Insights into protein function from evolutionary and conformational dynamics

by

Philip W. Bransford

B.Bm.E, University of Minnesota (2006)

Submitted to the Department of Biological Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

Author . . . .
Department of Biological Engineering
August 22, 2011

Certified by . . . .
Mark Bathe
Samuel A. Goldblith Assistant Professor of Applied Biology
Thesis Supervisor

Certified by . . . .
Roger D. Kamm
Singapore Research Professor of Biological and Mechanical Engineering
Thesis Supervisor

Accepted by . . . .
Forest White
Chairman, Department Committee on Graduate Theses

Thesis Committee

Mark Bathe

Samuel A. Goldblith Assistant Professor of Applied Biology

Thesis Supervisor

Roger D. Kamm

Singapore Research Professor of Biological and Mechanical Engineering

Thesis Supervisor

Bruce Tidor

Professor of Biological Engineering and Computer Science

Thesis Committee Chairman

Krystyn J. Van Vliet

Paul M. Cook Career Development Associate

Professor of Materials Science and Engineering

Thesis Committee Member

Insights into protein function from evolutionary and conformational dynamics

by

Philip W. Bransford

Submitted to the Department of Biological Engineering
on August 22, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The volume of protein structure data has grown rapidly over the past 30 years, leaving a wake of facts that still require explanation. We endeavored to answer a few open questions on the structure-function relationship of intriguing mechanochemical protein systems. To this end this thesis work contains five studies that offer novel insights into molecular biomechanical systems that may guide future basic research or applications development.

The first study concerns the biophysics of cadherin-mediated cell sorting observed in developing solid tissue. We investigated the evolutionary dynamics of the cadherin superfamily of cell-cell adhesion proteins to infer a structural basis for their paradoxical mixture of pairwise binding specificity and promiscuity. Our analysis predicts a small set of specificity-determining residues located within the protomer-protomer binding interface. The putative specificity-determinants form a design space with potential for engineering novel cell-cell adhesive interactions.

The second study addresses the open question of how to automatically identify regions within a protein that engage in allosteric communication. To identify allostery we developed and tested two computational tools that operate on protein conformational dynamics data. These tools are useful for generating testable hypotheses about proteins with multiple functional sites for the design of non-competitive protein inhibitors.

The third study asks, "What is the consequence of allosteric cooperation between the tandem binding sites in a class of proteins that bundle filamentous actin (F-actin)?" Through simulation we demonstrate that cooperative F-actin bundling tends to strengthen bundles by driving the formation of cross-links between neighboring filaments while depleting F-actin binding sites that are occupied but not cross-linked. We hence propose that allostery may be a natural feature of ABPs with tandem F-actin binding sites if nature indeed selects for sturdy F-actin bundles.

The final two studies examine the impact of two structural perturbations to F-actin on its mechanics. Using structure-based computer modeling we develop a simple explanation for the mechanism by which the structure of actin's polymorphic subdomain 2 mediates 4-fold changes in F-actin's flexibility. We further demonstrate that two calponin homology domains stabilize F-actin by binding in a configuration that tends to relax the stress concentration at actin-actin interfaces.

Thesis Supervisor: Mark Bathe
Title: Samuel A. Goldblith Assistant Professor of Applied Biology

Thesis Supervisor: Roger D. Kamm
Title: Singapore Research Professor of Biological and Mechanical Engineering

Acknowledgments

First and foremost I want to thank my supervisors, Roger and Mark. Together they provided the support and guidance that I needed throughout my PhD thesis and beyond. Along with my supervisors I thank the rest of my committee, Bruce Tidor and Krystyn Van Vliet, for providing advice, comments, encouragement, and patience.

It was great working alongside members of the Mechanobiology lab for my first two years of research followed by two more years in LCBB. Several "Kammsters" provided invaluable guidance when I first joined, so thanks to Seung, Aurore, Peyman, and Nate. In LCBB I benefited greatly from conversations with Do-Nyun and Reza. All of the other lab members I thank for making 500 Tech Square a great place to work.

I am grateful to the BE community as a whole. Thanks to Course XX 2006 for battling through the first year of courses along with me. Special thanks are due to Kristin, Scott, Lorena, Ranjani, Chris, Ta, and Luke.

Finally, thanks to my family for cheering me in all endeavors.

Contents

List of Figures

List of Tables

Chapter 1

Motivation

As of this writing, a decade has passed since the initial publications of the human genome [101, 175]. The Human Genome Project required 13 years and almost $3 billion [2]. The massive time and financial cost incurred to complete the project, in part, reflects the value of resolving the structure of biomolecules for the purpose solving problems in biology.

A key challenge for solving biological problems is interpreting the structural data that is published and deposited in databases. As J. Onuchic said of biology a year after the human genome publications, "[Biology] is faced with a lot of facts that need explanation." (quoted in [96]). Although new data is typically released along with a scientific publication containing some analysis and interpretation, the amount of time, money and effort required to obtain structural data demands that we extract as much information as possible from the data. Indeed, without thorough interpretation of biomolecules, structural biology would be little more than "high-tech stamp collecting" [96].

The technologies that resolve biomolecular structures seem much more efficient at creating data then humans are at interpreting it. Databases of structural information like the RCSB Protein Databank [3], UniProtKB/TrEMBL protein sequences [5], and the Electron Microscopy Database [1] have had their number of entries increase exponentially over the past 40, 30, and 10 years, respectively. Meanwhile, the number of new drugs that the United States Food and Drug Administration (FDA) has approved has increased linearly since 1940 when the FDA was created [4]. The implication is not that every new molecule sequenced or solved by crystallography need be a drug target for humans, but that a drug represents a rigorous benchmark of understanding a biological target. By the standards of the FDA, our deep understanding of biology is increasing only linearly, and thus lagging drastically behind the rate at which we acquire data.

So what is to be done to make use of the data stored in data banks? Computer modeling is a promising approach for obtaining novel insight from biomolecular data. In this thesis we apply a variety of computational techniques to either derive or test hypotheses on the function of proteins and protein assemblies. Out key findings are the following:

Key findings

- The cadherin-cadherin dimer interface is enriched with putative specificity determining residues.
- There still exists an unmet need for unsupervised methods and benchmarks for detecting allostery in proteins from conformation dynamics.
- Cooperative binding of actin binding proteins to bundled F-actin promotes cross-linking over other modes of F-actin decoration.
- The structure of actin subdomain 2 mediates F-actin flexibility.
- Fimbrin and alpha-actinin relax strain energy at protomer-protomer interfaces in F-actin.

Thesis outline

This thesis offers five contributions in the field of molecular biophysics. In Chapter 2 we develop a structural basis for cadherin-mediated cell sorting from the primary structure of the cadherin superfamily. In the process of studying cadherins, we generated questions about the conformational dynamics of proteins in general that we did not know how to solve using an unbiased approