beta_{1-40}/D23N$	11%	52%	37%
Observed A β_{11-25} [142]	✓	-	-
Observed A β_{16-22} [185]	-	✓	-

Table 5.1: **AmyloidMutants A β_{1-40} /A $\beta_{1-40}/D23N$ predictions suggest conformational switch:** Predictions using schema \mathcal{A} agree with published experimental evidence [186] showing an antiparallel, “17+k↔21-k” registry β -sheet in A $\beta_{1-40}/D23N$ amyloid fibrils. An ensemble was predicted for each full-length A β sequence and sampled structures were classified into one of three categories dependent on β -sheet registry (the two major conformation’s residue/residue bonding interactions shown, with **X** indicating position 23). The percent makeup of conformations within each predicted ensemble is shown (boldface), indicating a strong bias for each sequence to adopt different specific conformations within schema \mathcal{A} . Check marks also indicate A β registrations that have been experimentally observed in peptide fragments.

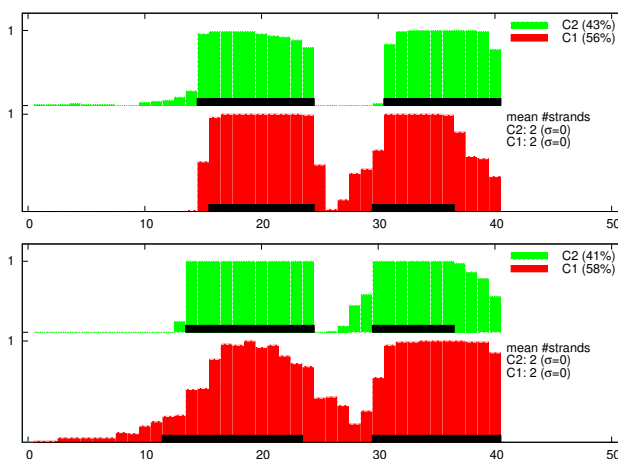


Figure 5-1: **AmyloidMutants ensemble predictions of A β_{1-40} and A $\beta_{1-40}/D23N$:** Schema \mathcal{A} results clustered into three populations allowing β -strand lengths to range between 9–10. **(a)** A β_{1-40} , **(b)** A $\beta_{1-40}/D23N$.

predictions for WT A β_{1-40} are quite heterogeneous, although with the largest cluster of structures forming “17+k↔22-k” registry, in agreement with observed A β_{11-25} structure. More strikingly, the “17+k↔21-k” registry conformation favored by A $\beta_{1-40}/D23N$ appears to be strongly disfavored by A β_{1-40} (and oppositely A $\beta_{1-40}/D23N$ appears to disfavor “17+k↔22-k” registry). These predictions and the divergence in ensemble makeup between A β_{1-40} and A $\beta_{1-40}/D23N$ supports the notion that the D23N mutation results in a singular energetically favorable conformational rearrangement from parallel β -sheets (in WT) to antiparallel β -sheets (in the D23N mutation). This is based on an as-

sumption that significantly low-energy conformations (manifested as a large ensemble population) may offer good predictive fits. Inspection at the residue-level suggests that the adoption of this “17+k↔21-k” conformation may be driven by both the alignment of oppositely charged K16 and E22, and the stacking arrangement of Q15 and N23.

5.1.2 HET-s yeast-toxic mutants

We further apply our approach to predict putative conformational rearrangements between a set of HET-s mutants shown to exhibit toxicity in yeast. In recent studies [15, 48], structural differences were found in a toxic HET-s mutant (named m8) and compared against four other non-toxic mutants (m3, m4, m9, m11), and WT. Notably, m8 exhibits a marked change from WT in secondary structure makeup, showing a shift of approximately half of the β -strand structure from parallel to antiparallel interactions.

We attempt to distinguish these phenotypically different mutants by inspecting predicted results using different schemas and comparing the relative structural heterogeneity of the ensembles. Again we premise that sequence mutants which significantly alter the predicted ensemble makeup (away from WT) are more likely to exhibit a different high-level conformational arrangement, and that high-likelihood conformations within an ensemble offer good predictive fits. Conversely, predictions that do not particularly favor any single conformation may suggest a poor fit. Table 5.2 reports

schema/class.	WT	m4	m8	m3	m9	m11
\mathcal{P} 2-rung	75%	95%	72%	13%	49%	55%
\mathcal{P} 1-rung	25%	5%	28%	87%	51%	45%
\mathcal{A} 2-rung-A	45%	42%	81%	44%	56%	50%
\mathcal{A} 2-rung-B	25%	43%	0%	36%	22%	40%
\mathcal{A} 1-rung	30%	15%	19%	20%	22%	10%
aggregation [48]	ring	foci	foci	diff.	diff.	diff.
toxicity [48]	—	minor	severe	—	—	—

Table 5.2: **AmyloidMutants predictions distinguish HET-s m8 mutant as unique:** Sequence m8 predictions (the only toxic mutant [48]) suggest the possibility of a conformational rearrangement in fibril structure: from a structure compatible with schema \mathcal{P} (WT sequence) to a structure compatible with schema \mathcal{A} (m8 sequence). The predicted ensemble of structures cluster into two general classifications when using schema \mathcal{P} and three when using \mathcal{A} (rows) — the relative percent makeup given. Strong percent bias for a structure within each ensemble may predict particular energetic favorability, and suggests that the sequence may favor such a conformational arrangement consistent with that schema (such as in the case of m8, boldface). Observed phenotypic differences between mutants are also described [48].

predicted ensemble makeup of the given six mutants, comparing schemas \mathcal{P} and \mathcal{A} . Across all mutants, schema \mathcal{P} predict clusters of 2-rung and 1-rung structures, while schema \mathcal{A} predicts three clusters: two forms of 2-rung solenoids, and one with 1-rung.

At a high-level, the difference between schemas \mathcal{P} 2-rung and \mathcal{A} 2-rung correlates with the shift in secondary structure makeup observed — \mathcal{P} 2-rung contains only parallel β -sheet structures while \mathcal{A} 2-rung can contain an equal amount of parallel and antiparallel β -sheet structure. Under schema \mathcal{P} , we see that WT, m4, and m8 form better 2-rung solenoids than a 1-rung solenoid, whereas with m3, m9, and m11, the opposite is true or no preference is apparent. This discrimination of mutants based on the structural landscape mirrors phenotypic variation seen by GFP-tagged aggregates [48] (independent of predictive accuracy). Under schema \mathcal{A} , we see similarities between the structural distribution of WT, m4, m3, m9, and m11; however, the toxic m8 mutant appears to strongly prefer only one of the 2-rung conformations. Such a dramatic shift in the predicted ensemble could suggest that the m8 mutant is energetically inclined to form the structure in cluster \mathcal{A} 2-rung-A.

More specifically, Figure 5-2 presents these clustered population predictions in much greater detail. Schema \mathcal{P} appears to permit two major structural populations across all mutants, a “two-rung-per-chain” β -solenoid (with β -structure approximately at positions 7-17, 19-29, 43-53, and 55-65) or a “one-rung-per-chain” β -solenoid (with approximate β -structure positions 21-30 and 41-51). Schema \mathcal{A} permits three major structural populations across mutants, two types of “two-rung-per-chain” β -solenoids, distinguished by the location of their β -strands (approximate β -structure positions of the first type: 2-12, 20-30, 38-38, and 54-64, and approximate positions of the second type: 14-24, 29-39, 43-53, and 58-68), and a “one-rung-per-chain” β -solenoid structure (with approximate β -structure positions 22-32, 51-60). Note, mutant m8 only exhibits one of the two types of “two-rung-per-chain” β -solenoid, seen by the overlap of β -strand positions in the clustered populations. When using schema \mathcal{P} , some mutants show dramatic change in the ensemble population, such as m4’s increase in propensity to form a “two-rung-per-chain” β -solenoid over a “one-rung-per-chain” β -solenoid, or m8’s change from a more homogeneous WT structural population to a more heterogeneous population with many different β -strand/ β -strand registrations. Oppositely, schema \mathcal{A} indicates a more heterogeneous population given the WT sequence, and a more ordered population for m8.

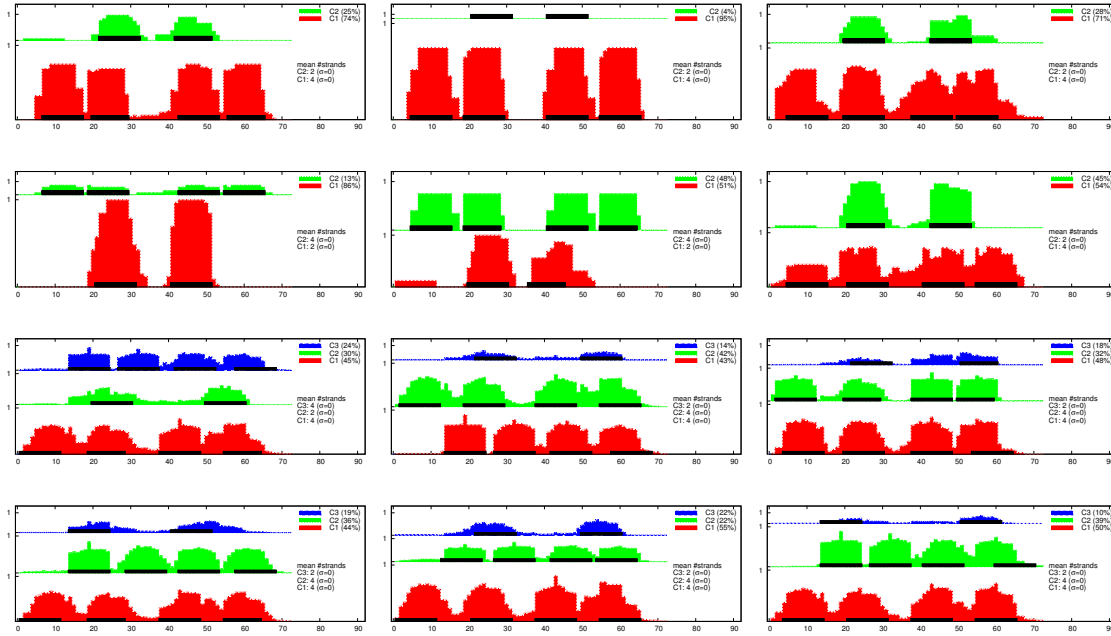


Figure 5-2: **AmyloidMutants** predictions of HET-s and yeast-toxic HET-s mutants: (a)-(f) Schema \mathcal{P} predictions clustered into two populations allowing β -strand lengths to range between 10–11. (g)-(l) Schema \mathcal{A} predictions clustered into two populations allowing β -strand lengths to range between 10–11. **WT**: (a) and (g). **m4**: (b) and (h). **m8**: (c) and (i). **m3**: (d) and (j). **m9**: (e) and (k). **m11**: (f) and (l).

5.2 Validation of predicted amyloidogenicity

Although our approach can provide detailed super-secondary structure information, we here evaluate its ability to accurately predict sequence-level per-residue amyloidogenic tendencies. This remains an important problem, and can be used to investigate whether a mutant is likely to form amyloid at all, rather than its particular conformation. To test our predictive accuracy, we perform an analysis of the 289-residue HET-s/HET-S natural homologs found in *Podospora anserina*, a combination of three A β scanning mutagenesis studies, and a set of 74 synthetic mutants of A β created by random mutagenesis. The amyloidogenicity of each mutation is predicted by computing a joint mutational landscape over WT and mutant sequences and quantifying sequence/structure state occupancy. This metric is described first.

5.2.1 Mutational occupancy:

Unfortunately, our ensemble approach describes both structural variation and sequence variation under a fixed assumption of an amyloid fibril fold, making the prediction of sequence amyloidogenic-

ity indirect. We therefore take the following approach. To determine whether a particular mutation makes a peptide more or less amyloidogenic, we quantify each sequence's energetic contribution to the ensemble as a whole and compare these quantities. In other words, if only two sequences are permitted during a prediction, the ensemble will contain 50% of the states with one sequence and 50% of the states with the other. However, the energetic weight of the structures resulting from these sequences will vary. In this case, we assert that the sequence with the larger energetic weight is more amyloidogenic since it forms better energy structures. For instance, when comparing predictions of a WT and mutant sequence, if the mutant sequence occupies 90% of the energetic weight of the ensemble, then the mutant is suggested as a better amyloid forming sequence. Note, however, such comparisons are only valid within a single prediction using one schema, where all possible states are accounted.

5.2.2 HET-s/HET-S:

The *Podospora anserina* HET-s allele forms an amyloid conformation in its prion form, while HET-S does not, despite differing by only three residues in the amyloid-forming 72-residue C-terminus, and 13 overall [47]. Predicting the joint HET-s/HET-S mutational landscape, AmyloidMutants found that approximately 72% of the ensemble favored HET-s, indicating that it's more amyloidogenic than HET-S. Although N-terminal mutations can induce a prion state in HET-S [47], our predictions suggests a sequence bias in HET-s permitting a more energetically favorable path for amyloid formation.

5.2.3 A β single-point proline mutagenesis:

Scanning mutagenesis studies have been performed on A β ₄₀ to detect the sequence position effect of proline-, alanine-, and cysteine-replacement on amyloid fibril formation, measured by WT/mutant $\Delta\Delta G$ [158, 208, 209]. Although P, A, and C-replacement $\Delta\Delta G$ values are difficult to interpret independently (due to experimental structural heterogeneity [209]), they support the broader conclusion that A β ₄₀ positions 18–21, 25–26, and 32–33 are particularly sensitive to P-replacement [209]. AmyloidMutants' predictions of the joint mutational landscape for individual proline replacements identified positions 16–25 and 31–35 as particularly disruptive, in agreement with these studies. Figure 5-3 plots this agreement along with similar predictions by other amyloid prediction tools TANGO and Zyggregator. However, we stress that only trends should be inferred from this plot as a direct one-to-one comparison between predictions and $\Delta\Delta G$ values would be inappropriate.

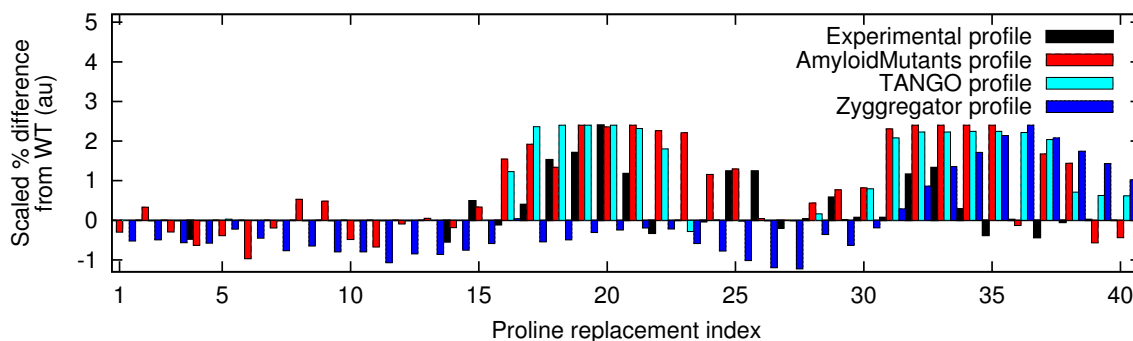


Figure 5-3: **A β ₄₀ scanning mutagenesis predictions compared with experimental data:** Comparison of AmyloidMutants, TANGO, and Zyggregator to experimental A β ₄₀ scanning mutagenesis data. Experimental “Pro-Ala $\Delta\Delta G$ ” values indicating the sensitivity of each sequence position to proline replacement [209]. All three predictors agree with this data around positions 32–33, AmyloidMutants and TANGO also agree with experimental data around positions 18–21, and AmyloidMutants agrees with experimental data around position 25. Since $\Delta\Delta G$ values and each predictor’s scores differ greatly in range, results are given in arbitrary units, scaled such that the maximum percent change in aggregation score of any predictor is 2.4 (the maximum experimental $\Delta\Delta G$ value). For all bars, positive values suggest that a proline replacement results in a less stable amyloid fibril, while negative values imply the opposite. Before scaling, AmyloidMutants values represented the percent difference in ensemble occupancy between WT and mutant sequences, TANGO values represented the inverted percent change in AGG score, and Zyggregator values represented the inverted percent change in Zagg score.

5.2.4 A β multiple-residue mutagenesis:

Finally, we evaluated AmyloidMutants amyloidogenicity prediction on a set of 74 A β mutants [100, 101, 212] whose relative aggregation levels were observed by GFP fluorescence relative to WT. Mutational occupancy scores identify which mutants forms amyloid more (or less) readily than WT in 81% of sequences (60 of 74). This is illustrated graphically in Figure 5-4, where relative change in GFP fluorescence is plotted as a function of predicted change in mutational occupancy.

5.3 Identification of amyloidogenicity bias of *Asn* over *Gln* in HET-s

Here we describe our use of AmyloidMutants to study a more fundamental question: the role chemically similar residues *Asn* and *Gln* play in fibril structure. Given the high propensity of Q/N-rich peptides to form amyloid [36], the amyloidogenic potential of *Asn* and *Gln* has often been considered equal — however, recent evidence suggests that N-rich proteins may have a slightly higher tendency to form amyloid [2]. We study this question by considering the effect of four ladder-forming

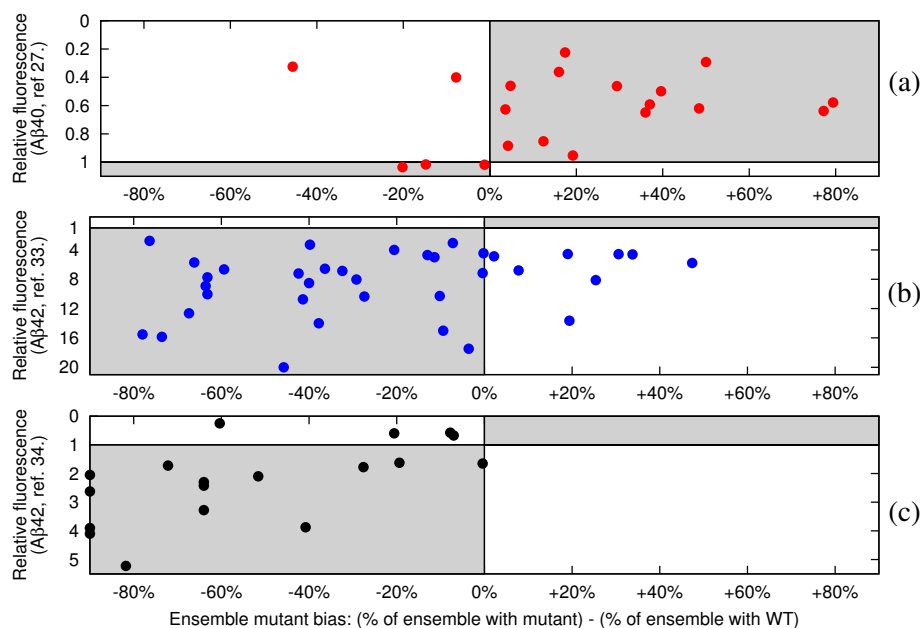


Figure 5-4: **AmyloidMutants amyloidogenicity predictions of multiple-residue A β mutants:** Mutants are either predicted to be more or less-amyloidogenic than wild-type — *positive x-values* indicate more-amyloidogenic while *negative x-values* indicate less-amyloidogenic than wild-type. Three studies are compared, **(a)** [101], **(b)** [212], **(c)** [100], that measure mutant vs. wild-type amyloidogenicity by GFP fluorescence — less fluorescence (lower numbers) indicates more amyloid formation. Shaded gray regions indicate where predictions agree with experimental observations.

asparagine residues in *Podospora anserina* HET-s (positions 226, 243, 262, and 279) which are believed important for fibril stabilization [205], and whose regions are conserved in a *Fusarium graminearum* homolog. AmyloidMutants sequence/structure landscapes were calculated permitting these four residues to mutate to *Gln* (“HET-s/4N \rightarrow Q”), and the likelihood and corresponding energetic weight of each sequence within the ensemble was compared. The WT HET-s sequence was much more energetically favorable than HET-s/4N \rightarrow Q, comprising approximately 96% of the ensemble, suggesting a greatly reduced ability of HET-s/4N \rightarrow Q to form fibrils, and a putatively higher amyloidogenic potential of *Asn* over *Gln*. Stochastic contact map predictions further illustrate this difference between sequences (Figure 5-6).

In collaboration, we tested these predictions experimentally, using purified recombinant WT and 4N \rightarrow Q HET-s proteins. Denatured proteins were diluted into a physiological buffer and allowed to form amyloid. While the WT protein readily did so, as detected by the retention of detergent-insoluble aggregates on a non-binding membrane, the mutant protein was recalcitrant to amyloid formation, shown in Figure 5-5.

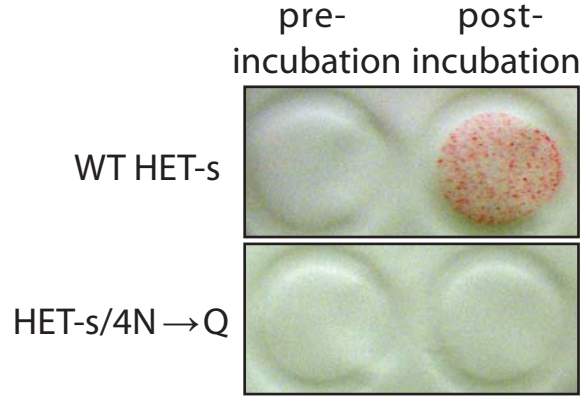


Figure 5-5: **HET-s/4N→Q is defective for amyloid assembly:** Purified proteins were filtered through a non-binding membrane either before or after incubation for 24 hrs in a physiological buffer. Protein aggregates that formed during the incubation are retained on the surface of the membrane, as visualized by Ponceau-S staining.

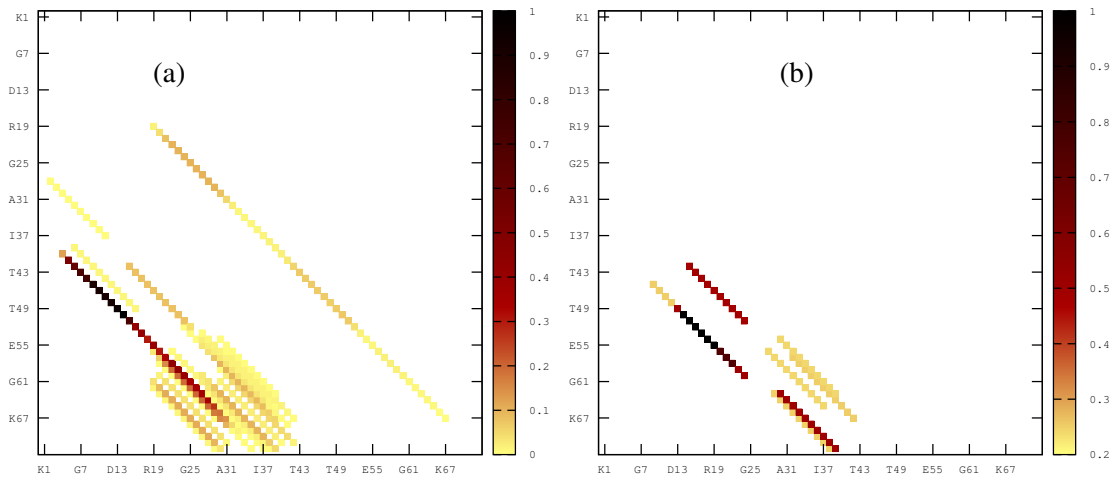


Figure 5-6: **Predicted contact maps highlight differences in WT HET-s and HET-s/4N→Q:** Predictions were made using schema \mathcal{P} , allowing either WT or HET-s/4N→Q mutations within the ensemble. The stochastic contact map of the WT HET-s cluster is given by (a) while the HET-s/4N→Q is given in (b). Sparse contacts in HET-s/4N→Q do not indicate a strong structure prediction, but most likely result from either sampling bias (as HET-s/4N→Q makes up only 4% of the ensemble), or a basic inability of HET-s/4N→Q to form fibril.

The specific protocol for these experiments involved the following. Sequences encoding HET-s WT and HET-s/4N→Q proteins were subcloned into pRH1 [2] to allow their expression in bacteria as 7xHis fusions. The proteins were expressed in *E. coli* strain BL21-AI and purified under denaturing conditions, as described [2]. Methanol-precipitated proteins were resuspended in 6 M GdnHCl, incubated for 5 min at 95°C, and then filtered through a YM-100 Microcon filter immediately prior

to use. Proteins were diluted to 20 μM (corresponding to approximately 60 mM GdnHCl in the assembly reactions) in assembly buffer (5 mM K₂HPO₄, pH 6.6; 150 mM NaCl; 5 mM EDTA; 2 mM TCEP) and allowed to incubate in 1.5 ml non-binding tubes, with 1000 rpm horizontal agitation, at 23°C for 24 hrs. Because HET-s amyloids are difficult to detect using the amyloid-specific dye thioflavin T, we instead used detergent insolubility as a measure for amyloid formation. Protein aggregates were detected by passaging the post-incubated reactions through a cellulose acetate membrane, followed by washing with 2% Sarkosyl, essentially as described [2]. Retained proteins were visualized by Ponceau-S staining.

5.4 Investigation of *E. coli* curli and biofilm inhibition

In collaboration, we have used our ensemble methodology to advance our understanding of *E. coli* curli proteins. Curli proteins are functional amyloid fibrils important for the physiology of *E. coli* and other enteric bacteria [10]. In particular, curli is localized to bacterial cell surfaces and mediates cell-cell and cell-surface contacts associated with biofilm formation — a problem of great magnitude in diverse medical and industrial settings. Curli are also involved in adhesion and invasion of mammalian cells, and are formed through a controlled process regulated by many factors. The major curli subunit, CsgA, is secreted as a soluble protein to cell surfaces where it is polymerized into amyloid fibrils by CsgB, an outer-membrane associated protein. However, beyond the identification of amyloidogenic domains in CsgA and CsgB [81, 204] relatively little is known about the specific molecular structure of this complex. In particular, the nucleation sequences in CsgB and corresponding interacting sites in CsgA are undetermined.

To create an effective therapeutic for biofilm inhibition we performed sequence and structure studies of CsgA, CsgB, and the CsgA/CsgB interface using AmyloidMutants, with the goal of using these results to design a targeted mechanism for disrupting polymerization. To study putative CsgA/CsgA structures we utilized schemas \mathcal{P} , \mathcal{A} , and \mathcal{S} (see Section 2.3). However, modeling potential CsgA/CsgB interfaces required the design of a new set of schemas capable of handling heterogeneous fibril structures involving multiple sequences, named \mathcal{P}^2 , \mathcal{A}^2 , and \mathcal{S}^2 . Figure 5-7 depicts schema \mathcal{P}^2 .

AmyloidMutants CsgA/CsgB and CsgA/CsgA structure predictions using \mathcal{P}^2 and \mathcal{P} identified 4 high-likelihood inter-peptide β -strand/ β -strand interaction sites among many β -strand regions that were common to high-scoring 3, 4, and 5 rung-per-CsgA-peptide β -solenoid conformations (other

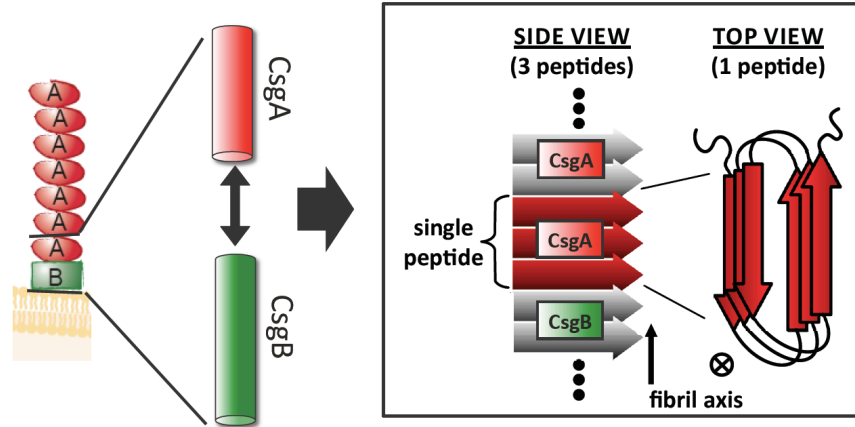


Figure 5-7: **Ensemble schema design for heterogeneous CsgA/CsgB amyloid fibril:** Illustration of schema \mathcal{P}^2 used to model CsgA/CsgB interactions, based on schema \mathcal{P} .

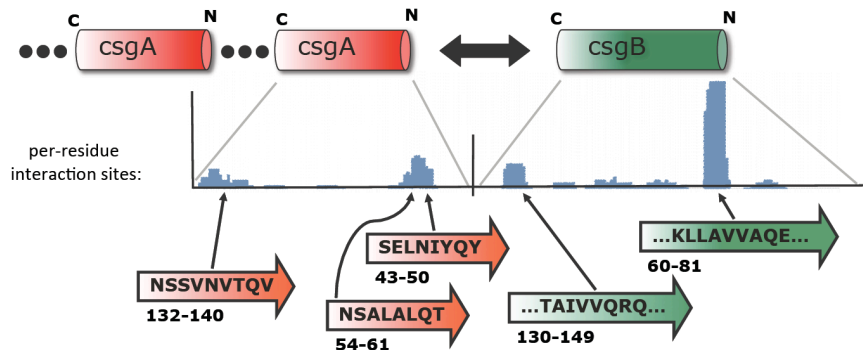


Figure 5-8: **AmyloidMutants predicted sequence regions for CsgA/CsgB interaction:** Illustrated are the 5 highest scoring regions for putative CsgA/CsgB interactions: residues 43-50, 54-61, and 132-140 in CsgA, and residues 60-81 and 130-149 in CsgB. Different N/C-terminal orientations were also calculated, with CsgA N-terminal to CsgB C-terminal depicted. Subsequent experimental studies support the importance of these regions to biofilm formation [105].

conformational shapes scored worse). Further, we artificially scaled our energetic scoring function to predict only the strongest β -strand sites by introducing a parameter that reduced the contribution of β -strand contacts with respect to coils by 8-fold. This was done to limit predictions to only those highest-scoring β -strand regions, and to examine any inter-peptide/intra-peptide energetic bias in our model. The resulting predictions identified the same 4 β -strand regions. Note, signal sequences were removed before prediction.

Figure 5-8 depicts the β -sheet propensity of this latter experiment and the 4 high-likelihood β -strand regions. Within CsgB, sequence positions 60–81 and 130–149 were predicted to form inter-peptide β -strands. An independent peptide array analysis of CsgB was conducted which re-

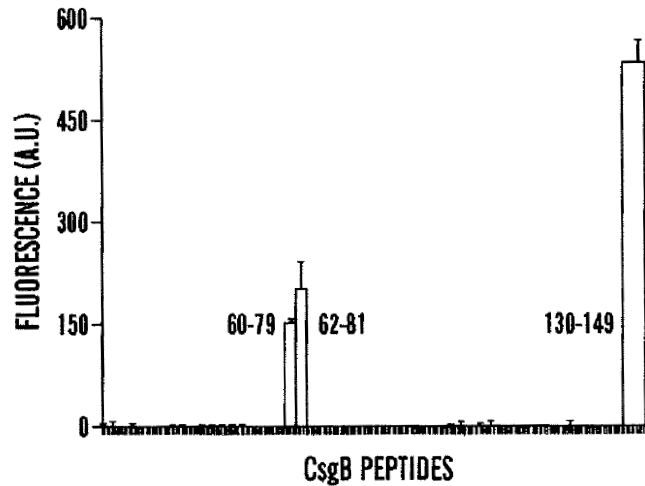


Figure 5-9: **CsgB peptide array seeds CsgA amyloid fibrils at two sequence positions:** Sequence regions 60-81 (LRQGGSKLLAVVAQEGSSNRAK) and 130-149 (GTQKTAIVVQRQSQ-MAIRVT) are able to nucleation fibril formation when CsgA is washed over the array. This figure and more detailed information can be found in Lu et al. [105, 113].

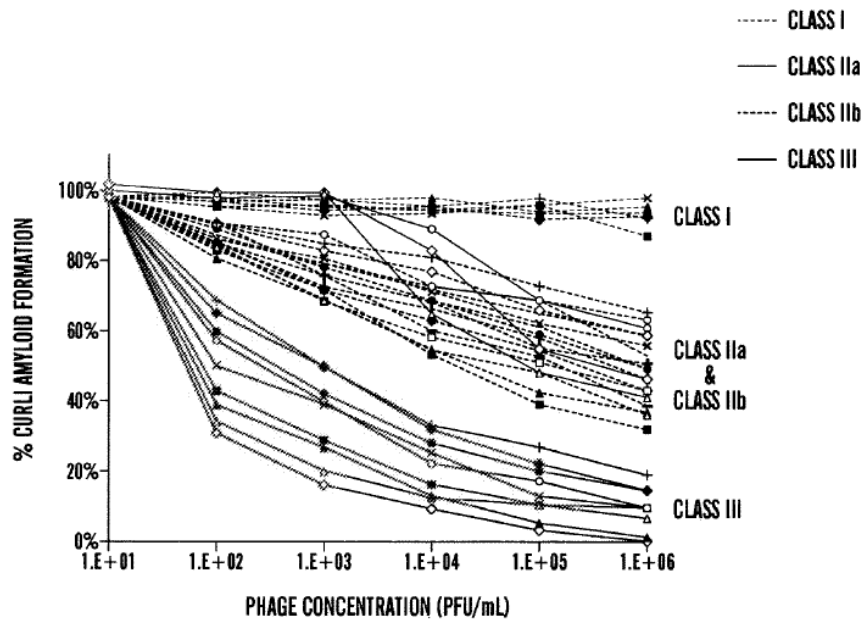


Figure 5-10: **Reduction in curli amyloid formation using phage-display peptide library:** Our library of bacteriophage-expressed peptides offered varying degrees amyloid fibril inhibition, categorized into three major classes. This figure and more detailed information can be found in Lu et al. [105, 113].

vealed the same two nucleation regions as predicted [105, 113] (Figure 5-9). Inter-peptide CsgA β -strands were predicted at regions 43–61 (with two distinct likelihoods at 43–50 and 54–61), and

132–140. For example, potential CsgA/CsgB interactions could include inter-peptide contacts at CsgA_{43–50}/CsgB_{134–141} (...NSELN_{43–50}YQ.../...TAIVV_{134–141}QRQ...), and so forth. The location of these predicted CsgA β -strands falls within repeats R1 and R5, repeats shown necessary for fiber formation via deletion experiments [203, 204]. Site-specific point lysine mutations were made within the first locus (positions 43–50 and 54–61), resulting in *E. coli* biofilm inhibition, along with additional mutations made adjacent to these regions (including the region 51–53 in between) that resulted in no phenotypic change [105].

Based on these predictions, a library of peptides were designed to target and inhibit these nucleation site candidates, in agreement with our algorithmic and mutational analyses. With this, our collaborators were able to design an efficient therapeutic delivery system based on phage-display. These bacteriophage candidates were shown to reduce *in vitro* curli assembly, decrease *E. coli* biofilm formation, block *E. coli* invasion of mammalian cells, and retard *E. coli* colony growth [105]. A reduction in curli amyloid formation using our library of bacteriophage can be seen in Figure 5-10. More information can be found in Lu et al. [113]. The subject of this investigation remains ongoing work.

Chapter 6

Simultaneous alignment and folding: consensus prediction

6.1 Goals and Overview

Chapters 2 and 4 described the use of our ensemble approach to create informative models of protein sequence and structure landscapes. Here we describe the use of ensembles for the purposes of comparative modeling, with the goal of identifying structural and sequential alignments with two (or more) proteins. This can be seen as a specialized application of mutational landscapes for the case of two distinct protein sequences, with the addition of a sequence comparison step. This specialization allows for powerful new analyses.

Conceptually, we introduce an algorithmic framework for simultaneously computing the structural landscape of two proteins based on an objective function that combines energetic structural interactions scores with sequence alignment scores. The resulting landscape is populated with *consensus structures*: low-energy conformations that both sequences can adopt given a sample-specific sequence alignment mapping. Consensus folds are an important consideration in structural bioinformatic analyses. In structure-function relationship studies, proteins that have the same consensus fold are likely to have the same function and be evolutionarily related [167]; in protein structure prediction studies, consensus fold predictions can guide tertiary structure predictors; and in sequence alignment algorithms [62], consensus fold predictions can improve alignments. The primary limitations in achieving accurate consensus folding, however, is the difficulty of obtaining reliable sequence alignments for divergent protein families and the inaccuracy of folding algorithms.

The specific problem we address is predicting consensus folds of proteins from their unaligned sequences. This definition of consensus fold should not be confused with the agreed structure between unrelated predictors [165]. Our approach succeeds by *simultaneously* aligning and folding protein sequences. By concurrently optimizing unaligned protein sequences for both sequence homology and structural conservation, both higher fidelity sequence alignment and higher fidelity structure prediction can be obtained. For sequence alignment, this sidesteps the requirement of correct initial profiles (because the best sequence aligners require profile/profile alignment [72]). For structure prediction, this harnesses powerful evolutionary corollaries between structure.

While this class of problems has received much attention in the RNA world [9, 55, 82, 86, 119, 159], it has not yet been applied to proteins. Applying these techniques to proteins is more difficult and less clearly defined. For proteins, the variety of structures is much more complicated and diverse than the standard RNA structure model, necessitating our approach of beginning with an abstract representation of structure, be it a TMB or amyloid fibril schema. Moreover, for proteins, there is no clear chemical basis for compensatory mutations [68], the energy models that define β -strand pairings are more complex, and the larger residue alphabet vastly increases the computational complexity of the problem.

This class of problems is also different than any that have been attempted for structure analysis. The closest related structure-prediction methods rely on sequence profiles, as opposed to consensus folds. Current protein threading methods such as RAPTOR [214] often construct sequence profiles of the “query” sequence before threading it onto solved structures in the PDB; however, given two “query” sequences, even if they are functionally related, it will output two structure matches but does not try to form a consensus from these. There are β -structure specific methods that “thread” a profile onto an abstract template representing a class of structures [20], but do not generate consensus folds.

In this section we focus on the simultaneous alignment and folding of pairs of unaligned protein sequences. This is an important first step as pairwise alignment is an important component in achieving reliable multiple alignments. Using this tool, we obtain significantly better pairwise sequence alignments than other alignment techniques for the case of proteins with low sequence identity. Our approach also obtains improved structural prediction accuracy, particularly in cases where single-sequence results are poor. We also describe the ability to predict consensus folds in amyloid fibrils. These results are detailed in Chapter 7. Given the broad generality of this approach and its proven impact on the RNA world, we hope that this idea will be used much more prevalently in protein structure prediction.

6.2 Transmembrane β -barrel consensus modeling

In this section we describe an algorithm for simultaneous alignment and folding of transmembrane β -barrel proteins, which has been implemented as a publically-accessible web-based tool named `partiFoldAlign`¹. To design an algorithm for simultaneous alignment and folding we must overcome one fundamental problem: predicting a consensus fold (structure) of two unaligned protein sequences requires a correct sequence alignment on hand, however, the quality of any sequence alignment depends upon the underlying unknown structure of the proteins. We adopt our solution to this issue from the approach introduced by Sankoff [159] to solve this problem in the context of RNAs — by predicting *partial* structural information that is then aligned through a dynamic programming procedure. In our approach this partial information is effectively a predicted stochastic contact map.

Similar to the method outlined in Section 2.2 we use a bottom-up recursion to describe the space of TMB structures and sequence alignments and use dynamic programming schemes to efficiently sample this space. Optimal solutions in this space are identified by a convex combination of ensemble-derived contact probabilities and sequence alignment matrices [83, 153, 174]. Broadly speaking, our simultaneous alignment and folding procedure begins by predicting the ensemble-based probabilistic contact map of two unaligned sequences. Alignment is then broken into two structurally different parts: the alignment of β -sheets, and the alignment of coils (depicted in Figure 6.2). Coil alignments can be performed independently at each position, however β -sheet alignments must respect residue pair assignments. Finally, to decompose the problem (Figure 6-2), we first consider the optimal alignment of a single β -sheet with a given inclination, including the enclosed coil alignment. Once all single alignments have been found, we “chain” these subproblems to arrive at a single consensus alignment and structure.

To overcome the intractability of this problem, we exploit sparsity in the set of likely amino acid pairings and aligned residues, inspired by the `LocARNA` algorithm [207]. Therefore β -strand contacts below a parameterizable threshold are excluded to allow for an efficient alignment of the most likely interactions. With these optimization, `partiFoldAlign` is able to achieve effectively cubic time and space in the length of its input sequences. We note that this technique is also somewhat related to the problem of *maximum contact map overlap* [25], although in such problems, contact maps implicitly signify the biochemical strength of a contact in a *solved* structured and not a well-

¹Available at <http://partifold.csail.mit.edu/>

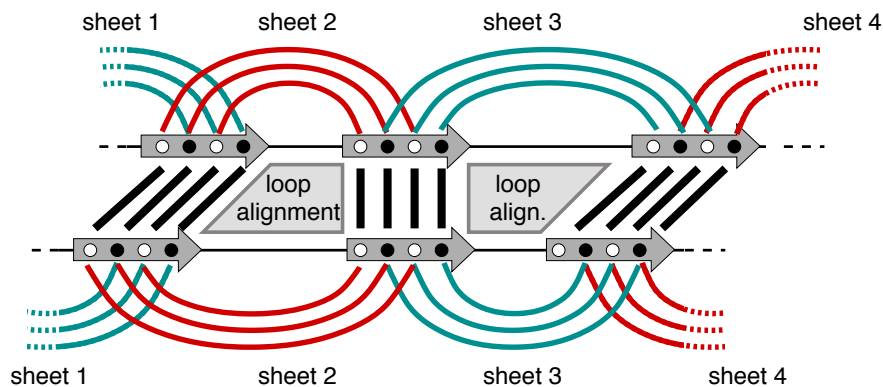


Figure 6-1: **Elements of a TMB sequence/structure alignment:** Differently colored amino acids in the sheet denote exposure to the membrane and to the channel, respectively. In a valid sheet alignment, only amino acids of the same type can be matched, whereas no further constraint (except length restriction) are applied to the loop alignment.

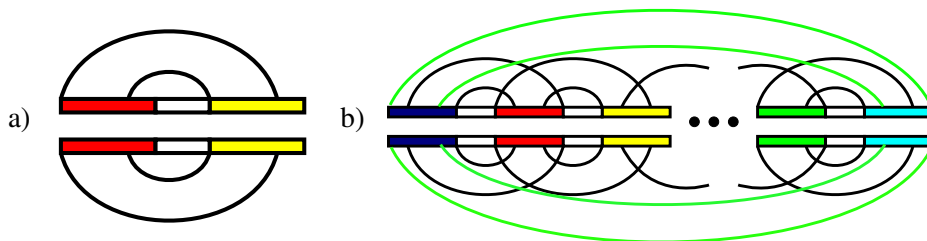


Figure 6-2: **Decomposition strategy for TMB alignment:** (a) alignment of a single sheet including the enclosed loop with positive shear; (b) chaining of single sheet alignment to form a β -barrel. Green arcs indicate the closing sheet connecting beginning and end.

distributed likelihood of interaction taken from a complete ensemble of possible structures.

6.2.1 Representing consensus structure ensembles

Due to the specificity of the bottom-up approach used in Section 2.2, we must redefine our representation of TMB sequence/structure states to enable a simultaneous alignment and folding algorithm. Formally, we define an alignment \mathcal{A} of two sequences a, b as a set of pairs $\{(p_1, p_2) \mid p_1 \in [1..|a|] \cup \{-\} \wedge p_2 \in [1..|b|] \cup \{-\}\}$ such that (i) for all $(i, j), (i', j') \in (\mathcal{A} \cap [1..|a|] \times [1..|b|])$ we have $i < j \implies i' < j'$ (non-crossing) and (ii) there is no $i \in [1..|a|]$ (resp. $j \in [1..|b|]$) where there are two different p, p' with $(i, p), (i, p') \in \mathcal{A}$ (resp. $(p, j), (p', j) \in \mathcal{A}$). Furthermore, for any position in both sequences, we must have an entry in \mathcal{A} . We say that \mathcal{A} is a *partial alignment* if there are some sequence positions for which there is no entry in \mathcal{A} . In this case, we denote with $\text{def}(a, \mathcal{A})$ (resp. $\text{def}(b, \mathcal{A})$) the set of positions in a (resp. b) for which an entry in \mathcal{A} exists. This provides us a mapping for putative residue/residue β -strand contacts to alignments.

As shown in Figure 6-1, we allow two possible side chain orientations for any given β -strand amino acid: facing the channel (C) and facing the membrane (M). Since contacts can form only if both amino acids share the same orientation, a *TMB probabilistic contact map* P of any TMB a is a matrix $P = (P(i, i', x))_{1 \leq i < i' \leq |a|, x \in \{C, M\}}$ where $P(i, i', x) = P(i', i, x)$ and $\forall x \in \{C, M\} : \sum_i P(i, i', x) \leq 1$. To overcome the intractability of this problem, we use only those entries in the matrix P which have a likelihood above a parameterizable threshold.

We weight the alignments with a scoring function that sums a folding energy term $\mathcal{E}()$ with an alignment score $\mathcal{W}()$, where the energy term $\mathcal{E}()$ corresponds to the sum of the folding energies of the consensus structure mapped onto the two sequences (see Section 2.4). To allow a convex optimization of this function, we introduce a parameter α distributing the weights of the two terms. Thus, given two sequences a, b , an alignment \mathcal{A} and a consensus TMB structure \mathcal{S} of length $|\mathcal{A}|$, the score of the alignment is:

$$\text{score}(\mathcal{A}, \mathcal{S}, a, b) = (1 - \alpha) \cdot \mathcal{E}(\mathcal{A}, \mathcal{S}, a, b) + \alpha \cdot \mathcal{W}(\mathcal{A}, a, b)$$

The effects of selecting different values of α are explored in Section 7.1.2.

Let $E_{ct}(x, y)$ be the energy value of a pairwise residue contact. Since all residue/residue contacts in a consensus structure are aligned (by definition), we define the energy component of the score() as:

$$\mathcal{E}(\mathcal{A}, \mathcal{S}, a, b) = \sum_{\substack{\binom{i}{j} \in \mathcal{A}, \binom{i'}{j'} \in \mathcal{A} \\ (i, i') \in \mathcal{S}_a^{\text{arcs}}, (j, j') \in \mathcal{S}_b^{\text{arcs}}}} \tau(i, i', j, j'), \text{ where } \tau(i, i', j, j') = E_{ct}(i, i') + E_{ct}(j, j')$$

Although in practice our tool implements both a pairwise and stacking pair energy model (see Section 2.4), we assume for clarity here that only pairwise contacts are allowed.

Let $\sigma(x, y)$ be the substitution score of the amino acids x by y , and $g(x)$ an insertion/deletion cost. Then, the sequence alignment component of the score() is given by:

$$\mathcal{W}(\mathcal{A}, a, b) = \sum_{\binom{i}{j} \in \mathcal{A}} \sigma(a_i, a_j) + \sum_{\binom{i}{-} \in \mathcal{A}} g(a_i) + \sum_{\binom{-}{j} \in \mathcal{A}} g(a_j)$$

Again, in practice, a penalty for opening gaps is added but not described here for clarity. Finally,

the optimization problem our algorithm solves is, given two sequences a and b :

$$\arg \max_{\substack{\mathcal{A} \text{ TMB alignment of } a \text{ and } b, \\ \mathcal{S} \text{ TMB structure of length } |\mathcal{A}|}} \{\text{score}(\mathcal{A}, \mathcal{S}, a, b)\}.$$

To account for the sidechain orientation of residues in TM β -strands toward the channel or the membrane, the $\mathcal{E}()$ and $\mathcal{W}()$ recursion equations require a slightly more detailed version of the scoring. An additional condition is that contacts only happen between amino acids with the same orientation, and that this orientation alternates between consecutive contacts. Hence, we introduce in τ an additional parameter env standing for this sidechain orientation environment feature. The same holds for the alignment edit scores σ and g , where the orientation can also be the loop environment. For strands we use $\sigma_s(i, j, env)$, while for loops we distinguish inner from outer loops (indicated by the loop type lt) with the amino acids in the loops scored using $\sigma_l(i, j, lt)$. Gaps are treated analogously. We also note that, for simplicity and computational reasons, we do not account for a strand extension term as considered in Section 2.2.2.

6.2.2 Computing the partition function tables

Here we define a decomposition of the simultaneous alignment and folding problem into a description amenable to efficient dynamic programming. We describe the construction of basic aligned β -sheet states, and then show how to combine these states into a global solution. Finally, we provide a complexity analysis of the algorithm.

Alignment decomposition

The alignment of a single antiparallel strand pair as shown in Figure 6-2(a) has nested arcs and an outdegree of at most one. To account for this fact, we introduce a table $\text{SHA}()$ (where SHA stands for *sheet alignment*) aligning pairs of subsequences $a_{i..i'}$ and $b_{j..j'}$. Another parameter to account for is the shear number which represents the inclination of the strands in the TM β -barrel. Since the strand pair alignments also include a loop alignment, and the scoring function of this loop depends on the loop type (inner/outer loop), we need to set the loop type as an additional parameter. Similarly, we need to know the orientation of the final contact to ensure the succession of channel and membrane orientations. Given an orientation environment of a contact env , the term $\text{next}_c(env)$ return the orientation of the following contact. Thus, we have a table $\text{SHA}(i, i'; j, j'; env; lt; s)$ with

the following recursion:

$$\text{SHA}(i, i'; j, j'; env; lt; s) = \max \begin{cases} \text{SHA}_{gap}(i, i'; j, j'; env; lt; s) \\ \text{SHA}_{shear}(i, i'; j, j'; env; lt; s) & \text{if } s \neq 0 \\ \text{SHA}_{contact}(i, i'; j, j'; env; lt) & \text{if } s = 0 \\ \text{LA}(i, i'; j, j'; lt) & \text{if } s = 0 \end{cases}$$

where

$$\begin{aligned} \text{SHA}_{contact}(i, i'; j, j'; env; lt) = & \text{SHA}(i + 1, i' - 1; j + 1, j' - 1; \text{next}_c(env); lt; 0) \\ & + \tau(i, i'; j, j'; env) + \sigma_s(a_i, b_j, env) + \sigma_s(a_{i'}, b_{j'}, env) \end{aligned}$$

$$\begin{aligned} \text{SHA}_{gap}(i, i'; j, j'; env; lt; s) = & \max \begin{cases} \text{SHA}(i + 1, i'; j, j'; env; lt; s) + g_s(a_i, env) \\ \text{SHA}(i, i' - 1; j, j'; env; lt; s) + g_s(a_{i'}, env) \\ \text{SHA}(i, i'; j + 1, j'; env; lt; s) + g_s(b_j, env) \\ \text{SHA}(i, i'; j, j' - 1; env; lt; s) + g_s(b_{j'}, env) \end{cases} \\ \text{SHA}_{shear}(i, i'; j, j'; env; lt; s) = & \max \begin{cases} \text{SHA}(i + 1, i'; j + 1, j'; env; lt; s + 1) \\ \quad + \sigma_s(a_i, b_j, env) & \text{if } s < 0 \\ \text{SHA}(i, i' - 1; j, j' - 1; env; lt; s - 1) \\ \quad + \sigma_s(a_{i'}, b_{j'}, env) & \text{if } s > 0 \end{cases} \end{aligned}$$

SHA_{gap} , $\text{SHA}_{contact}$ and SHA_{shear} are introduced for better readability and will not be tabulated. The matrix $\text{LA}(i, i'; j, j'; lt)$ represents an alignment of two loops $a_{i..i'}$ and $b_{j..j'}$, with a loop type lt . This table can be calculated using the usual sequence alignment recursion. Thus, we have

$$\text{LA}(i, i'; j, j'; lt) = \begin{cases} \text{LA}(i, i' - 1; j, j'; lt) + g_l(a_{i'}, lt) \\ \text{LA}(i, i'; j, j' - 1; lt) + g_l(b_{j'}, lt) \\ \text{LA}(i, i' - 1; j, j' - 1; lt) + \sigma_1(a_{i'}, b_{j'}, lt) \end{cases}$$

As already mentioned, we use a probability threshold to reduce both the space and time complexity of the problem. Thus, we tabulate only values in the SHA-matrix for positions i, i' and j, j' where the contact probability is above a given threshold in both sequences. This is tabulated at the granularity of β -strand/ β -strand pairs in practice to further reduce computation time.

Alignment Chaining

Given the alignment of single β -sheets defined above, our next task is to combine these different alignments by what we call *chaining* (Figure 6-2(b)). To build a valid global alignment, we have to

guarantee that the sub-alignments agree on their overlapping regions. We define a *strand alignment* \mathcal{A}_s as a partial alignment, and extend the matrices for sheet alignments by an additional entry allowing the alignment of strand regions. Since our model assumes no β -strand bulges, one can insert or delete only a complete contact instead of a single amino acid. When chaining sheet alignments, the gap in one strand is then transferred to the chained sheet (by the agreement of sub-alignments).

Formally, we extend the matrices of sheet alignments by an alignment descriptor which is used to ensure the compatibility of sub-solutions used in the recursion. Note that although the alignment is fixed for the strands of a sheet, the scoring is not. Therefore, the new matrix is then formulated as $\text{SHA}(i, i'; j, j'; env; lt; s; \mathcal{A}_s)$, where we enforce \mathcal{A}_s to satisfy $\text{def}(a, \mathcal{A}_s) = [i..l_1] \cup [r_1..i']$ and $\text{def}(b, \mathcal{A}_s) = [j..l_2] \cup [r_2..j']$ for some $i < l_1 < r_1 < i'$ and $j < l_2 < r_1 < j'$. The new version of $\text{SHA}()$ is then

$$\text{SHA}(i, i'; j, j'; env; lt; s; \mathcal{A}_s) = \max \begin{cases} \text{SHAgap}(i, i'; j, j'; env; lt; s; \mathcal{A}_s) \\ \text{SHAshear}(i, i'; j, j'; env; lt; s; \mathcal{A}_s) & \text{if } s \neq 0 \\ \text{SHAcontact}(i, i'; j, j'; env; lt; \mathcal{A}_s) & \text{if } s = 0 \\ \text{LA}(i, i'; j, j'; lt) & \text{if } s = 0 \end{cases}$$

$\text{LA}(i, i'; j, j'; lt)$ does receive an additional parameter since sub-alignment agreement during chaining is restricted to strands. Therefore we modify the definitions of $\text{SHAgap}()$, $\text{SHAcontact}()$ and $\text{SHAshear}()$ to ensure that the associated alignment operations are compatible with \mathcal{A}_s . Thus, the new definition of $\text{SHAcontact}()$ is

$$\text{SHAcontact}(i, i'; j, j'; env; lt; \mathcal{A}_s) = \max \begin{cases} \text{SHA}(i+1, i'-1; j+1, j'-1; env; lt; 0; \mathcal{A}_s) & \text{if } (i, j) \in \mathcal{A}_s \\ + \tau(i, i'; j, j'; env) + \sigma_s(a_{i'}, b_{j'}, env) & \text{and } (i', j') \in \mathcal{A}_s \\ -\infty & \text{else} \end{cases}$$

If all entries are incompatible with \mathcal{A}_s , then $-\infty$ is returned. Note that we add an amino acid match score only for a single specified end of the contact. Thus, $\sigma_s(a_i, b_j)$ is skipped. The reason is simply that otherwise this score would be added twice in the course of chaining. The new definition

of *SHAshear* is then

$$SHAshear(i, i'; j, j'; env; lt; s, \mathcal{A}_s) = \max \begin{cases} \text{SHA}(i + 1, i'; j + 1, j'; env; lt; s + 1; \mathcal{A}_s) & \text{if } s < 0 \wedge (i, j) \in \mathcal{A}_s \\ \text{SHA}(i, i' - 1; j, j' - 1; env; lt; s - 1; \mathcal{A}_s) & \text{if } s > 0 \wedge (i', j') \in \mathcal{A}_s \\ + \sigma_s(a_{i'}, b_{j'}, env) \end{cases}$$

The new variant of *SHAgap*() is defined analogously. Now we can define the matrix *Dchain*() for chaining the strand pair alignments. At the end of its construction, the sheet is closed by pairing its first and last strands to create the barrel. To construct this, we need to keep track of the leftmost and rightmost strand alignments $\mathcal{A}_s^{\text{chain}}$ and $\mathcal{A}_s^{\text{cyc}}$ of the sheet. We further add two parameters, *ct* and *nos*. The variable *ct* is used to determine if the closing strand pair has been added or not. Here, $ct = c$ means that the sheet is not closed while $ct = l_f$ indicates that the barrel has been built. To control the number of strand in the barrel, we add the variable *nos* storing the number of strands in the β -sheet.

We now initialize the array *Dchain* for every i, j and any strand alignment $\mathcal{A}_s^{\text{cyc}}$ such that $\text{def}(a, \mathcal{A}_s^{\text{cyc}}) = [i..i']$ and $\text{def}(b, \mathcal{A}_s^{\text{cyc}}) = [j..j']$. This initializes the array to a non-barrel solution. Then

$$Dchain(i, j; \mathcal{A}_s^{\text{cyc}}; \mathcal{A}_s^{\text{cyc}}; c; lt; 1) = \text{LA}(i, |a|; j, |b|; lt; 1),$$

where *lt* represents the orientation environment (although the strand alignment has not yet been scored).

We can now describe rules used to build an unclosed β -barrel sheet. To account for the alignment of the first β -strand of this sheet (thus far unscored by SHA) we introduce the function $\text{SHA}_{\text{start}}(\mathcal{A}, nos)$ returning the cost of this alignment when $nos = 2$, and returning 0 otherwise. A function *prev*() returning the previous loop type is also used to alternate loop environments between both sides of the membrane. In addition, given two alignments $\mathcal{A}_s, \mathcal{A}'_s$, we say that $\mathcal{A}_s, \mathcal{A}'_s$ agree on the strands $i..i'$ in the first sequence and $j..j'$ in the second sequence, written $\text{agr}(\mathcal{A}'_s; \mathcal{A}_s; i, i'; j, j')$. With this notation, the recursion used to build the unclosed sheet is:

$$\begin{aligned}
Dchain(i, j; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; c; lt; nos) = & \\
& \max_{i', j', \mathcal{A}'_s, s, lt', env} \left(\begin{array}{l} \text{SHA}(i', i; j, j'; env; lt'; s; \mathcal{A}'_s) \\ + Dchain(r_1, r_2; \mathcal{A}'_s; \mathcal{A}_s^{\text{cyc}}; c; prev(lt); nos - 1) \\ + \text{SHA}_{start}(\mathcal{A}'_s, nos) \end{array} \right) . \\
& \text{with} \\
& \text{SHA}(i, i'; j, j'; lt'; s; \mathcal{A}'_s) > -\infty, \\
& \text{def}(a, \mathcal{A}_s) = [i..l_1] \cup [r_1..i'], \\
& \text{def}(b, \mathcal{A}_s) = [j..l_2] \cup [r_2..j'], \\
& \text{and } agr(\mathcal{A}'_s; \mathcal{A}_s; i, l; j, l')
\end{aligned}$$

Finally, we describe the recursions necessary to close the barrel and perform a sequence alignment of the N-terminal sequences. Since the antiparallel or parallel nature of the closing strand pair depends on the number of strands in the barrel, we separately define a function $ShAclose()$ which returns the folding energy of the parallel strand pairings of the leftmost and rightmost strands of the sheet if the number of strands nos is odd, and folding energy of the antiparallel strand pairings if nos is even.

$$\begin{aligned}
Dchain(i, j; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; l_f; lt) = & \\
\max \left\{ \begin{array}{l} \max \left\{ \begin{array}{l} Dchain(i+1, j; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; l_f; lt) + g_l(a_i, lt) \\ Dchain(i, j+1; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; l_f; lt) + g_l(b_j, lt) \\ Dchain(i+1, j+1; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; l_f; lt) + \sigma_l(a_i, b_j, lt) \end{array} \right. \\ \max_{i', j', env, nos} \left\{ \begin{array}{l} Dchain(i, i'; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; c; lt) \\ + ShAclose(i, i'; j, j'; env; s; \mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; dir(nos)) \end{array} \right. \end{array} \right.
\end{aligned}$$

This represents the final recursion used for consensus folding for some lt and $\mathcal{A}_s, \mathcal{A}_s^{\text{cyc}}$ with $agr(\mathcal{A}_s; \mathcal{A}_s^{\text{cyc}}; 1, i; 1, j)$, where $\text{def}(a, \mathcal{A}_s) = [1..i] \cup [r..i']$ and $\text{def}(b, \mathcal{A}_s) = [1..j] \cup [r..j']$. Sampling sequence/structure solutions can be done through a classical backtracking procedure (similar to Section 2.2.4).

We note that these equations assume that strand inclinations, identified by the shear number s , are independent. However, in practice this parameter must be used to determine when a strand pair can be concatenated at the end of an existing sheet to ensure the coherency of the barrel structure and conserve a constant inclination of the strands (see Figure 6-1).

Alignment Complexity Analysis

We present here a complexity analysis of the approach defined above, and then further discuss refinements made to improve efficiency. Let n and m denote the lengths of the two sequences. For the analysis, loop type, orientation, and shear number are negligible as they are constantly bounded. Thus, there are $O(n^2m^2)$ entries $\text{LA}(i, i', j, j')$ of loop alignments, each computed in constant time. For a fixed strand alignment \mathcal{A}_s , there are $O(n^2 \cdot m^2)$ many entries $\text{SHA}(i, i', j, j'; or; lt; s; \mathcal{A}_s)$, also computed in constant time using our recursions. Given that we model standard TMB structure, we assume that the maximum length of a strand alignment l_{\max} and the maximum number of gaps g_{\max} in a strand alignment can be bounded by small constants. Therefore we say that the number of such bounded alignments, ν , which are in $O(l_{\max}^{g_{\max}})$ is constant for fixed parameters l_{\max} and g_{\max} . As a result, there are $O(n^2m^2\nu)$ entries of $\text{SHA}(i, i', j, j'; or; lt; s; \mathcal{A}_s)$ in total.

In the chaining step there are $O(nm\nu^2)$ entries of $Dchain(i, j; \mathcal{A}_s, \mathcal{A}_s^{\text{cyc}}; ct, lt)$, each of which is computed by maximizing over left boundaries i' and j' , orientation, loop type, shear number and strand alignment of an entry SHA. There are $O(nm\nu)$ such combinations. The final cyclic closing is computed by searching over all $O(nm\nu)$ alignments $\mathcal{A}_s^{\text{cyc}}$ and pairs of positions i and j , where the last strand alignment ends. This results in a complexity of $O(n^2m^2 + n^2m^2\nu + n^2m^2\nu^3)$ in time and $O(n^2m^2 + n^2m^2\nu + nm\nu^2)$ in space.

To yield a practical results, this is reduced through the use of a threshold p_{cutoff} for the probabilities in our probabilistic contact map. As a result, the contact degree is bounded by $1/p_{\text{cutoff}}$ and the quadratically many contacts considered for the above analysis are thus reduced to linearly many “significant” ones.

Since we compute only entries of $\text{SHA}(i, i', j, j'; or; lt; s; \mathcal{A}_s)$ where all positions i, i', j and j' are within a narrow range r from a significant contact (p, p') , and r is bounded by the shear number s and g_{\max} , there remain only $O(4r^2nm\nu)$ entries. This means each entry can be computed in only $O(4r^2\nu)$ time due to the constant contact degree. Time and space complexity are thus reduced by a factor of $O(nm)$.

Since we also assume that no β -sheet bulges exist, all strand alignments have equal length and their gaps are located at the same position. This further restricts the choice of overlapping strand alignments \mathcal{A}_s , leaving the final complexity of our approach to be $O(n^2m^2 + 4r^2nm\nu + 4r^2nm\nu) = O(n^2m^2 + 4r^2nm\nu)$ in time and $O(n^2m^2 + 4r^2nm\nu + 4r^2nm\nu) = O(n^2m^2 + 4r^2nm\nu)$ in space.

6.3 Amyloid fibril consensus modeling

The concept of simultaneous alignment and folding of protein structures can similarly be applied to amyloid fibril schemas. However, while its implementation is may be simplified through the use of the recursive rules described in Section 2.3.3, the potential for algorithmic optimizations are more limited.

We have implemented consensus folding in our tool *AmyloidMutants* by effectively tracking the alignment and energetic scores for two sequences instead of one throughout every invocation of a *C-rule*, *M-rule*, or *N-rule* (Section 2.3.3). State space dimensionality is therefore doubled, and each state's score is calculated via a convex combination of contact probabilities and alignment matrix scores (similar to that used for the simultaneous alignment and folding of TMBs). Since this straightforward approach notably increases the time and space complexity of the algorithm, an optimization of this technique for amyloid fibrils is a matter of ongoing work.

Chapter 7

Evaluation of simultaneous alignment and folding

In this chapter we demonstrate the benefits of our ensemble algorithm for simultaneous alignment and folding, evaluating both the problems of pairwise sequence alignment and protein structure prediction. For these analyses, we focus on transmembrane β -barrel proteins using our implementation of partiFoldAlign. In summary, our sequence alignment accuracy performs comparably to existing alignment techniques, and significantly surpasses state-of-the-art alignment tools in the case of low homology sequences. It is also shown that consensus fold can better predict secondary structure when aligning proteins within the same superfamily. Finally, we propose a novel use of consensus folding for the case of amyloid fibril cross-seeding modeling and describe future applications.

We believe this technique to be generally applicable to many classes of proteins where the structure can be defined through a chaining procedure as described in Section 6.2.2. This could open new areas of analysis that were previously unattainable given current tools' poor ability to construct functional alignments on low sequence homology proteins.

7.1 Transmembrane β -barrel consensus modeling

In this section we validate the accuracy of our method in performing specific TMB sequence alignments and structural prediction. However, we begin with a description of the test dataset and scoring metrics used, as well as an analysis of the structural and energetic parameters used by our algorithm (described in Section 6.2.1).

7.1.1 Dataset and evaluation technique

Transmembrane β -barrels represent a particularly interesting protein class for study due to the relatively little that is known about their structure and their highly divergent sequences — posing difficulties for current alignment tools. Specifically, only approximately 20 non-homologous TMB structures have been solved via X-ray crystallography or NMR to date, and often TMB sequences can exhibit less than 20% sequence similarity, despite sharing structure and function.

To evaluate our approach we select 13 proteins from five superfamilies of TMBs found in the Orientation of Proteins in Membranes (OPM) database [112] (using the OPM database definition of class, superfamily, and family). This constitutes all solved TMB proteins with a single, transmembrane, β -barrel domain, excludes proteins with significant extracellular or periplasmic structure, and limits the sequence length to a computationally-tractable maximum of approximately 300 residues. With the assumption that structural alignment best mimics the intended goal of identifying evolutionary and functional similarities, we perform structural alignments between all pairs of proteins within large superfamilies, and across smaller superfamilies (28 alignments, with the breakdown illustrated in Table 7.1), and for sequence alignment testing purposes consider these the “correct” pairwise alignment. For generate good structural alignments, the *Matt* [122] algorithm is used, which has demonstrated state-of-the-art accuracy. We then sort the resulting alignments (using *Matt*) according to relative sequence identity for presentation. Here sequence identity is defines as

$$\text{Sequence Identity } \% = \frac{\text{Identical positions}}{\text{aligned positions} + \text{internal gap positions}}.$$

Sequence alignments derived from consensus folds are then compared against structural alignments using the Q_{Cline} [39, 95] scoring metric, restricted to transmembrane regions as defined by the OPM (since structural predictions in the algorithm only contribute to transmembrane β -strand alignments; coils are effectively aligned on sequence-alone). A more simplistic metric to gauge alignment accuracy would be the $Q_{combined}$ score

$$Q_{combined} = \frac{\# \text{ correct pairs}}{\# \text{ unique pairs in sequence \& structure alignments}}.$$

However, we instead focus on Q_{Cline} percent accuracy since it measures the combined under- and over-prediction of aligned pairs in a more fair manner, accounts for off-by- n alignments. Such shifts often occur from energetically-favorable off-by- n β -strand pairings that would still remain useful

Number of strands	Seq. identity range	Pairwise seq. identity	Protein pair	Classification
8-stranded	0-4%	4%	1BXW-1THQ	OMPA-like / OMPA-like (LAA)
		4%	1QJ8-1THQ	
		4%	1THQ-2F1V	OMPA-like (LAA) / OMPA-like
		4%	2ERV-2F1V	
	5-9%	6%	1P4T-1THQ	OMPA-like / OMPA-like (LAA)
		6%	1THQ-2ERV	OMPA-like (LAA) / OMPA-like (LAA)
		6%	1THQ-2JMM	OMPA-like (LAA) / OMPA-like
		6%	1BXW-2ERV	OMPA-like / OMPA-like (LAA)
		6%	1QJ8-2ERV	
		7%	2ERV-2JMM	OMPA-like (LAA) / OMPA-like
		7%	1P4T-2ERV	OMPA-like / OMPA-like (LAA)
		8%	2F1V-2JMM	OMPA-like / OMPA-like
		9%	1BXW-2F1V	
		9%	1P4T-2JMM	
	10%	1P4T-2F1V		
	10-20%	10%	1QJ8-2F1V	OMPA-like / OMPA-like
		14%	1P4T-1QJ8	
15%		1BXW-1P4T		
15%		1QJ8-2JMM		
17%		1BXW-1QJ8		
50%	50%	1BXW-2JMM		
10-stranded	6%	6%	1I78-1K24	OMPT-like / OMPT-like
12-stranded	0-5%	3%	1TLY-2QOM	Nucleoside-spec. porin / Autotransp.
		3%	1QD6-1TLY	OM phosph. / Nucleoside-spec. porin
		5%	1QD6-2QOM	OM phosph. / Autotransp.
	6-10%	6%	1QD6-1UYN	OM phosph. / Autotransp.
		6%	1TLY-1UYN	Nucleoside-spec. porin / Autotransp.
		9%	1UYN-2QOM	Autotransp. / Autotransp.

Table 7.1: **Breakdown of OPM database TMB pairwise alignments:** For all 28 we list their corresponding sequence identities and subfamily classifications. LAA distinguishes a family within the OMPA-like superfamily of proteins involved with Lipid A Acylation

alignments for many purposes. The Q_{Cline} parameter ϵ is chosen to be 0.2, which allows alignments displaced by up to five residues to contribute (proportionally) toward the total accuracy. The higher the Q_{Cline} score, the more closely the alignments match (ranging $[-\epsilon, 1]$).

To judge the accuracy of our consensus structure predictions single-sequence structure predictions, we use the same OPM database of proteins described above. For each of the 13 proteins, a structure prediction is computed using the exact same ensemble structure prediction methodology as in the consensus predictions, only applied to a single sequence. The transmembrane-region Q_2 secondary structure prediction score between the predicted structures and the solved PDB structure (annotated by STRIDE [76]) can then be computed; where $Q_2 = (TP + TN) / (\text{sequence length})$.

7.1.2 Model parameter selection

Our approach for modeling TMBs requires both structural and energetic constants to parameterize the ensemble (see Sections 3.1.1 and 6.2.1). The choice of structural parameters, such as the allowable β -strand and coil region lengths, as well as shear numbers can be assigned based on biological quantities such as membrane thickness, etc. However, other algorithmic parameters, such as the pairwise contact threshold (which filters which β -strand pairs are used in the alignment, see Section 6.2.1), the sequence alignment gap penalty, the choice of substitution matrix, and the α balance parameter require selection without as clear a biological interpretation. For example, the substitution matrix used in this evaluation is a combination of the BATMAS [174] matrix for transmembrane regions, and the BLOSUM [83] matrix for coils.

For the evaluation described below, we choose three sets of structural parameters according to 3 protein classifications: 8-, 10-, or 12-stranded TMBs (Table 7.2). Parameters were again chosen from general TMB characteristics [164], however in practice these would be derived from existing experimental data already available for a protein under test. A well-formulated machine-learning approach for parameter optimization would also make an ideal fit for this problem. Further, we varied the β -strand pair probability threshold used in the initial step of the algorithm and the α score-balancing parameter based on whether the pairwise sequence identity was above or below 10%. Below 10%, $p_{\text{cutoff}} = 1 \times 10^{-5}$ and $\alpha = 0.6$, and above, $p_{\text{cutoff}} = 1 \times 10^{-10}$ and $\alpha = 0.7$. This reduces signal degradation from low-likelihood β -strand pairs with very distant sequence similarities, and boosts the contribution of the structural predictor when less sequence homology can be exploited. As seen in Figure 7-1, consensus predictions from lower α parameters more closely resemble predictions based solely on structural scores, and thus, an optimal alignment should correlate α with sequence homology.

Constraint	8-strand	10-strand	12-strand
Number of β -strands	8	10	12
Min/max β -strand length	8-13	10-14	9-14
Min/max shear value	0-2	0-3	0-3
Min/max periplasmic loop length	2-15	2-10	2-20
Min/max extra-cellular loop length	2-35	20-45	5-40

Table 7.2: **TMB structural constraints used for consensus folding:** Structural constraints were chosen according to one of three TMB classifications.

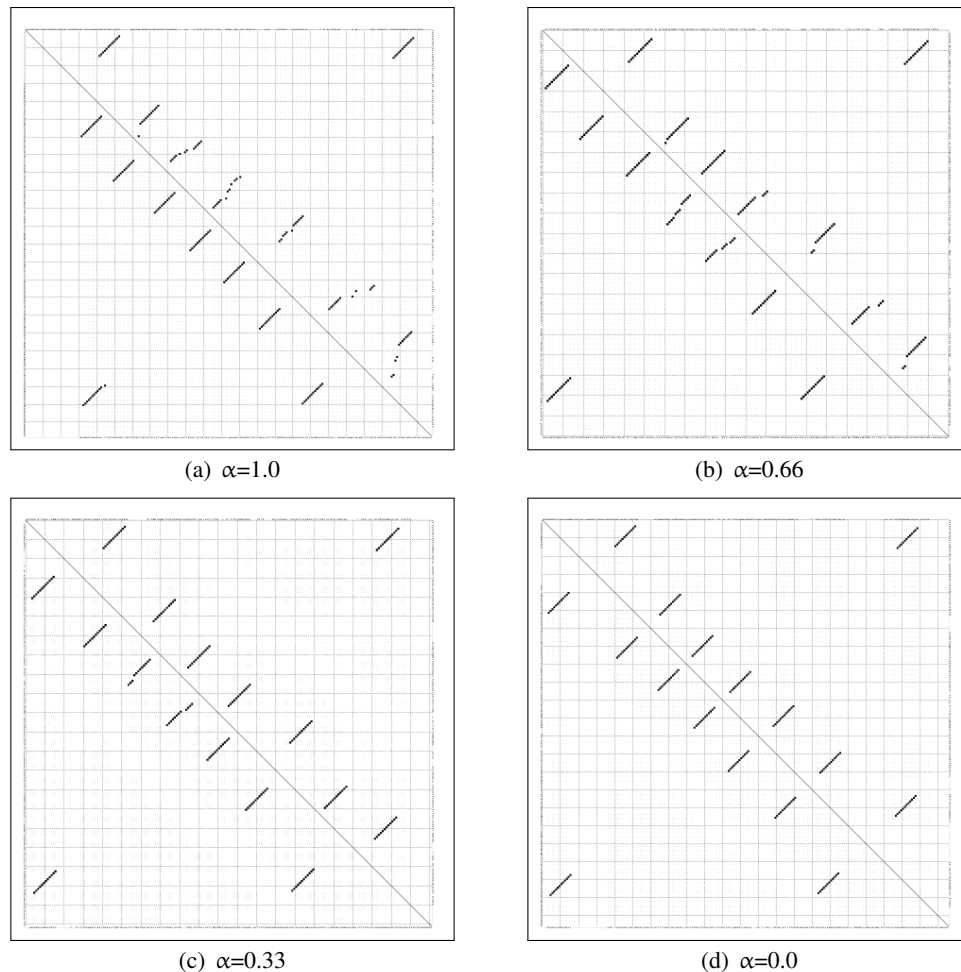


Figure 7-1: **Alignment depends on a balance between sequential and structural information:** Stochastic contact maps from a `partiFoldAlign` prediction of the proteins 1BXW and 2F1V. For each of the four plots, the sequence of 1BXW and 2F1V is given on the axes (with gaps), and high probability residue-residue interactions indicated for 1BXW on the lower left half of the graph and 2F1V on the upper right half (i.e., the single-sequence probabilistic contact maps). Structural contact map alignment can be judged by how well the plot is mirrored across the diagonal. **(a)** ($\alpha = 1.0$) shows an alignment which ignores the contribution of the structural contact map, while **(d)** ($\alpha = 0.0$) shows an alignment wholly-dependent on the structural contact map, and ignorant of sequence alignment information.

7.1.3 Validation of alignment accuracy under low sequence identity

Here we compare our consensus fold-derived pairwise sequence alignments against the output of two leading sequence alignment algorithms: EMBOSS [153] and MUSCLE [60,61]. EMBOSS may be considered the best Needleman-Wunsch style global sequence alignment algorithm (a straightforward, widely applicable method of alignment), while MUSCLE is widely thought the most accurate of the “fast” alignment tools, though it incorporates several position-specific gap penalty

heuristics similar to MAFFT and LAGAN [23] which our approach does not include (although no technical reasons prevent incorporation of MUSCLE's gap penalty heuristics into our model).

Since our algorithm utilizes Needleman-Wunsch style dynamic programming, comparisons between EMBOSS and our partiFoldAlign tool represent a fair analysis of what simultaneous folding and alignment algorithms specifically contribute to the problem. Although we note that while we use a bipartite, BATMAS/BLOSUM, sequence alignment scoring matrix, and EMBOSS uses only BLOSUM, Forrest et al. [72] have shown that BATMAS-style matrices do not show improvement for EMBOSS-style algorithms (likely due to a lack of extra information directing which matrix to use when). Comparisons with MUSCLE alignment scores are included to portray the practical benefits of consensus folding.

Figure 7-2 presents transmembrane Q_{Cline} accuracy scores for EMBOSS, MUSCLE, and partiFoldAlign across 27 TMB pairwise alignments. (The absent 28th alignment, between 1BXW and 2JMM (50% sequence-homologous), is aligned with a nearly-perfect Q_{Cline} score of 0.98 by all three algorithms). Results are separated into the 3 categories according to the number of circling strands within a protein's β -barrel: seven 8-stranded OMPA-like proteins account for 21 alignments, two 10-stranded OMPT-like proteins account for one alignment, and finally, four 12-stranded Auto-transporters, OM phospholipases, and Nucleoside-specific porins make up the final six alignments (see Table 7.1). Equal-sized clusters of pairwise alignments are then generated and ordered according to sequence identity, with cluster mean Q_{Cline} and standard deviation reported. All individual alignment-pair statistics for both the Q_{Cline} and $Q_{combined}$ metrics can be found in Figure 7-3 and Figure 7-4, respectively.

For all TMBs, partiFoldAlign alignments improve upon EMBOSS alignments by an average Q_{Cline} of 16.9% (4.5x). Most importantly, partiFoldAlign greatly improves upon the EMBOSS Q_{Cline} score for alignments with a sequence identity lower than 9% (by a Q_{Cline} average of 28%), and roughly matches or improves 24/28 alignments overall. If 12-strand alignments are excluded, which align proteins across different superfamilies, intra-superfamily alignments exhibit improvements over EMBOSS by a Q_{Cline} difference of 20.3% (27.4% versus 7.1%). Compared with MUSCLE alignments, our approach achieves a 4% increased Q_{Cline} on average, even without any gap penalty heuristics.

Although here we demonstrate *pairwise* consensus folding, we also believe this approach can translate into considerable improvements in multiple sequence alignments. This is because many multiple alignment procedures use pairwise alignment information at their core [95]. Such an ex-

tension would be an obvious next step for our approach to be added in combination with other, more elaborate techniques found in sequence alignment algorithms (e.g., MUSCLE).

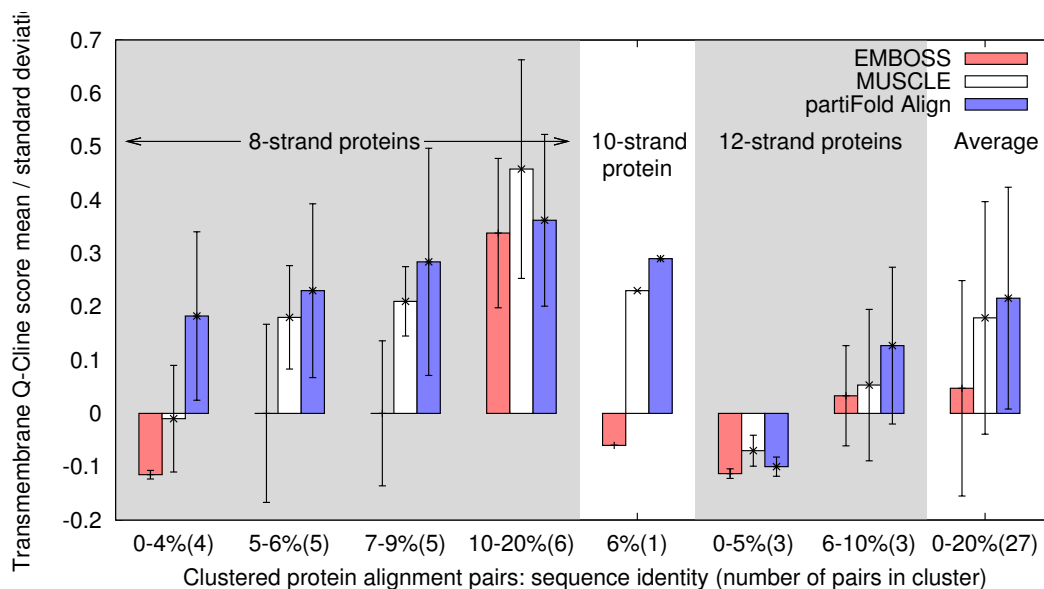


Figure 7-2: **partiFoldAlign alignment accuracy for 8-, 10-, and 12-stranded TMBs:** Mean and standard deviation Q_{Cline} scores plotted. Each of the 3 categories of proteins are clustered and ordered according to sequence identity, with the number of alignments in each cluster in parentheses. Note: By definition, Q_{Cline} scores range between $-\epsilon$ and 1.0, where $\epsilon = 0.2$; negative indicating very poor alignments.

7.1.4 Secondary structure prediction accuracy of consensus folds

Consensus folds can not only be used to derive more accurate sequence alignments, but also serve as a mechanism for incorporating sequence homology information into structural prediction. Here we compare the accuracy predicted structures of TMB proteins using consensus folding against identical predictions using a single sequence alone. Table 7.3 lists the Q_2 accuracies computed from consensus folding of all pairs of TMB sequences within the same n -stranded category. For each protein, the Q_2 score from the single sequence minimum folding energy structure is given, and compared against the Q_2 score from the best alignment partner and the average Q_2 score obtained when aligning that protein with all others in its category. Single sequence structure prediction is performed using a modification of the exact same partiFoldAlign algorithm, removing the contribution of sequence.

The results for 8- and 10-stranded categories show a clear improvement (more than 8%) by the best consensus fold in 6 of 9 instances (1P4T, 2F1V, 1THQ, 2ERV, 1K24, 1I78), and roughly equiv-

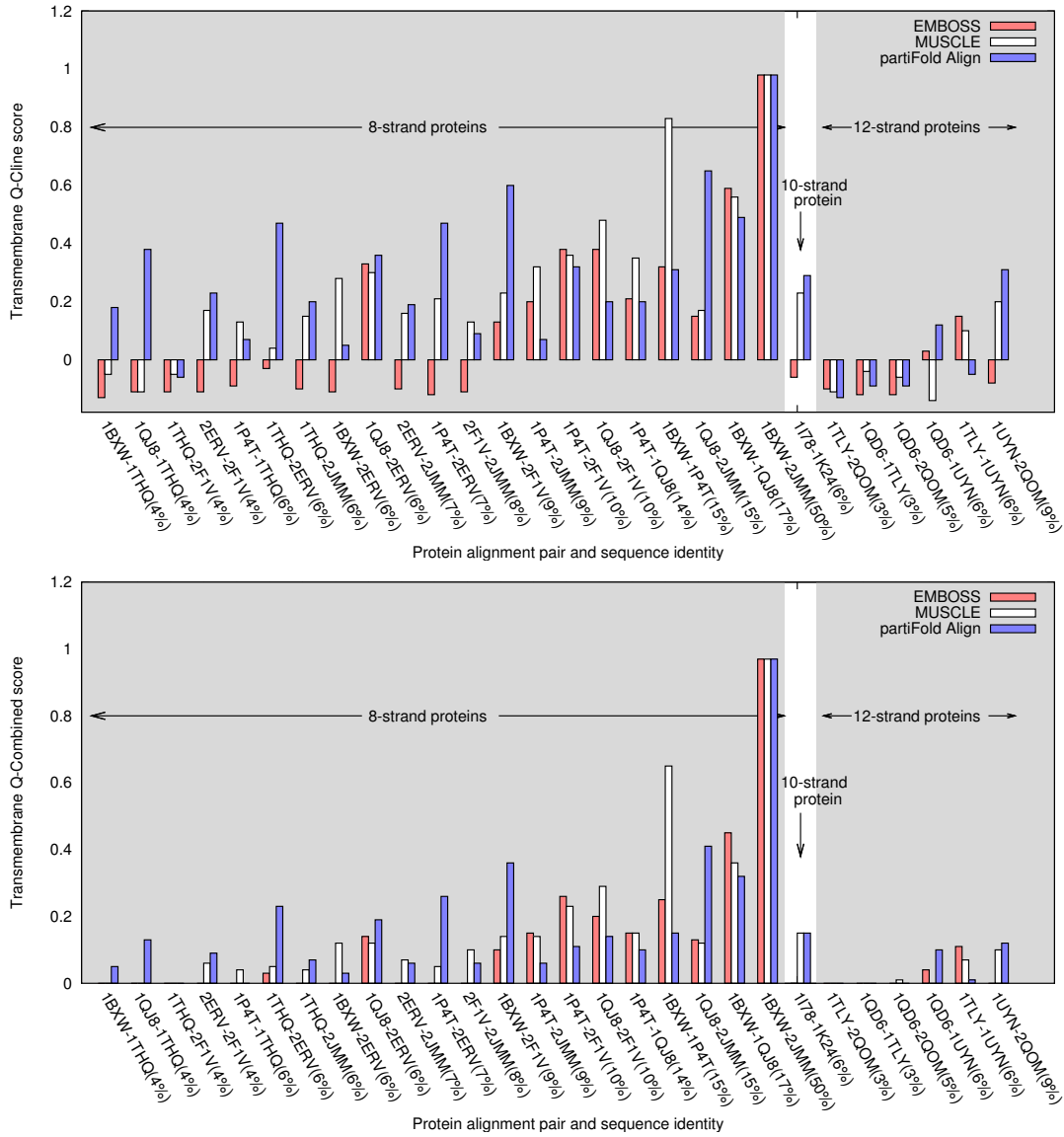


Figure 7-3: **partiFoldAlign alignment accuracy for individual TMBs (transmembrane only):** Listed in order of increasing sequence identity, all 28 TMB pairwise alignments for the 3 classes of proteins, along with their corresponding transmembrane Q_{Cline} and $Q_{combined}$ score. From this we see that Q_{Cline} and $Q_{combined}$ mirror the same general trend.

alent results for the remaining 3 (2F1V, 1K24, I178). Further, on average, nearly all proteins show equivalent or improved scores when aligned with any other protein in their group, with the exception of 1BXW. However, the single sequence structure prediction Q_2 for 1BXW is not only high, but significantly higher than all other 8-stranded proteins; the contact maps of any other aligning partner may simply add noise, diluting accuracy. Conversely, the proteins which have poor single sequence structure predictions benefit the greatest from alignment (e.g., 2F1V). This relationship is not unidirectional, though, as we see that the consensus fold of 1K24 and I178 improves upon both

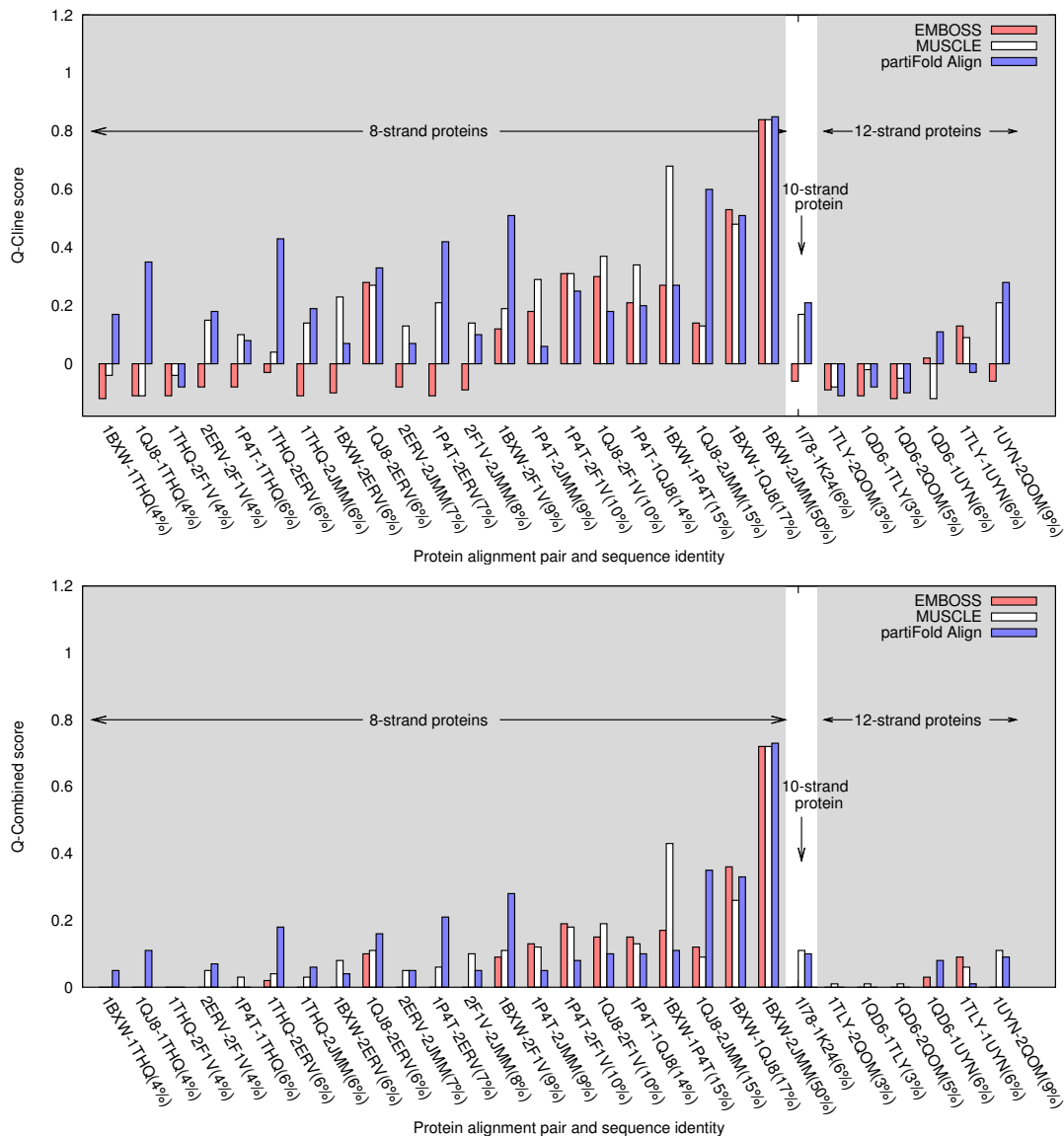


Figure 7-4: **partiFoldAlign alignment accuracy for individual TMBs (whole protein):** Listed in order of increasing sequence identity, all 28 TMB pairwise alignments for the 3 classes of proteins, along with their corresponding whole-protein Q_{Cline} and $Q_{combined}$ score. From this and Figure 7-3 we see that transmembrane and whole-protein alignment scores follow the same trend.

proteins' single sequence structure prediction.

In contrast, the results compiled on the 12-strands category do not show any clear change in the secondary structure accuracy. However, recalling that this category covers 3 distinct superfamilies in the OPM database, such results may make sense. The Autotransporter, OM phospholipase, and Nucleoside-specific porin families all exhibit reasonably different structures, and perform quite unrelated tasks. Further, the ensemble representation used in partiFoldAlign does not take into account β -strand extensions (see Section 2.2.2), which also reduces the structure prediction accuracy

Category	PDB id	single seq.	consensus	
			best	average
8-stranded	1BXW	72	70(-2)	63(-9)
	1P4T	60	68(+8)	58(-2)
	1QJ8	65	68(+3)	66(+1)
	2F1V	47	63(+22)	62(+15)
	1THQ	50	69(+13)	52(+2)
	2ERV	57	67(+10)	59(+2)
	2JMM	62	65(+3)	62(+0)
10-stranded	1K24	60	69(+9)	69(+9)
	1I78	76	83(+7)	83(+7)
12-stranded	1QD6	54	61(+7)	56(+2)
	1TLY	59	59(+0)	58(-1)
	1UYN	56	56(+0)	53(-3)
	2QOM	51	55(+4)	53(+2)

Table 7.3: **partiFoldAlign secondary structure assignment accuracy:** Shown are Q_2 percentages of secondary structure predictions (transmembrane and non-transmembrane regions). The third column reports the performance of a single sequence prediction (i.e., involving no alignments). The fourth and fifth columns report the best and average Q_2 scores of a consensus structure over all possible alignment pairs for this PDB ID.

of these more complex TMBs.

From this benchmark we conclude that the consensus folding approach can be used to improve the structure prediction of low homology sequences, provided both sequences share some putative evolutionary connection (such as belonging to the same superfamily). However, we emphasize the importance parameter selection may play in these results; a different parameter selection method may enable accuracy improvement for higher-level classes of proteins.

7.2 Amyloid fibrils consensus modeling

Unfortunately, a thorough evaluation of our approach’s sequence alignment accuracy or consensus structure prediction sensitivity in amyloids is not possible. This is due to the paucity of amyloid fibril structural data, the drastic sequence differences between known amyloid conformations, and the existence of the amyloid strain phenomena (in other words, multiple sequences can fold to one structure and one sequence can fold to many structures). However, biological investigations into many amyloidal proteins could be aided through the application of our methodology to predict amyloid consensus folds and corresponding sequence alignments. This may even help to speed the discovery of new physiological amyloid structures.

One particularly interesting phenomenon worth study using this approach is that of amyloid fibril “cross-seeding.” In this biological process some prion proteins that fold into amyloid structure can “cross-seed” homologous proteins in other organisms [179, 180], initiating a fibril assembly that is templated by the original protein conformation — sometimes even in cases with a sequence identity as low as 38% [206]. The most famous case of cross-seeding may be the ability of “mad cow” bovine prion proteins to infect humans, causing variant Creutzfeld-Jakob disease. However, the specific structure of these interactions, nor the conformation of possible intermediate states, is unknown. Predicted consensus structures of amyloid fibril sequences describe the set of likely amyloid conformations two sequences can both adopt, suggesting potential cross-seeded states that may or may not differ from the conformation of either sequence’s individual, homogeneous fibril state. Understanding this molecular state could lead to targeted therapeutics, and greatly advance the study of prions and *in vivo* aggregation in general.

Chapter 8

Ensemble prediction of folding dynamics

In this chapter we present the use of ensemble techniques to solve an altogether different problem: the prediction of protein folding pathways. Unlike earlier chapters, which model proteins at a steady-state assuming a Boltzmann distribution equilibrium, here we use are interested in kinetics. Specifically, we model the transition from random coil to native state as a Markov process, and use ensemble predictions and a master equation to simulate population dynamics of folding over time. This algorithmic framework has been implemented with collaborators as a publically-accessible web-based tool named tFolder¹. Through this demonstration, we argue for the general applicability of ensemble modeling concepts throughout many problems in bioinformatics.

8.1 Goals and overview

Protein folding and unfolding is a key mechanism used to control biological activity and molecule localization [57]. The simulation of folding pathways is thus helpful to decipher cell behavior. Classical molecular dynamics (MD) methods [97] have been used to produce reliable predictions of folding pathways, but unfortunately the heavy computational load required by these techniques limits their application to inputs tens of amino acids long and prevents their application to large sequences (i.e., hundreds of amino acids). Our goal is instead to model protein folding kinetics using the kinds of coarse-grained representations introduced earlier in this work, estimating approximate folding pathways.

Simplified representations of protein structure motions have been widely used to circumvent computational limitations [109], most recently through the motion planning techniques of Amato et

¹Available at <http://csb.cs.mcgill.ca/tfolder/>

al. [3, 177]. These are significantly faster than classical MD techniques, but requires the three-dimensional structure of the native state to compute potential intermediate structures and unfolding pathways. Thus such methods cannot be applied to proteins with unknown structures, and few insights can be gained into off-pathway kinetics, such as aggregation. In fact, nearly all previously described methods suffer from a common difficulty: efficiently sampling the conformational landscape. MD algorithms explore the landscape through force-directed search and progressive modification of the structure, while motion planning techniques predict structural intermediates only in terms of the known native fold. Recently, Hosur et al. [88] have combined motion planning techniques with machine-learning to model proteins as an ensemble, but this approach is effective only in the local neighborhood of the input structure. Similarly, Faccioli et al. proposed a solution of the Fokker-Planck equation to compute dominant protein folding pathways [65], but the mentioned efficiency limitations remain.

These obstacles have been addressed by the development of ensemble modeling techniques for RNAs [53, 121, 184] and for TMBs and amyloid fibrils (introduced in this thesis). Further, Wolfinger et al. [211] has demonstrated how an RNA energy landscape can be constructed by connecting composable ensemble states and estimating transition rates. The resulting ordinary differential equation (ODE) system can be solved to predict and characterize folding pathways. The method has since been improved to analyze the motion of larger molecules [176].

We adopt this general approach and expand the methodology to explore the folding pathways of proteins. We design an algorithm to calculate the partition function of an ensemble, sample, and cluster configurations according to contact distance metrics. We associate each cluster with an intermediate folding state and use the difference between cluster energy scores to compute transition rates and build an ODE system modeling folding pathways. Solving this ODE system estimates the distribution of conformations over folding time. As a proof of principle we model simple single- β -sheet proteins rather than much more complicated TMBs or amyloid fibrils, however we believe that the algorithmic integration of ensemble prediction with Markovian dynamics can be applied to many other problems.

This methodology is meant to reconcile the MD and motion planning approaches for studying folding pathways. With this approach, pathway simulations can be performed in minutes on proteins with unknown structure (although in tFolder, restricted to single β -sheets). Further, although our approach only predicts coarse folding transitions, its strength lies in its ability to quickly separate conformational transitions that are critical to folding from those transitions that could simply result

from minor structural fluctuations. This complements the use of MD simulations as MD can be used to explore nuanced structural interactions that certainly occur near a transition highlighted by our coarse model. Section 8.2 describes the algorithm and we validate the method in Section 8.3 by applying it to predict the folding pathways of the well-studied B1 domain of Protein G.

8.2 Modeling single β -sheet protein dynamics

In this section we introduce a new coarse grained structural representation for single β -sheet structures of arbitrary parallel or antiparallel composition, named permutable β -sheet schemas (or templates). Using this, we describe the new algorithmic techniques required to construct a folding pathways from standard ensemble methods (introduced in Chapter 2). Our approach proceeds in three steps:

1. Given an arbitrary peptide sequence, we compute the partition function of all possible β -sheet structures and sub-structures using permutable β -sheet schemas.
2. We energetically sample conformations from the predicted ensembles.
3. We compute conformational compatibility metrics for all samples, derive the likelihood of dynamic state-to-state transitions, and assemble a set of folding paths using the Fokker-Planck equation.

Predicted pathways are then ranked from an unfolded conformation to a fully folded conformation.

8.2.1 Representing permutable β -sheet ensembles

We introduce the concept of permutable β -sheet schemas to enable the calculation of the partition function of a single β -sheet with an arbitrary number of β -strands and any combination of parallel or antiparallel pairing configurations. Importantly, this removes sequence order dependencies between β -strand/ β -strand interactions via an enumeration of all possible permutations, rather than a fixed subset of allowable permutations as for amyloid fibril schemas (Section 2.3). Necessarily, this enumeration increases the computational complexity.

Protein conformations are represented using the same coarse grain super-secondary structure model used for TMBs and amyloids fibrils (Chapter 2), identifying residue/residue pairing partners in β -sheets as well as coil region assignments. To efficiently encode shape variants, each β -strand is labeled by sequence order $\{1\dots n\}$ to allow discrete enumeration, and a signed permutation is

defined such that each β -strand is assigned to be parallel or antiparallel relative to the first strand in the sheet (Figure 8-1). As with the case of TMBs and amyloid fibrils, we impose steric and biologically-inspired constraints on the recursive description of potential β -sheet structures to limit the exploration of unrealistic conformations and minimize computation time. This also enables directed investigation into specific structural motifs. For the permutable β -sheet schemas used here, these parameters include a minimum and maximum β -strand length, a maximum shear between neighboring β -strands, and a minimum inter- β -strand loop size. This last loop size parameter is particularly important to remove infeasible physical conformations since we dissociate strand ordering from sequence — for example if β -strands 1 and 4 in Figure 8-1 had too short a coil between them.

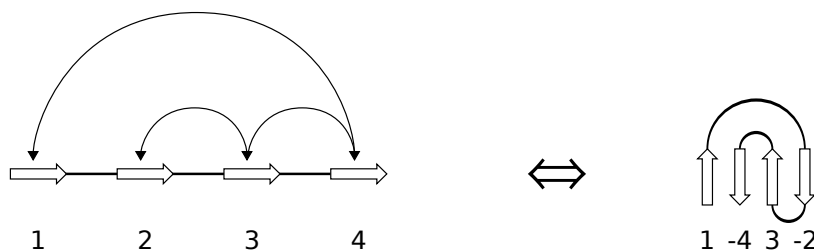


Figure 8-1: **Permutable β -sheet schemas encoding using signed permutation:** β -strands are ordered according to sequence, and permutations list the β -strands in the order that they occur in the β -sheet. The sign indicates whether the strand is parallel (+) or antiparallel (-) relative with the first β -strand.

8.2.2 Computing the partition function

We define the Boltzmann partition function of a single permutable β -sheet schema in much the same way as before (Section 2.1) where for any structural state s , $\mathcal{Z} = \sum_{i=1}^n e^{-\frac{E_{s_i}}{RT}}$, with the relative likelihood of any structure state s given by $p(s_i) = \frac{e^{-E_{s_i}}}{\mathcal{Z}}$. Similarly, our energetic model uses the same structure as described in Section 2.4.

We recursively define the energy of a permutable β -sheet structure with n strands as $E(s_n) = E(s_{n-1}) + \text{Pairing}(s_{n-1}, s_n)$, where $E(s_{n-1})$ is the interaction energy between the first $n - 1$ strands, and $\text{Pairing}(s_{n-1}, s_n)$ is the energy of the pairing of strand $n - 1$ with strand n (See Figure 8-2). To exploit the shared structure between instances in the ensemble, the result of each recursive call is stored in a memoization table indexed by the parameters of the call. Thus, subsequent recursive calls made with the same parameters perform a table lookup instead of re-computing the value of the recursion.

$$E(S_n) = E(S_{n-1}) + \text{Pair}(s_{n-1}, s_n)$$

Figure 8-2: **Decomposition of recursion for permutable β -sheet schemas:** The energy of each structure is recursively defined as the sum of the contribution of the current subsolution along with the next pairwise β -strand/ β -strand interaction.

For a sheet of n strands, our memoization table has n rows, where the k^{th} row has entries corresponding to valid configurations of the first k strands. For the k^{th} strand, these configurations are partitioned by the location of four indices k_1, k_2, k_3, k_4 , which denote the boundaries of the region occupied by the k strands (Figure 8-3). To begin, the algorithm enumerates all possible positions of the first two β -strands, and for each stores the strand pair interaction energy in entry $E_{2_1 2_2 2_3 2_4}$ of the table. For each subsequent strand k , the value of $E_{k_1 k_2 k_3 k_4}$ is computed as:

$$E_{k_1 k_2 k_3 k_4} = \sum_{i_1 i_2 i_3 i_4} E_{i_1 i_2 i_3 i_4} + \text{Pairing}(i, k),$$

where i_1, i_2, i_3, i_4 are enumerated for all valid boundaries of the preceding strands, given the boundaries of the k^{th} strand. The partition function \mathcal{Z} over all permutations is thus calculated by summing over all possible settings of n_1, n_2, n_3, n_4 ,

$$\mathcal{Z} = \exp \left(- \sum_{n_1 n_2 n_3 n_4} E_{n_1 n_2 n_3 n_4} \right).$$

Finally, this calculation is performed across all possible permutation β -sheet schemas. That is to say, separate ensembles are calculated for β -sheets containing 1, 2, 3, ... strands, and for each different signed permutation β -strand interaction topology.

8.2.3 Boltzmann distribution sampling

Sampling conformations from the ensembles is also carried out in a manner much similar to that described for amyloid fibrils (Section 2.3.4). Using the notation above, we sample via a traceback through the memoization table, where, at each i^{th} step we sample from the indices within the first i

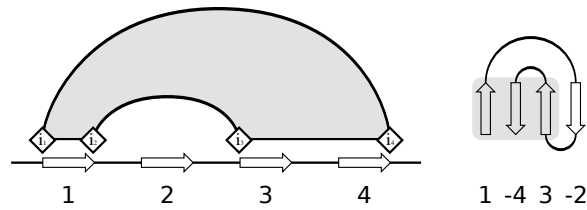


Figure 8-3: **Indices of intermediate permutable β -sheet structures:** Here we illustrate the indices used to store the energies of intermediate structures during the recursion.

strands according to the Boltzmann energy of these i -stranded structures (Figure 8-4).

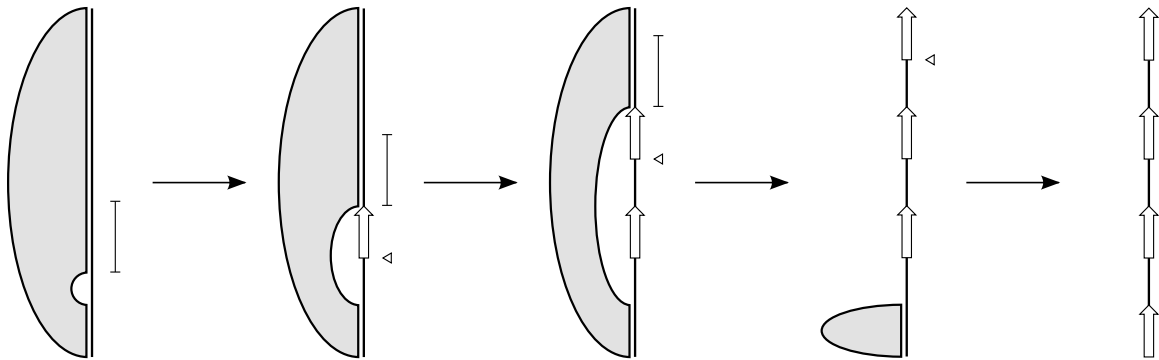


Figure 8-4: **Illustration of permutable β -sheet schemas sampling:** During each step of the sampling procedure, the location of a single β -strand is sampled from the region indicated by the vertical bars. The triangles denote the location of the β -strand sampled during the previous step.

8.2.4 Predicting dynamics using the Fokker-Planck equation

Conceptually, we model the folding process as a path through a graph of varyingly folded protein conformation states. In this graph, states that can inter-convert via a folding event are connected by an edge, analogous to work with RNA described previously [211]. The Boltzmann-weighted ensemble sampling method described above can provide a means to populate this graph with energetically accessible conformational states. However we must now propose a method to determine the inter-connectivity between states. This is based on two broad rules: for every pair of states we add an transition edge if

1. the states come from compatible permutation schema topologies, and
2. the states show structural similarity.

To satisfy the first requirement, we assert that two permutation schema topologies are compatible if they are identical to each other, modulo the addition or removal of a single strand pairing. For example, this could involve the growth of a core β -sheet structure, or the nucleation of an independent β -strand pair forming a (conceptually) separate β -sheet (see Figure 8-5). To satisfy the second criteria we use a residue pairing contact distance metric to deem two structure sufficiently structurally similar. Note that this choice in metric can strongly impact the resulting folding pathway predictions — in our tests pair contact distance performed the best over segment overlap [216] and mountain metric [127].

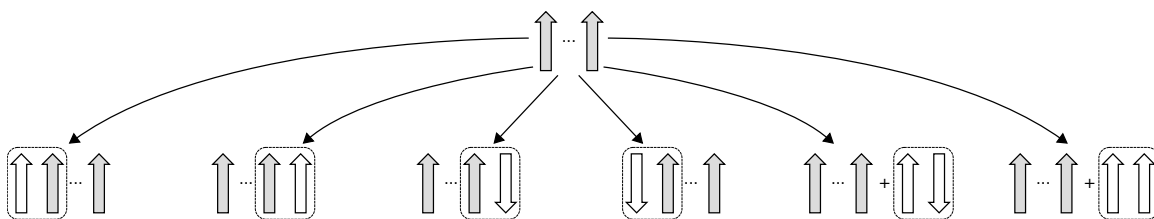


Figure 8-5: **Illustration of permutable β -sheet schema compatibility:** Compatible topologies of any given state (shaded gray) result from the addition of a single β -strand pairing (dashed box). Additions can either be extensions of a current β -sheet or the addition of an unconnected β -strand pairing (indicated by '+').

Given the construction of this graph, the change in the probability of the system being in state i at time t is calculated from the total flux into and out of state i

$$\frac{dp_i}{dt} = \sum_{j \in X} r_{ij} p_j(t),$$

where p_i is the probability of state i , X is the state space, and r_{ij} is the rate of transition from state i to state j . Given that two states are connected in the graph, the rate at which two states inter-convert is proportional to the difference between free energies of the states (ΔG) so that the system tends toward energetically favorable states. We calculate the transition rate r_{ij} between states i and j using the Kawasaki rule to align best with this free energy model (with parameter r_0 to scale the time dimension):

$$r_{ij} = r_0 \exp(-\Delta G_{ij}/2RT).$$

A comparison of the use of the Kawasaki rule versus the Metropolis rule is given by Sauerwine and Widom [160].

The dynamics of the system are calculated by treating the folding process as a continuous time discrete state Markov process. Given the matrix of folding rates R , where $R_{ij} = r_{ij}$ and initial state density $\vec{p}(0)$, the distribution over states $\vec{p}(t)$ of the system at time t is given by the explicit solution to the system of linear differential equations,

$$\vec{p}(t) = \exp(Rt) \vec{p}(0).$$

Since we sample hundreds of states from each β -strand topology, we partition the state space into macro states using clustering, in order to work with a tractably sized system. Under this approximation, we consider two clusters in the graph to be connected if the minimum distance between any two states from each cluster are connected. We define the ensemble free energy difference ΔG_{ij} between two macrostates i and j by summing over the states from which they are composed.

$$\Delta G_{ij} = E(\chi_i) - E(\chi_j) = \sum_{x \in \chi_i} E(x) - \sum_{x \in \chi_j} E(x).$$

We note that although this approximation lessens the computational burden, it means that the granularity achievable by our simulation is at the level of macrostates. Further, we point out that energy barriers and transition states are not explicitly incorporated into the model since entire β -strands are either added or removed between states without consideration of partially-formed intermediates.

8.3 Evaluation of single β -sheet protein dynamics prediction

In this section we validate the accuracy and utility of our algorithms through the implementation tFolder. Note, however, that the accuracy of the predicted folding pathways entirely on the accuracy of the underlying structural predictions. Therefore we begin with a comparison of ensemble super-secondary structure predictions against a data set of known single β -sheet proteins. We then evaluate tFolder folding pathways predictions by comparing against the extensively studied Protein G. In this case, the predicted folding pathways mirror other published reports. Finally, for our work to truly complement the use of MD simulations, these predictions must be achievable very efficiently. For this end we provide a brief empirical runtime analysis.

8.3.1 Validation of super-secondary structure prediction accuracy

To evaluate the contact prediction performance of tFolder, we tested it using a set of proteins selected from the Protein Data Bank which had single β -sheets with 4–6 β -strands, 100 residues or smaller, and a sequence identity less than 30%. As in Section 3.1.1 we compare against BETApro, along with SVMcon, two state-of-the-art β -sheet structure predictors. To ensure a fair comparison with these machine-learning based tools, we removed from this set those proteins that were used in the training sets of these methods. This resulted in a data set of 16 proteins.

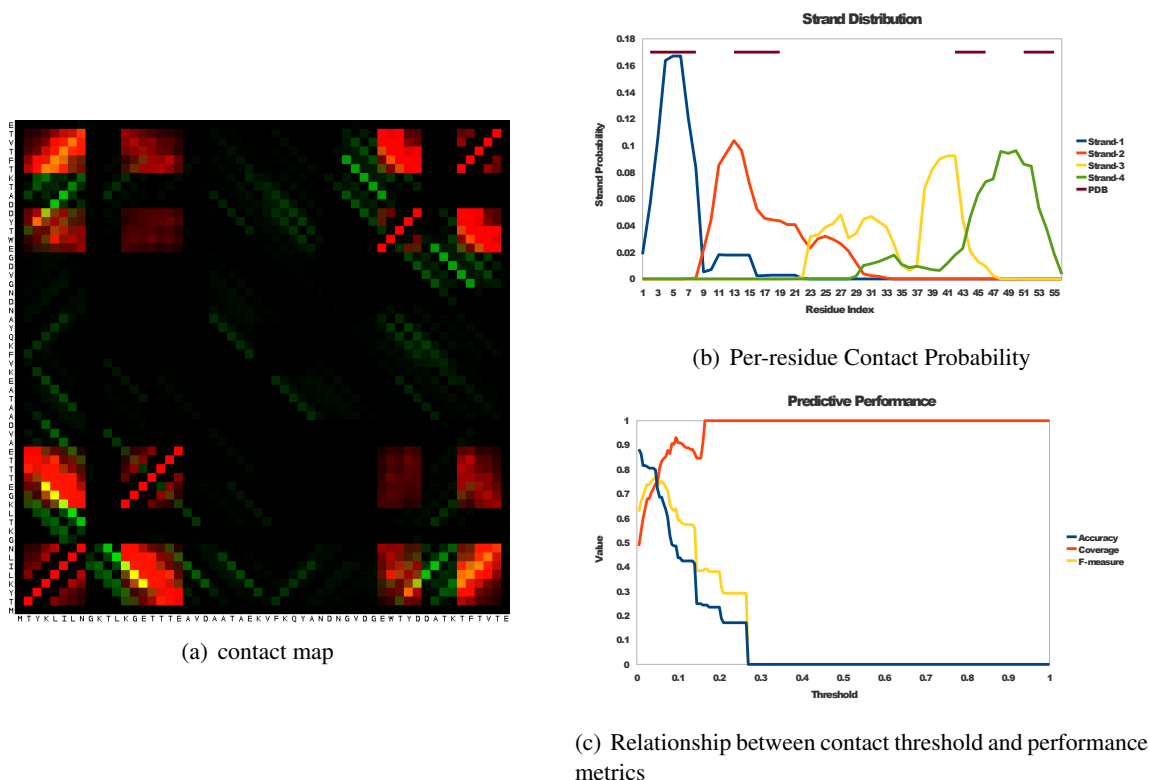


Figure 8-6: **Illustration of tFolder Protein G ensemble predictions:** Summary of the distribution of structures predicted by tFolder for Protein G (a) The stochastic contact map predicted by tFolder for all pairs of amino acids. Here green intensity indicate the likelihood of a contact as predicted by tFolder, while red points represent contact pairs observed in the experimental structure. Yellow point indicate agreement. (b) The predicted probability for the location of each numbered β -strand (using the pre-defined β -sheet schema topology). Black bars at the top indicate the location of β -strands in the experimentally determined structure. (c) The relationship between the choice in threshold t and the accuracy, coverage, and F-measure scores.

From each protein, the specific β -sheet schema topology (the number of β -strands and ordering of parallel or antiparallel interactions) was extracted and used as input for tFolder, along with the amino acid sequence and a fixed strand length of 5–8 residues. Since folding pathway predictions

permute over all β -sheet schema topologies, this demonstrates the expected accuracy of each folding state along the pathway. For each ensemble, 500 conformations were sampled, and a stochastic contact map and distribution of β -strand locations was computed (See Figure 8-6(a) and Figure 8-6(b) for the case of Protein G). As described in Section 3.1.1, we derive a set of predicted contacts by selecting a probability threshold value t that optimized the F-measure score (see Figure 8-6(c)). Similarly, our predictions are evaluated using accuracy ($\frac{\text{no. of correctly predicted contacts}}{\text{no. of predicted contacts}}$), coverage ($\frac{\text{no. of correctly predicted contacts}}{\text{no. of observed contacts}}$), and F-measure ($\frac{2 \cdot \text{accuracy} \cdot \text{coverage}}{\text{accuracy} + \text{coverage}}$).

A summary of the performance of tFolder on the set of 16 proteins is presented in Table 8.1 (along with the performance on Protein G). To highlight the ability of our approach to correctly predict long-range contacts, we distinguish between results for contacts with a sequential distance greater than 0, 12, or 24 residues apart. Table 8.2 provides a comparison of tFolder with BETApro and SVMcon, showing comparable results. In particular, although these other methods sometimes outperform our approach, tFolder appears less sensitive to the distance of contact separation. Since critical protein folding steps can involve both short-range and long-range β -sheet contacts, it is important to correctly predict both to allow the construction of accurate folding pathways.

	16 protein benchmark			Protein G		
	≥ 0	≥ 12	≥ 24	≥ 0	≥ 12	≥ 24
F-measure	0.25	0.27	0.23	0.36	0.37	0.45
Accuracy	0.25	0.27	0.28	0.34	0.33	0.41
Coverage	0.28	0.32	0.25	0.39	0.42	0.50

Table 8.1: **tFolder super-secondary structure contact prediction performance:** Shown are accuracy, coverage, and F-measure of experimentally observed contacts. Results are reported for contacts that are more than 0, 12, or 24 residues apart in sequence.

Method	≥ 12			≥ 24		
	F-measure	Accuracy	Coverage	F-measure	Accuracy	Coverage
tFolder	0.27	0.27	0.32	0.23	0.28	0.25
BETApro	0.22	0.40	0.16	0.05	0.14	0.07
SVMcon	0.32	0.31	0.50	0.24	0.21	0.44

Table 8.2: **Comparison of predictive performance of tFolder, BETApro, and SVMcon:** Results are reported for contacts that are more than 0, 12, or 24 residues apart in sequence.

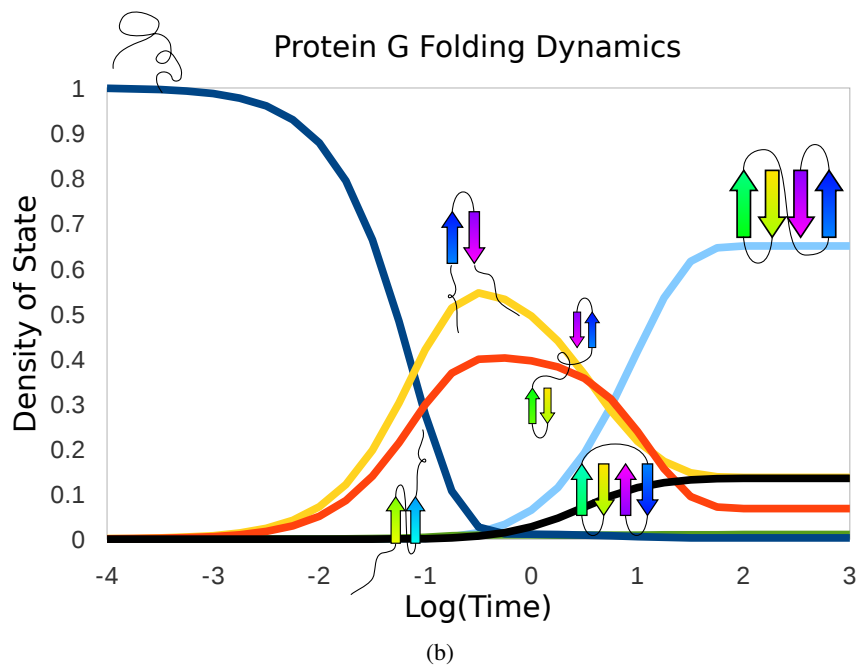
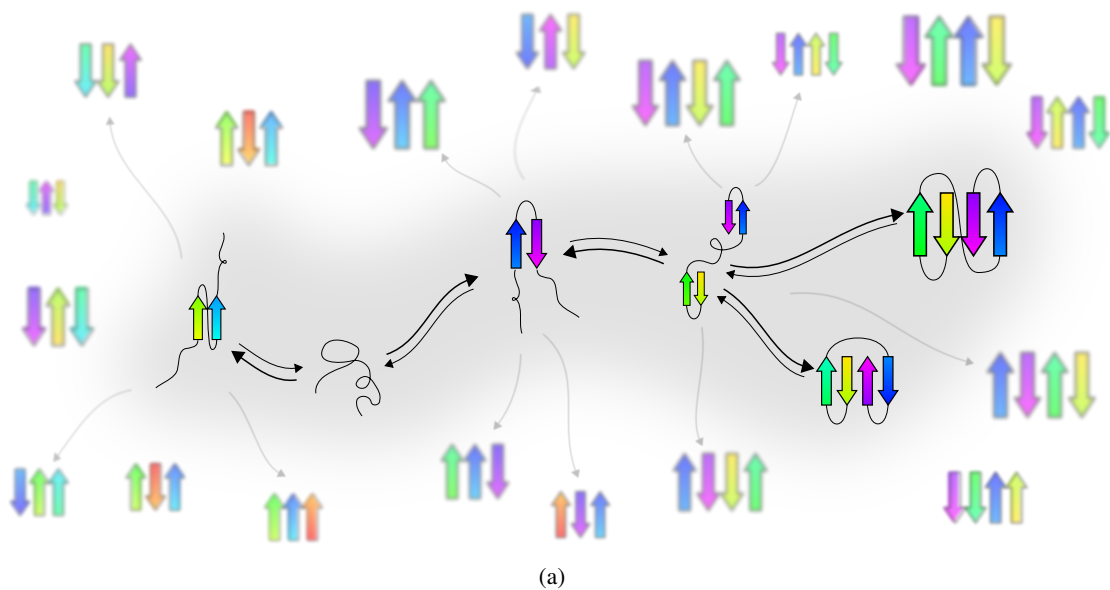


Figure 8-7: **Predicted folding pathways of B1 domain of Protein G:** (a) The gray shaded region indicates the states predicted to be reachable from the unfolded state. The dark arrows indicate transitions between states, and the size of the arrow indicates the favored direction of transition along each edge. Faded arrows are drawn between states that have compatible topologies but do not reach our selected transition threshold. The size of each state indicates its relative representation at equilibrium. Faded structures indicate states that are unreachable from the unfolded state. (b) The folding dynamics of Protein G shows how the probability of observing any of the reachable states changes over the time. Each line is annotated with an image of the state it represents in (a).

8.3.2 Case analysis of folding pathway prediction of the B1 domain of Protein G

To test the efficacy of our folding pathways prediction techniques, we reconstruct the folding landscape of the B1 domain of Protein G — a well-studied protein for which the pathway has been elucidated through many experimental studies and MD simulations. To do this, all possible schema topologies of a 4-strand β -sheet were sampled and clustered. For each of these sets of structures, the cluster with the highest probability of being observed was selected as representative of each topology.

The graph of the folding pathway was constructed by considering all pairs of clusters. If the minimum distance between two clusters was less than a fixed transition threshold, we considered there to be a potential exchange between two states. The resulting graph of protein conformations is illustrated in Figure 8-7(a). Inspection of this graph, along with the folding dynamics computed from this graph in Figure 8-7(b), reveals folding intermediates consistent with those previously reported by Song et al. [171]. It should also be noted that although we compute other configurations of the sequence that are energetically favorable (faded states in Figure 8-7(a)), they are not predicted to form because they are unreachable from the initial unfolded state. Interestingly, a four-stranded off-pathway structure is also less-favorably predicted to form, which has not been observed previously.

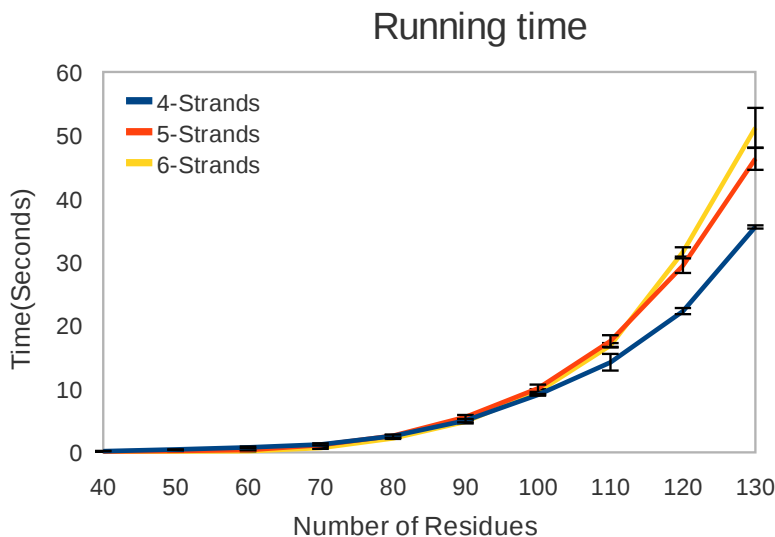


Figure 8-8: **tFolder running time performance:** The time required to compute the partition function increases with increasing size of amino acid sequence and number of strands. The time was computed by averaging over 3 trials for sequences ranging from 40–130 residues in length with 4–6 strands.

Furthermore, our results also agree with the work of Hubner et al., who show that the antiparallel beta-hairpin, predicted to form an interaction between residues 39–44 and 50–55, center around known nucleation points W43, Y50, F54 [90].

8.3.3 tFolder running time

For our tool to complement the use of MD simulations, predictions must be able to be computed quickly. The computational bottleneck of our algorithm is the computation of each β -sheet schema topology partition function, primarily influenced by sequence length. To evaluate the time needed for this step, we calculate the partition function for sequences between 40–130 residues and 4–6 strands were calculated using a single 2.66GHz processor with 512 MB of RAM. The effect of these two parameters on the computation time is depicted in Figure 8-8. Note, computing the partition functions of multiple β -sheet schemas is trivially parallelizable.

Chapter 9

Web-based ensemble prediction tools

An important goal in computational biology is the dissemination of novel algorithmic ideas so that other research groups can take advantage of such developments and build upon them. In the case of protein modeling algorithms and predictors this is often best achieved through the creation of publically-available web-based tools that offer simple to understand inputs and outputs. Accordingly, nearly all of the computational algorithms introduced in this thesis have been implemented as online tools that will be discussed in this chapter.

We present a summary of the four major tools that have been developed and made publically available based on the algorithms in this thesis: `partiFold`, `AmyloidMutants`, `partiFoldAlign`, and `tFolder`. In particular, we outline their goals and provide brief instruction in their use. We note that changes may be made in the future to these websites to add new features or alter ease of use.

9.1 `partiFold`

Our ensemble algorithms for predicting the structure of transmembrane β -barrel proteins have been implemented by the software `partiFold` and made available as a web-based tool at the URL <http://partiFold.csail.mit.edu>. In its simplest form the tool is designed to take as input an amino acid sequence and output a prediction of likely ensemble TMB structures, most easily visualized using a stochastic contact map. (Section 2.2.5). By default, the set of TMB ensemble parameters are fixed, however more advanced use allows these to be changed to best suit the current prediction problem. More detailed usage instructions are provided on the website.

Figure 9-1: **Screenshot of partiFold online tool:** The only input necessary is the amino acid sequence. Standard structural constraints are automatically selected, however these can be changed via the lower text boxes.

9.1.1 Input

Figure 9-1 provides a screenshot of the partiFold web front-end. From this initial webpage, TMB structural predictions can be produced by the following steps:

1. Enter the amino acid sequence using the FASTA format.

2. Choose different structural restraints (optional).
3. Choose different energetic parameters (optional).
4. Submit the request to the server.

In step 2, biologically motivated TMB structure constraints refine the specific kinds of conformations the predicted ensemble will take into account. These include the number of TM β -strands in the barrel, the length of TM β -strands, the shear number (Section 2.2.2), the size of periplasmic and extra-cellular loops, and the hydrophobic profile of TM β -strands.

Step 3 selects which energetic scoring function should be used (e.g., pairwise potentials or stacking pairs), along with other optional energetic bonuses such as the use of additional polarity scores to filter loops between β -strand pairs, or additional hydrophobicity scores help identify TM β -strands.

9.1.2 Output

Since ensemble predictions are designed to describe patterns across multiple sampled conformations, partiFold generates multiple forms of output representations to aid analysis. However, a textual representation of every sampled conformation is also available to download for further analysis.

The primary output presents a graphical representation of a stochastic contact map, similar to that which is seen in Figure 2-6(a). In addition, partiFold generates a per-residue 2-dimensional plot of the probability that any specific residue index is involved in a β -sheet interaction. Similarly, the per-residue contact entropy $E(i)$ of a residue at index i is computed as $-\log(P(i, j))$. The inclusion of 3-dimensional PDB files for clustered mediod structures is a subject of ongoing work. Finally, for completeness partiFold also displays a textual representation of the single minimum folding energy structure

9.2 AmyloidMutants

Our ensemble algorithms for predicting amyloid fibril structure and mutational landscapes have been implemented by the software AmyloidMutants and made available as a web-based tool at <http://amyloid.csail.mit.edu>. This tool serves two purposes, to predict amyloid fibril structure from sequence given no mutations, and to output the mutational landscape of a given sequence. As with partiFold, the goal is to output predicted results without the need to fix any specific parameters, while

at the same time allowing these ensemble and energetic parameters to be changed in an advanced mode. More detailed usage instructions are provided on the website.

AmyloidMutants

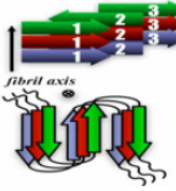
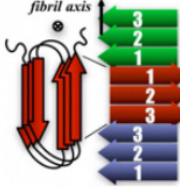
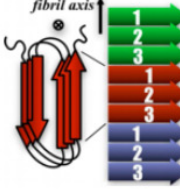
Input Rerun Clustering Instructions Examples References Contact

Basic Input Landscape Parameters Sampling Advanced Mutations

Sequence:
Please paste your sequence:

or [load a sample sequence](#)

Structural Schema:
Please choose how you want to fold your sequence.

Cross-beta-pleat (serpentine) Beta-solenoidal Anti-parallel inter-peptide interface Beta-solenoidal Parallel inter-peptide interface

Mutational Schema:
Please choose a residue index and select possible mutations (max 2). The *Advanced Mutations* tab allows for more mutant entries.

position	residue	or	residue
• <input type="text"/>	→ <input type="text" value="NA"/>	or	<input type="text" value="NA"/>
• <input type="text"/>	→ <input type="text" value="NA"/>	or	<input type="text" value="NA"/>
• <input type="text"/>	→ <input type="text" value="NA"/>	or	<input type="text" value="NA"/>

or [load a sample mutation](#)

Email (opt):

Click to Submit: Reset Defaults:

Figure 9-2: **Screenshot of AmyloidMutants online tool:** The initial screen allows the input of the basic features necessary to generate amyloid fibril predictions. More advanced options can be input through alternate tabs.

9.2.1 Input

Figure 9-2 provides a screenshot of the AmyloidMutants web front-end. From this initial webpage, amyloid fibril sequence/structural predictions can be generated using the following steps:

1. Enter the amino acid sequence.

2. Choose a schema.
3. Enter a set of potential mutations (optional).
4. Enter an email address (optional).
5. Submit the request to the server.

Schema selection in step 2 is mandatory as it defines the type of amyloid ensemble: schema \mathcal{S} , \mathcal{A} , or \mathcal{P} . Mutation protocol selection in step 3 is optional and serves as a simplified interface for specifying mutations. The *Advanced Mutations* tab allows the insertion of arbitrary mutation definitions, as described in Section 4.2.1, such as “13 \rightarrow APZ,” meaning index 13 can be *Ala*, *Pro*, or its original WT value (“Z”), “V \rightarrow VP,” meaning all *Val* in the sequence can be *Val* or *Pro*, or “14=A 15=V, 14=T 15=P” meaning either position 14 is *Ala* and 15 is *Val* or position 14 is *Thr* and 15 is *Pro*. Prediction results are both displayed on the screen and emailed to an email address if entered.

The tabs *Landscape Parameters* and *Sampling* allow more advance selection of ensemble parameters. For example, the *Landscape Parameters* tab contains structural parameters such as minimum and maximum β -strand lengths, N-/C-terminus and inter-strand coil lengths, whether to enable “kinks,” and β -sheet “slip” (Section 2.3.2). Further, the choice of energetic scoring function can be selected here, as well as an optional use of thresholding (Section 2.3.7). The *Sampling* tab contains options that effect sampling and the post-process clustering ensemble sequence/structure states. This includes the number of samples, whether to use unique sampling, the number of clusters, the type of clustering distance metric to use, and an optional choice of random seed. Since clustering is a post-processing step, *AmyloidMutants* also allows generated samples to be reclustered with different parameters via the *Rerun Clustering* option seen at the top of Figure 9-2.

9.2.2 Output

Again, since ensemble predictions typically describe conformational population patters, multiple forms of output are visualized by *AmyloidMutants* to aid in data interpretation. A textual representation of every sampled conformation is also made available, which can be reclustered using the *Rerun Clustering* feature.

The primary output of *AmyloidMutants* depicts the per-residue β -sheet propensity of each cluster in single graph indicating the Boltzmann weight of each cluster (as is seen in Figure 3-8(b)).

Supporting figures display a stochastic contact map for each cluster, and the per-sequence-index mutational frequency of each cluster. Additionally, the rank-ordered pseudo-energetic scores of each cluster are plotted, offering a simplistic illustration of predicted ensemble energy wells.

9.3 partiFoldAlign

Our ensemble algorithms for performing simultaneous alignment and folding of transmembrane β -barrel proteins has been implemented by the software `partiFoldAlign` and made available as a web-based tool at <http://partiFold.csail.mit.edu>. At present, only a preliminary interface has been constructed which accepts two amino acid sequences and produces the same types of results as described for `partiFold`. However, unlike `partiFold`, `partiFoldAlign` stochastic contact maps depict the contact probability of both sequences at once (displayed in either the lower-left or upper-right triangles), as is seen in Figure 7-1. The integration of `partiFold` and `partiFoldAlign` into a single web interface is a subject of ongoing work.

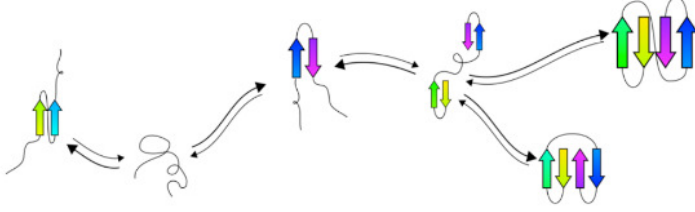
9.4 tFolder

With our collaborators, our ensemble algorithms for predicting the folding pathways of single β -sheet proteins have also been implemented by the software `tFolder` and made available as a web-based tool at <http://csb.cs.mcgill.ca/tFolder/>. The goal of this tool is to demonstrate the ability to predicted folding pathways for relatively simple single- β -sheet proteins provided only sequence information. As a result, at present relatively few structural parameters can be changed, and only β -sheets with as many as 6 β -strands are calculated.

9.4.1 Input

Figure 9-3 provides a screenshot of the `tFolder` web front-end. In its simplest form, folding pathway predictions can be generated by entering an amino acid sequence and submitting the request to the server. However, the webserver also allows the specification of four structural parameters that alter the size and makeup of the ensemble: the maximum number of parallel or antiparallel β -strand in the β -sheet, the minimum length of loops connecting two β -sheets, and the minimum and maximum permitted β -strand lengths. Prediction results are then displayed on the screen and emailed to an email address if provided.

tFolder is a program that enables you to compute coarse grained representations of the energy landscape of β -sheet proteins and to predict their folding pathways. All you need is to enter your sequence and select the maximal number of strands allowed.



Sequence:

Maximal number of Strands: Minimum loop length:

Minimum strand length: Maximum strand length:

Email:

Figure 9-3: **Screenshot of tFolder online tool:** The input screen accepts an amino acid sequence and basic changes to the size and makeup of the ensemble via four structural parameters.

9.4.2 Output

The primary output of tFolder is a plot indicating the probability of observing reachable states over simulation time, as is shown in Figure 8-7(b). Similarly, a graph of the pathway connectivity of states is also automatically generated, similar to the stylized graph seen in Figure 8-7(a).

Chapter 10

Conclusion

In this chapter we summarize the major contributions of this thesis and address its broader impact. We then describe algorithmic improvements that may be applied to our techniques in the future, as well as other biological problems that may benefit from an ensemble modeling approach.

10.1 Summary

This thesis introduces new algorithms for efficiently modeling β -sheet protein structures and their sequence variants. Our structure prediction algorithms compute the Boltzmann-distributed energy of *all* conformational states within a particular protein family, and output sampled high-likelihood structures. This statistical mechanics-based “*ensemble*” approach is further employed to describe sequence/structure variation, enabling sequence and energy-driven mutational and comparative analyses. Additionally, we demonstrate how ensemble predictions can be incorporated into algorithms approximating kinetic protein folding pathways. These algorithms have advanced the state-of-the-art in structure prediction, mutational analysis, and sequence alignment for two enigmatic β -sheet protein families: transmembrane β -barrels and amyloid fibrils. Further, we have used these methods to help guide experiments revealing structural characteristics of amyloid fibrils and to design efficient therapeutics for bacterial biofilm inhibition.

At a high-level, this thesis demonstrates two broad points. First, that coarse protein structure representations can be used to characterize complete ensembles of potential states, enabling accurate predictions and new forms of analysis. Moreover, this trade-off between representation detail and computation tractability has broad applicability to bioinformatics problems in general. Second, that constructs such as schemas can allow a model to be easily updated as new experimental

data becomes available. Such an iterative approach is essential for the successful integration of computational predictions throughout a biological investigation.

Note, one of the key factors enabling our efficient calculations is that predictions are strictly conditioned on the assumption that the protein family or other schema information is known. In many biological investigations this may be true, however, when no data exists the only recourse is to hypothesize that such information is true and use the resulting computational predictions to guide experimental validation. This can present both benefits (highly specific experiments can be designed to quickly validate predictions) and disadvantages (assays may be done wastefully if the underlying assumption is incorrect). As a result, it may be useful to integrate non-parametric, data-driven techniques with our methods when studying datasets for which no prior information is known (e.g., genomic-level studies).

Finally, while the study of TMBs and amyloid fibrils has been the subject of this thesis, fundamentally, any protein, RNA, or other molecular representation can be treated using an ensemble approach. Statistical mechanics-based predictions can be calculated in the same way as long as the representation is sufficiently coarse (e.g., non-3-dimensional) and an energetic model is used that assumes independence between substates. In particular, nearly all of the algorithms described could be directly applied to other all- β -sheet protein families through the design of new recursive schema encodings and the use of similar statistical potentials-based energetics (for example, the various β -sandwich folds found in SCOP [130] and CATH [140]). Further, while we demonstrate ensemble methods for structure prediction, mutant analysis, sequence alignment, and folding pathway prediction, many other bioinformatics problems could be well served by such an approach. For example, signaling and regulatory network analysis, protein/protein interactions, and DNA transcription factor occupancy could all be characterized via a Boltzmann distribution of states.

10.2 Future research

The ensemble methods described in this thesis provide accurate, useful tools for the study of TMBs and amyloid fibrils, and have been used in ongoing open biological investigations. However, there are still computational research opportunities to enhance the accuracy of the predictions and the descriptiveness of the models. Further, many other interesting biological systems could be studied with straightforward modifications to the algorithms. Here we describe future directions for improving our energy model and schema design, as well as the potential for studying α/β proteins and

autotransporters.

10.2.1 Energy model improvements

We have shown that the statistical potentials-based energy model described in Section 2.4 can enable sensitive and specific predictions for TMBs and amyloid fibrils. However, the use of these statistical potentials has necessary drawbacks: pairwise and stacking pair residue frequencies only incorporate the effects of a relatively small number of local interactions, and further, the precise physical interpretation of database-derived potentials is unclear.

Improved predictive accuracies may be possible through the use of statistical potentials incorporating complex, non-pairwise interactions (e.g., long aromatic stacking chains or residue sidechain rotamer consideration). Moreover, it may be possible to integrate energetic scores from non-uniform divisions of substructure. Two key questions must be addressed to accomplish this: the identification of which interactions are most influential to protein stability, and a method to prevent undersampling if these complex interactions are infrequent in the PDB.

Alternately, as has been achieved in the case of RNA [213], improved thermodynamic parameters could be added from experimental study (e.g., chemical modification or site directed mutagenesis). Similarly, the energetic potential of water or the fluctuations of coil regions could be integrated into the model through the use of independent atomistic force field calculations [108]. Properly integrating multiple data sources into a coherent physical or statistical model presents one of the larger challenges in these approaches.

10.2.2 Constraint-based schema design

In the present ensemble algorithms, protein conformational space is encoded using recursive techniques. This allows for an efficient computation of the partition function using dynamic programming while still providing flexibility in the type of structural configurations that can be described. However, some state spaces, such as those combining multiple unrelated structures, can be difficult to efficiently implement. An interesting solution may be to describe protein structure schemas using constraint-based techniques such as SAT [16], or more effectively, using satisfiability modulo theories (SMTs) [16]. Indeed, the efficiency of modern SAT solvers has been shown quite useful for other bioinformatics problems such as protein design [134]. Such an approach offers significant descriptive freedom, although the specific mapping of 2-dimensional or 3-dimensional structure to

logical clauses (such as for SAT) or richer constraints (such as bitvector arithmetic for SMTs [77]) must be done carefully, so as not to introduce a large number of variables.

10.2.3 α/β protein schemas

In this thesis, we address modeling techniques for mainly- β -sheet proteins such as TMBs and amyloid fibrils. This focus was primarily chosen because experimental characterization of such proteins has proven extremely difficult, suggesting a need for accurate computational methods that can leverage their semi-regular β -sheet structure. However, a far greater number of proteins contain significant amounts of both α -helices and β -sheets, and even TMBs and amyloid fibrils can contain α -helices within β -strand loop connection regions. At present, if one wishes to include α -helices in TMB or amyloid fibril models, we use a standard secondary structure predictor [145] as a preprocessor, and specifically exclude highly-likely α -helix regions from forming β -sheet (e.g., Section 2.3.2). The inclusion of sequentially local α -helix residues interactions, and sequentially distant α/α and α/β -sheet interactions within our schemas would provide a significant improvement to our model. Crucial to this would be the inclusion of α -helical substructure states. Unfortunately, while the inclusion of sequentially local α -helical interactions may be straightforward, sequentially distant α/α interactions may be problematic due to their irregularity. Further, α -helix/ β -sheet-face interfaces offer further complications to recursive state definitions and the construction of a joint α -helix/ β -sheet interaction energetic model.

10.2.4 Autotransporters

Finally, one particularly interesting area of future research would be the application of our ensemble algorithms to autotransporter proteins. This diverse class of proteins from gram-negative bacteria contain a β -barrel porin domain that typically embeds itself within the outer membrane and facilitates the transport of a passenger peptide to extracellular space (in the absence of ATP) [50]. The passenger domain is often solenoidal and can act as a virulence factor, adhesin, or degradative enzyme. However, the exact mechanisms for peptide translocation are not entirely clear.

The integration of a β -barrel and β -solenoid within a single structural ensemble seems a direct extension of the present algorithms for TMBs and amyloid fibrils. This could be used to identify critical features of the secretion process, such as barrel flexibility or channel/passenger compatibility, furthering our understanding of the secretory pathway. Mutational landscape analysis may also help identify new strategies for the inhibition or activation of secretion.

Bibliography

- [1] A. Aguzzi, M. Heikenwalder, and M. Polymenidou. Insights into prion strains and neurotoxicity. *Nat. Rev. Mol. Cell Biol.*, 8:552–561, 2007.
- [2] S. Alberti, R. Halfmann, O. King, A. Kapila, and S. Lindquist. A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. *Cell*, 147:146–158, 2009.
- [3] N.M. Amato and G. Song. Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9(2):149–68, 2002.
- [4] I. Andre, P. Bradley, C. Wang, and D. Baker. Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci.*, 104:17656–17661, 2007.
- [5] I. André, C.E.M. Strauss, D.B. Kaplan, P. Bradley, and D. Baker. Emergence of symmetry in homooligomeric biological assemblies. *Proc. Natl. Acad. Sci.*, 105(42):16148–16152, 2008.
- [6] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.
- [7] M. Asogawa. Beta-sheet prediction using inter-strand residue pairs and refinement with hopfield neural network. *Bioinformatics, Proc. of ISMB 1997*, 5:48–51, 1997.
- [8] A.W. Bryan Jr. and M. Menke and L.J. Cowen and S. Lindquist and B. Berger. BETASCAN: Probable β -amyloids Identified by Pairwise Probabilistic Analysis. *PLoS Comput. Biol.*, 5:e1000333, 2009.
- [9] R. Backofen and W. Sebastian. Local sequence-structure motifs in RNA. *J. Bioinform. Comput. Biol.*, 2(4):681–698, 2004 Dec.
- [10] M. M. Barnhart and M.R. Chapman. Curli biogenesis and function. *Annu. Rev. Microbiol.*, 60:131–147, 2006.
- [11] A.I. Bartlett and S.E. Radford. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat. Struct. Mol. Biol.*, 16:582–588, 2009.
- [12] V. Batagelj. The quadratic hash method when the table size is not a prime number. *Comm. of the ACM*, 18, 1975.
- [13] O.G. Berg and P.H. Hippel. Selection of dna binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193(4):723–750, 1987.
- [14] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.
- [15] K. Berthelot, F. Immel, J. Géan, S. Lecomte, R. Oda, B. Kauffmann, and C. Cullin. Driving amyloid toxicity in a yeast model by structural changes: a molecular approach. *FASEB J.*, 23:2254–2263, 2009.

- [16] A. Biere, M.J.H. Heule, H. van Maaren, and T. Walsh, editors. *Handbook of Satisfiability*, volume 185. IOS Press, February 2009.
- [17] H. Bigelow, D. Petrey, J. Liu, D. Przybylski, and B. Rost. Predicting transmembrane beta-barrels for entire proteomes. *Nucleic Acids Res.*, 32(8):2566–2577, 2004.
- [18] R. Blossey. *Computational Biology: A Statistical Mechanics Perspective*. Chapman & Hall/CRC, 2006.
- [19] F.C. Botelho, R. Pagh, and N. Ziviani. Simple and space-efficient minimal perfect hash functions. *Lect. Notes. Comput. Sci., Proc. of WADS 2007*, 4619, 2007.
- [20] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. BETAWRAP: Successful prediction of parallel Beta-helices from primary sequence reveals an association with many microbial pathogens. *Proc. Natl. Acad. Sci.*, 98(26):14819–24, 2001.
- [21] P. Bradley, K. M. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309:1868–1871, 2005.
- [22] C. Branden and J. Tooze. *Introduction to Protein Structure, Second Edition*. Garland Publishing, Inc., 1999.
- [23] M. Brudno, C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, and S. Batzoglu. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Res.*, 13(4):721–731, 2003 Apr.
- [24] A.A. Canutescu, A.A. Shelenkov, and R.L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, 12:2001–2014, 2003.
- [25] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz. 1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap. *J. Comput. Biol.*, 11(1):27–52, 2004.
- [26] Critical Critical Assessment of Techniques for Protein Structure Prediction. <http://predictioncenter.org/casp7/>.
- [27] A.W. Chan, E.G. Hutchinson, D. Harris, and J.M. Thornton. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.*, 2(10):1574–1590, 1993.
- [28] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987.
- [29] P. Chasignat, J. Waldspühl, and J.-M. Steyaert. personal communication.
- [30] B. Chazelle, J. Kilian, R. Rubinfeld, and A. Tal. The bloomier filter: An efficient data structure for static support lookup tables. *Proc. of SODA 2004*, 2004.
- [31] J. Cheng and P. Baldi. Three-stage prediction of protein β -sheets by neural networks, alignments and graph algorithms. *Bioinformatics, Proc. of ISMB 2005*, 21(suppl. 1):i75–i84, 2005.
- [32] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinf.*, 8:113–121, 2007.
- [33] R.V. Chereji, D. Tolkunov, G. Locke, and A.V. Morozov. Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. *Phys. Rev. E*, 83(5):050903, 2011.
- [34] D. Chiang, A. K. Joshi, and K.A. Dill. A grammatical theory for the conformational changes of simple helix bundles. *J. Comput. Biol.*, 13(1):27–42, 2006.

- [35] D. Chiang, A.K. Joshi, and D.B. Searls. Grammatical representations of macromolecular structure. *J. Comput. Biol.*, 13(5):1077–100, Jun 2006.
- [36] F. Chiti and C.M. Dobson. Protein Misfolding, Functional Amyloid, and Human Disease. *Annu. Rev. Biochem.*, 75:333–366, 2006.
- [37] C. Chotia. The Nature of the Accessible and Buried Surfaces in Proteins. *J. Mol. Biol.*, 105(1):1–14, 1975.
- [38] P.Y. Chou and G.D. Fasman. Prediction of protein conformation. *Biochemistry*, 13:222–245, 1974.
- [39] M. Cline, R. Hughey, and K. Karplus. Predicting reliable regions in protein sequence alignments. *Bioinformatics*, 18(2):306–314, 2002 Feb.
- [40] P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, 2000.
- [41] P. Clote, J. Waldispühl, B. Behzadi, and J.-M. Steyaert. Energy landscape of k-point mutants of an RNA molecule. *Bioinformatics*, 21(22):4140–4147, 2005.
- [42] E.J. Cohn and J.T. Edsall, editors. *Proteins, Amino Acids and Peptides*. Reinhold Publishing, 1943.
- [43] The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, 39:D214–D219, 2011.
- [44] M.R. Cookson. The Biochemistry of Parkinson’s Disease. *Annu. Rev. Biochem.*, 74:29–52, 2005.
- [45] A. Cooper. Thermodynamic fluctuations in protein molecules. *Proc. Natl. Acad. Sci.*, 74(8):2740–2741, 1976.
- [46] V. Coustou, C. Deleu, S. Saupe, and J. Begueret. The protein product of the het-s heterokaryon incompatibility gene of the fungus *podospira anserina* behaves as a prion analog. *Proc. Natl. Acad. Sci.*, 94(18):9773–9778, 1997.
- [47] V. Coustou, C. Deleu, S.J. Saupe, and J. Bégueret. Mutational Analysis of the [HET-s] Prion Analog of *Podospira anserina*: a Short N-Terminal Peptide Allows Prion Propagation. *Genetics*, 153:1629–1640, 1999.
- [48] J. Couthouis, K. Rébora, F. Immel, K. Berthelot, M. Castroviejo, and C. Cullin. Screening for Toxic Amyloid in Yeast Exemplifies the Role of Alternative Pathway Responsible for Cytotoxicity. *PLoS ONE*, 4(3):e4539, 2009.
- [49] L. Cowen, P. Bradley, M. Menke, J. King, and B. Berger. Predicting the beta-helix fold from protein sequence data. *J. Comput. Biol.*, pages 261–276, 2001.
- [50] N. Dautin and H.D. Bernstein. Protein Secretion in Gram-Negative Bacteria via the Autotransporter Pathway. *Annu. Rev. Microbiol.*, 61:89–112, 2007.
- [51] M. Dietzfelbinger, A. Karlin, K. Mehlhorn, F. Meyer auf der Heide, H. Rohnert, and R. Tarjan. Dynamic perfect hashing: Upper and lower bounds. *SIAM J. Comput.*, 23, 1994.
- [52] K.A. Dill and S. Bromberg. *Molecular Driving Forces*. Garland Science, Taylor & Francis, 2003. New York.
- [53] Y. Ding and C.E. Lawrence. A bayesian statistical algorithm for RNA secondary structure prediction. *Comput. Chem.*, 23(3-4):387–400, Jun 1999.
- [54] Y. Ding and C.E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31(24):7280–7301, 2003.
- [55] C.B. Do, C-S. Foo, and S. Batzoglou. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, 24:i68–i76, 2008.

- [56] C.B. Do, D.A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–8, 2006.
- [57] C.M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.
- [58] D.A. Drummond and C.O. Wilke. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.*, 10:715–724, 2009.
- [59] H.J. Dyson and P.E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, 2005.
- [60] R.C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.
- [61] R.C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf.*, 5:113, 2004 Aug 19.
- [62] R.C. Edgar and S. Batzoglou. Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, 16(3):368–373, 2006 Jun.
- [63] R. Elber and M. Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235(4786):318–321, 1987.
- [64] U. Erlingsson, M. Manasse, and F. McSherry. A cool and practical alternative to traditional hash tables. *Proc. of WDAS 2006*, 2006.
- [65] P. Faccioli, M. Sega, F. Pederiva, and H. Orland. Dominant pathways in protein folding. *Phys. Rev. Lett.*, 97(10):108101, Sep 2006.
- [66] B. Fain and M. Levitt. A novel method for sampling alpha-helical protein backbones. *J. Mol. Biol.*, 305(2):191–201, 2001.
- [67] B. Fain and M. Levitt. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proc. Natl. Acad. Sci. USA*, 100(19):10700–5, 2003.
- [68] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins Struct. Funct. Bioinf.*, Suppl 5:157–162, 2001.
- [69] A.M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol. e-pub*, 22(10):1302–1306, 2004.
- [70] A. Fiser and A. Sali. MODELLER: generation and refinement of homology-based protein structure models. *Methods Enzymol.*, 374:461–491, 2003.
- [71] B.C. Foat, A.V. Morozov, and H.J. Bussemaker. Statistical mechanical modelign of genmoe-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–149, 2006.
- [72] L.R. Forrest, C.L. Tang, and B. Honig. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J.*, 91(2):508–517, 2006 Jul 15.
- [73] D. M. Fowler and J. W. Kelly. Aggregating knowledge about prions and amyloid. *Cell*, 137(1):146–158, 2009.
- [74] H. Frauenfelder, S.G. Sligar, and P.G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

- [75] M.L. Fredman, J. Komolós, and E. Szemerédi. Storing a sparse table with $o(1)$ worst case access time. *J. of the ACM*, 31:538–544, 1994.
- [76] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinf.*, 23:566–579, 1995.
- [77] V. Ganesh and D.L. Dill. A decision procedure for bit-vectors and arrays. In *Proc. of CAV 2007*, pages 519–531, 2007.
- [78] O. Grana, D. Baker, R.M. MacCallum, J. Meiler, M. Punta, B. Rost, M.L. Tress, and A. Valencia. Casp6 assessment of contact prediction. *Proteins Struct. Funct. Bioinf.*, 61(7):214–224, 2005.
- [79] M. Gromiha and M. Suwa. A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, 27(7):961–968, 2005.
- [80] H. Heise and W. Hoyer and S. Becker and O.C. Andronesi and D. Riedel and M. Baldus. Molecular-level secondary structure, polymorphism, and dynamics of full-length α -synuclein fibrils studied by solid-state NMR. *Proc. Natl. Acad. Sci.*, 102(44):15871–15876, 2005.
- [81] N.D. Hammer, J.C. Schmidt, and M.R. Chapman. The curli nucleator protein, csgB, contains an amyloidogenic domain that directs csgA polymerization. *Proc. Natl. Acad. Sci.*, 104:12494–12499, 2007.
- [82] J.H. Havgaard, E. Torarinsson, and J. Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, 3(10):1896–1908, 2007 Oct.
- [83] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *PNAS*, 89:10915–10919, 1992.
- [84] M. Herlihy, N. Shavit, and M. Tzafrir. Hopscotch hashing. In *DISC'08*, pages 350–364, 2008.
- [85] T. Hirokawa, S. Boon-Chieng, and S. Mitaku. Sosisu: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–379, 1998.
- [86] I.L. Hofacker, S.H.F. Bernhart, and P.F. Stadler. Alignment of RNA Base Pairing Probability Matrices. *Bioinformatics*, 20:2222–2227, 2004.
- [87] B. Honig. Protein folding: from the Levinthal paradox to structure prediction. *J. Mol. Biol.*, 293:283–293, 1999.
- [88] R. Hosur, R. Singh, and B. Berger. Sparse estimation for structural variability. *Algorithms Mol. Biol.*, 2011.
- [89] T.J.P. Hubbard. Use of beta-strand Interaction Pseudo-Potentials in Protein Structure Prediction and Modeling. In *Proceedings of the Biotechnology Computing Track, Protein Structure Prediction MiniTrack of the 27th HICSS*, pages 336–344, 1994.
- [90] I.A. Hubner, J. Shimada, and E.I. Shakhnovich. Commitment and nucleation in the protein G transition state. *J. Mol. Biol.*, 336:745–761, 2004.
- [91] P.M. Hwang, W.-Y. Choy, E.I. Lo, L. Chen, J.D. Forman-Kay, C.R.H. Raetz, G.G. Privé, R.E. Bishop, and L.E. Kay. Solution structure and dynamics of the outer membrane enzyme PagP by NMR. *Proc. Natl. Acad. Sci.*, 99(21):13560–13565, 2002.
- [92] G.B. Irvine, O.M. El-Agnaf, G.M. Shankar, and D.M. Walsh. Protein Aggregation in the Brain: The Molecular Basis for Alzheimer's and Parkinson's Diseases. *Mol. Med.*, 14(7-8):451–464, 2008.

- [93] S. Istrail. Statistical mechanics, three-dimensionality and NP-completeness: I. Universality of intractability of the partition functions of the ising model across non-planar lattices. In *Proc. of STOC 2000*, pages 87–96, 2000.
- [94] S. Istrail and F. Lam. Combinatorial Algorithms for Protein Folding in Lattice Models: A Survey of Mathematical Results. *Communications in Information and Systems*, 9(4):303–346, 2009.
- [95] R.L. Dunbrack Jr. Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.*, 16(3):374–384, 2006 Jun.
- [96] A.V. Kajava, U. Aepli, and A.C. Steven. The Parallel Superpleated Beta-structure as a Model for Amyloid Fibrils of Human Amylin. *J. Mol. Biol.*, 348:247–252, 2005.
- [97] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9(9):646–52, Sep 2002.
- [98] Y. Kato, T. Akutsu, and H. Seki. Dynamic programming algorithms and grammatical modeling for protein beta-sheet prediction. *J. Comput. Biol.*, 16(7):945–57, Jul 2009.
- [99] A. Kernysky and B. Rost. Static benchmarking of membrane helix prediction. *Nucleic Acids Res.*, 31(13):3642–3644, 2003.
- [100] W. Kim and M.H. Hecht. Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer’s A β 42 peptide. *Proc. Natl. Acad. Sci.*, 103(43):15824–15829, 2006.
- [101] W. Kim and M.H. Hecht. Mutations Enhance the Aggregation Propensity of the Alzheimer’s A β Peptide. *J. Mol. Biol.*, 377:565–574, 2008.
- [102] M. Kimura. Diffusion models in population genetics. *J. Appl. Prob.*, 1:177–232, 1964.
- [103] R. Koebnik. Membrane assembly of the *Escherichia Coli* outer membrane protein ompa: Exploring sequence constraints on transmembrane β -strands. *J. Mol. Biol.*, 285:1801–1810, 1999.
- [104] W.A. Koppensteiner and M.J. Sippl. Knowledge-based potentials – back to the roots. *Biochemistry*, 63:247–252, 1998.
- [105] R. Krishnan, T.K. Lu, and J. Goodman. Personal communication.
- [106] A. Krogh, G. von Heijne, and E.L.L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, 305:567–580, 2001.
- [107] J. Kyte and R. F. Doolittle. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, 157:105, 1982.
- [108] A. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001.
- [109] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–8, Feb 1975.
- [110] J. Liu, C. Lillo, P.A. Jonsson, C. Vande Velde, C.M. Ward, T.M. Miller, J.R. Subramaniam, J.D. Rothstein, S. Marklund, P.M. Andersen, T. Brännström, O. Gredal, P.C. Wong, D.S. Williams, and D.W. Cleveland DW. Toxicity of Familial ALS-Linked SOD1 Mutants from Selective Recruitment to Spinal Mitochondria. *Neuron*, 43(1):5–17, 2004.
- [111] Q. Liu, Y.-S. Zhu, B.-H. Wang, and Y.-X. Li. A HMM-based method to predict the transmembrane regions of β -barrel membrane proteins. *Comput. Biol. Chem.*, 27:69–76, February 2003.

- [112] M.A. Lomize, A.L. Lomize, I.D. Pogozheva, and H.I. Mosberg. OPM: Orientations of Proteins in Membranes database. *Bioinformatics*, 22:623–625, 2006.
- [113] T.K. Lu, S. Lindquist, R. Krishnan, J. Collins, C.W. O’Donnell, B. Berger-Leighton, and S. Devadas. Bacteriophages Expressing Amyloid Peptides and Uses Thereof. International Patent No. WO/2011/014693, Feb. 3, 2011.
- [114] S. Luca, W.M. Yau, R. Leapman, and R. Tycko. Peptide Conformation and Supramolecular Organization in Amylin Fibrils: Constraints from Solid State NMR. *Biochemistry*, 46(47):13505–13522, 2007.
- [115] T. Lührs, C. Ritter, M. Adrien, D. Riek-Loher, B. Bohrmann, H. Döbeli, D. Schubert, and R. Riek. 3D structure of Alzheimer’s amyloid- β (1-42) fibrils. *Proc. Natl. Acad. Sci.*, 102(48):17342–17347, 2005.
- [116] S.K. Majil., M.H. Perrin, M.R. Sawaya, S. Jessberger and dK. Vadodaria, R.A. Rissman, P.S. Singru, K.P. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek. Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules. *Science*, 325(5938):328–332, 2009.
- [117] H. Mamitsuka and N. Abe. Predicting location and structure of beta-sheet regions using stochastic tree grammars. *Proc. of ISMB 1994*, pages 276–284, 1994.
- [118] P.L. Martelli, P. Fariselli, A. Krogh, and R. Casadio. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. In *Bioinformatics, Proc. of ISMB 2002*, volume 18, pages S46–S53, 2002.
- [119] D.H. Mathews and D.H. Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, 317(2):191–203, 2002 Mar.
- [120] S. Maurer-Stroh, M. Debulpaep, N. Keummerer, M. Lopez de la Paz, I.C. Martins, J. Reumers, K.L. Morris, A. Copland, L. Serpell, L. Serrano, J.W. Schymkowitz, and F. Rousseau. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, 7(3):237–242, 2010.
- [121] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
- [122] M. Menke, B. Berger, and L. Cowen. Matt: local flexibility aids protein multiple structure alignment. *PLoS Comp Bio*, 4(1):e10, 2008 Jan.
- [123] L. Mirny and E. Shakhnovich. Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.*, 30:361–96, 2001.
- [124] S. Miyazawa and R.L. Jernigan. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, 18:534–552, 1985.
- [125] B. Morel, S. Casares, and F. Conejero-Lara. A Single Mutation Induces Amyloid Aggregation in the α -Spectrin SH3 Domain: Analysis of the Early Stages of Fibril Formation. *J. Mol. Biol.*, 356:453–468, 2006.
- [126] A.V. Morozov, J.J. Havranek, D. Baker, and E.D. Siggia. Protein-dna binding specificity predictions with structural models. *Nucleic Acids Res.*, 33(18):5781–5798, 2005.
- [127] V. Moulton, M. Zuker, M. Steel, R. Pointon, and D. Penny. Metrics on RNA secondary structures. *J. Comput. Biol.*, 7:277–292, 2000.

- [128] D.W. Mount. *Bioinformatics: Sequence and Genome Analysis, Second Edition*. Cold Spring Harbor Laboratory Press, 2004.
- [129] M.D. Mukrasch, S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow, and M. Zweckstetter. Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. *PLoS Biol.*, 7(2):e1000034, 2009.
- [130] A.G. Murzin, S.E. Brenner, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.
- [131] K.N. Navjyot, K. Harpreet, and G.P.S. Raghava. Prediction of transmembrane regions of β -barrel proteins using ANN- and SVM-based methods. *Proteins Struct. Funct. Bioinf.*, 56:11–18, 2004.
- [132] R. Nelson, M.R. Sawaya, M. Balbirnie, A.O. Madsen, C. Riek, R. Grothe, and D. Eisenberg. Structure of the cross-beta spine of amyloid-like fibrils. *Nature*, 435:773–778, 2005.
- [133] C.W. O'Donnell, J. Waldispühl, M. Lis, R. Halfmann, S. Devadas, S. Lindquist, and B. Berger. A method for probing the mutational landscape of amyloid structure. *Bioinformatics, Proc. of ISMB 2011*, 2011.
- [134] N. Ollikainen, E. Sentovich, C. Coelho, A. Kuehlmann, and T. Kortemme. Sat-based protein design. *Proc. of ICCAD 2009*, pages 128–135, 2009.
- [135] E.A. Ortlund, J.T. Bridgman, M.R. Redinbo, and J.W. Thornton. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science*, 317(5844):1544–1548, 2007.
- [136] V.G. Ostapchenko, M.R. Sawaya, N. Makarava, R. Savtchenko, K.P. Nilsson, D. Eisenberg, and I.V. Baskakov. Two amyloid states of the prion protein display significantly different folding patterns. *J. Mol. Biol.*, 400(4):908–921, 2010.
- [137] R. Pagh and F.F. Rodler. Cuckoo hashing. *Lecture Notes in Computer Science*, 2161:121–133, 2001.
- [138] A.K. Paravastu, I. Qahwash, R.D. Leapman, S.C. Meredith, and R. Tycko. Seeded growth of β -amyloid fibrils from Alzheimer's brain-derived fibrils produces a distinct fibril structure. *Proc. Natl. Acad. Sci.*, 106(18):7443–7448, 2009.
- [139] B. Park and M. Levitt. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J. Mol. Biol.*, 258:367–392, 1996.
- [140] F. M. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, and C. A. Orengo. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, 31(1):452–455, 2003.
- [141] J.S. Pedersen and D.E. Otzen. Amyloid — a state in many guises: Survival of the fittest fibril fold. *Protein Sci.*, 17:2–10, 2009.
- [142] A.T. Petkova, G. Buntkowsky, F. Dyda, R.D. Leapman, W.M. Yau, and R. Tycko. Solid State NMR Reveals a pH-dependent Antiparallel β -sheet Registry in Fibrils Formed by a β -Amyloid Peptide. *J. Mol. Biol.*, 335:27–260, 2004.
- [143] A.T. Petkova, Y. Ishii, J.J. Balbach, O.N. Antzutkin, R.D. Leapman, F. Delaglio, and R. Tycko. A structural model for alzheimer's beta-amyloid fibrils based on experimental constraints from solid state nmr. *Proc. Natl. Acad. Sci.*, 100(2):383–385, 2003.

- [144] A.T. Petkova, R.D. Leapman, Z. Guo, W.M. Yau, M.P. Mattson, and R. Tycko. Self-Propagating, Molecular-Level Polymorphism in Alzheimer's β -Amyloid Fibrils. *Science*, 307(5707):262–265, 2005.
- [145] G. Pollastri, D. Przybylski, B. Rose, and P. Baldi. Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins Struct. Funct. Bioinf.*, 47:228–235, 2002.
- [146] Y. Ponty. Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: The boustrophedon method. *J. Math. Biol.*, 56(1-2):107–127, Jan 2008.
- [147] B. Punta and B. Rost. Profcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968, 2005.
- [148] C. Puorger, O. Eidam, G. Capitani, D. Erilov, M.G. Grutter, and R. Glockshuber. Infinite Kinetic Stability against Dissociation of Supramolecular Protein Complexes through Donor Strand Complementation. *Structure*, 16:631–642, 2008.
- [149] The PyMOL Molecular Graphics System, Schrödinger, LLC.
- [150] B. Qian, S. Raman, P. Bradley, A.J. McCoy, and D. Baker. High-resolution structure prediction and the crystallographic phase problem. *Nature*, 450:259–264, 2007.
- [151] H. Remaut, R.J. Rose, T.J. Hannan, S.J. Hultgren, S.E. Radford, A.E. Ashcroft, and G. Waksman. Donor-Strand Exchange in Chaperone-Assisted Pilus Assembly Proceeds through a Coordinated β -strand Displacement Mechanism. *Mol. Cell*, 22(6):831–842, 2006.
- [152] G. Rhodes. *Crystallography Made Crystal Clear*. Academic Press: San Diego, 2nd edition edition, 2000.
- [153] P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends. Genet.*, 16(6):276–277, 2000 Jun.
- [154] D. Ringe and G.A. Petsko. Study of protein dynamic by X-ray diffraction. *Methods Enzymol.*, 131:389–433, 1986.
- [155] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, and D. Baker. Protein Structure Prediction Using Rosetta. *Methods Enzymol.*, 383:66–93, 2004.
- [156] B. Rost, J. Liu, D. Przybylski, R. Nair, H. Bigelow, K.O. Wrzeszczynski, and Y. Ofran. *Prediction of protein structure through evolution*. Wiley-VCH, 2003.
- [157] J. Rumbley, L. Hoang, L. Mayne, and S.W. Englander. An amino acid code for protein folding. *Proc. Natl. Acad. Sci.*, 98(1):105–112, 2001.
- [158] S. Shivaprasad and R. Wetzel. Scanning Cysteine Mutagenesis Analysis of A β -(1-40) Amyloid Fibrils. *J. Biol. Chem.*, 281(2):993–1000, 2006.
- [159] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Comput.*, 45(5):810–825, 1985.
- [160] B. Sauerwine and M. Widom. Arrhenius lifetimes of rna structures from free energy landscapes. *J. Stat. Phys.*, 142:1337–1352, 2011.
- [161] M.R. Sawaya, S. Sambashivan, R. Nelson, M.I. Ivanova, S.A. Sievers, M.I. Apostol, M.J. Thompson, M. Balbirnie, J.J. Wiltzius, H.T. McFarlane, A. Madsen, C. Riek, and D. Eisenberg. Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, 447:453–457, 2007.

- [162] A. Schlessinger and B. Rost. Protein flexibility and rigidity predicted from sequence. *Proteins Struct. Funct. Bioinf.*, 61:115–126, 2005.
- [163] C.P. Schultz. Illuminating folding intermediates. *Nature Struct. Biol.*, 7:7–10, 2000.
- [164] G. Schulz. β -barrel membrane proteins. *Curr. Opin. Struct. Biol.*, 10:443–447, 2000.
- [165] J. Selbig, T. Mevissen, and T. Lengauer. Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics*, 15(12):1039–1046, 1999 Dec.
- [166] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.
- [167] B.E. Shakhnovich, E. Deeds, C. Delisi, and E. Shakhnovich. Protein structure and evolutionary history determine sequence space topology. *Genome Res.*, 15(3):385–392, 2005 Mar.
- [168] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, and W. Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330:341–346, 2010.
- [169] S. Shenker, C.W. O’Donnell, S. Devadas, B. Berger, and J. Waldispühl/Na. Efficient Traversal of Protein Folding Pathways using Ensemble Models. *Lect. Notes Comput. Sci., Proc. of RECOMB 2011*, 6577:408–423, 2011.
- [170] M.J. Sippl. Calculation of Conformational Ensembles from Potentials of Mean Force. *J. Mol. Biol.*, 213:859–883, 1990.
- [171] G. Song, S. Thomas, K.A. Dill, J.M. Scholtz, and N.M. Amato. A path planning-based study of protein folding with a case study of hairpin formation in protein G and L. *Pac Symp Biocomput.*, pages 240–51, 2003.
- [172] Spanner. <http://sysimm.ifrec.osaka-u.ac.jp/spanner/>.
- [173] D.M. Standley, H. Toh, and H. Nakamura. Ash structure alignment package: Sensitivity and selectivity in domain classification. *BMC Bioinf.*, 8, 2007.
- [174] R.A. Sutormin, A.B. Rakhmaninova, and M.S. Gelfand. Batmas30: amino acid substitution matrix for alignment of bacterial transporters. *Proteins Struct. Funct. Bioinf.*, 51:85–95, 2003.
- [175] L.K. Tamm, H. Hong, and B. Liang. Folding and assembly of β -barrel membrane proteins. *Biochim. Biophys. Acta, Biomembr.*, 1666:250–263, 2004.
- [176] X. Tang, S. Thomas, L. Tapia, D.P. Giedroc, and N.M. Amato. Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, 381(4):1055–67, Sep 2008.
- [177] L. Tapia, S. Thomas, and N.M. Amato. A motion planning approach to studying molecular motions. *Communications in Information and Systems*, 10(1):53–68, 2010.
- [178] G.G. Tartaglia and M. Vendruscolo. The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.*, 37:1395–1401, 2008.
- [179] P.M. Tessier and S. Lindquist. Prion recognition elements govern nucleation, strain specificity, and species barriers. *Nature*, 447:556–561, 2007.
- [180] P.M. Tessier and S. Lindquist. Unraveling infectious structures, strain variants and species barriers for the yeast prion [psi+]. *Nat. Struct. Mol. Biol.*, 16:598–605, 2009.

- [181] M.J. Thompson, S.A. Sievers, J. Karanicolas, M.I. Ivanova, D. Baker, and D. Eisenberg. The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci.*, 103(11):4074–4078, 2006.
- [182] V.D. Tran, P. Chassignet, S. Sheikh, and J.-M. Steyaert. Energy-based classification and structure prediction of transmembrane beta-barrel proteins. *Proc. of ICCABS 2011*, 2011.
- [183] A. Trovato, F. Seno, and S.C.E. Tosatto. The PASTA server for protein aggregation prediction. *Protein Eng., Des. Sel.*, 20:521–523, 2007.
- [184] D.H. Turner and D.H. Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, 38(Database issue):D280–2, Jan 2010.
- [185] R. Tycko and Y. Ishii. Constraints on supra-molecular structure in amyloid fibrils from two-dimensional solid state NMR spectroscopy with uniform isotopic labeling. *J. Am. Chem. Soc.*, 125:6606–6607, 2003.
- [186] R. Tycko, K.L. Sciarretta, J.P.R.O. Orgel, and S.C. Meredith. Evidence for Novel β -Sheet Structures in Iowa Mutant β -Amyloid Fibrils. *Biochemistry*, 48(26):6074–6084, 2009.
- [187] V.N. Uversky and A.K. Dunker. Understanding protein non-folding. *Biochim. Biophys. Acta.*, 1804(6):1231–1264, 2010.
- [188] M. Vilar, H.T. Chou, T. Lührs, S.K. Maji, D. Riek-Loher, R. Verel, G. Manning, H. Stahlberg, and R. Riek. The fold of α -synuclein fibrils. *Proc. Natl. Acad. Sci.*, 105(25):8637–8642, 2008.
- [189] D. Vitkup, D. Ringe, G.A. Petsko, and M. Karplus. Solvent mobility and the protein 'glass' transition. *Nat. Struct. Biol.*, 7(1):34–38, 2000.
- [190] V.A. Voelz, G.R. Bowman, K. Beauchamp, and V.S. Pande. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.*, 132(5):1526–8, Feb 2010.
- [191] M.J. Volles and P.T. Lansbury Jr. Relationships between the Sequence of α -Synuclein and its Membrane Affinity, Fibrillization Propensity, and Yeast Toxicity. *J. Mol. Biol.*, 366:1510–1522, 2007.
- [192] M. von Bergen, P. Friedhoff, J. Biernat, J. Heberle, E.M. Mandelkow, and E. Mandelkow. Assembly of τ protein into Alzheimer paired helical filaments depends on a local sequence motif (306 VQIVYK 311) forming β structure. *Proc. Natl. Acad. Sci.*, 97(10):5129–5134, 2000.
- [193] G. von Heijne. Membrane protein structure prediction: hydrophobicity analysis and the 'positive inside' rule. *J. Mol. Biol.*, 225:487–494, 1992.
- [194] G.P. Wagner, W. Otto, V. Lynch, and P.F. Stadler. A stochastic model for the evolution of transcription factor binding site abundance. *J. Theor. Biol.*, 247(3):544–553, 2007.
- [195] J. Waldispühl, B. Berger, P. Clote, and J.-M. Steyaert. Predicting Transmembrane β -barrels and Inter-strand Residue Interactions from Sequence. *Proteins Struct. Funct. Bioinf.*, 65:61–74, 2006.
- [196] J. Waldispühl, S. Devadas, B. Berger, and P. Clote. Efficient Algorithms for Probing the RNA Mutation Landscape. *PLoS Comput. Biol.*, 4:e1000124, 2008.
- [197] J. Waldispühl, C.W. O'Donnell, S. Devadas, P. Clote, and B. Berger. Modeling Ensembles of Transmembrane β -barrel Proteins. *Proteins Struct. Funct. Bioinf.*, 71:1097–1112, 2008.

- [198] J. Waldispühl, C.W. O'Donnell, S. Will, S. Devadas, R. Backofen, and B. Berger. Simultaneous Alignment and Folding of Protein Sequences. *Lect. Notes Comput. Sci., Proc. of RECOMB 2009*, 5541:339–355, 2009.
- [199] J. Waldispühl and Y. Ponty. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *Lect. Notes. Comput. Sci., Proc. of RECOMB 2011*, 6577:501–515, 2011.
- [200] J. Waldispühl and J.-M. Steyaert. Modeling and predicting all- α transmembrane proteins including helix-helix pairing. *Theor. Comput. Sci., Special issue on Pattern Discovery in the Post Genome*, 335(1):67–92, 2005.
- [201] J. Waldispühl and J.-M. Steyaert. Modeling and predicting all-alpha transmembrane proteins including helix-helix pairing. *Theor. Comput. Sci.*, 335(1):67–92, 2005.
- [202] Jun Wang and Wei Wang. A computational approach to simplifying the protein folding alphabet. *Nature Struct. Biol.*, 6(11):1033–1038, november 1999.
- [203] X. Wang, N.D. Hammer, and M.R. Chapman. The molecular basis of functional bacterial amyloid polymerization and nucleation. *J. Biol. Chem.*, 283:21530–21539, 2008.
- [204] X. Wang, D.R. Smith, J.W. Jones, and M.R. Chapman. In vitro polymerization of a functional escherichia coli amyloid protein. *J. Biol. Chem.*, 282:3713–3719, 2007.
- [205] C. Wasmer, A. Lange, H. Van Melckebeke, A.B. Simer, R. Riek, and B.H. Meier. Amyloid Fibrils of the HET-s(218-289) Prion Form a β Solenoid with a Triangular Hydrophobic Core. *Science*, 219(5869):1523–6, 2008.
- [206] C. Wasmer, A. Zimmer, R. Sabaté, A. Soragni, S.J. Saupe, C. Ritter, and B.H. Meier. Structural Similarity between the Prion Domain of HET-s and a Homologue Can Explain Amyloid Cross-Seeding in Spite of Limited Sequence Identity. *J. Mol. Biol.*, 402:311–325, 2010.
- [207] S. Will, K. Reiche, I.L. Hofacker, P.F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65, 2007 Apr 13.
- [208] A.D. Williams, E. Portelius, I. Kheterpal, J.T. Guo, K.D. Cook, Y. Xu, and R. Wetzel. Mapping A β Amyloid Fibril Secondary Structure Using Scanning Proline Mutagenesis. *J. Mol. Biol.*, 335:833–842, 2004.
- [209] A.D. Williams, S. Shivaprasad, and R. Wetzel. Alanine Scanning Mutagenesis of A β (1-40) Amyloid Fibril Stability. *J. Mol. Biol.*, 357:1283–1294, 2006.
- [210] W. C. Wimley and S. H. White. Reversible unfolding of β -sheets in membranes: A calorimetric study. *J. Mol. Biol.*, 342:703–711, 2004.
- [211] M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, and P.F. Stadler. Efficient computation of RNA folding dynamics. *J. Phys. A Math. Gen.*, 37(17), 2004.
- [212] C. Wurth, N.K. Guimard, and M.H. Hecht. Mutations that Reduce Aggregation of the Alzheimer's A β 42 Peptide: an Unbiased Search for the Sequence Determinants of A β Amyloidogenesis. *J. Mol. Biol.*, 319:1279–1290, 2002.
- [213] T. Xia, J. Jr. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, and D.H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735, 1998.
- [214] J. Xu, Y. Xu, D. Kim, and M. Li. RAPTOR: Optimal Protein Threading by Linear Programming. *J. Bioinform. and Comput. Biol.*, 1(1):95–117, 2003.

- [215] S. Yang, H. Levine, J. Onuchic, and D.L. Cox. Structure of infectious prions: stabilization by domain swapping. *FASEB J.*, 19:1778–1782, 2005.
- [216] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct. Funct. Bioinf.*, 34(2):220–3, 1999.
- [217] Y. Zhang. Template-based modeling and free modeling by i-tasser in casp7. *Proteins Struct. Funct. Bioinf.*, 69 Suppl 8:108–117, 2007.
- [218] Y. Zhang and J. Skolnick. SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.*, 25:865–871, 2004.
- [219] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.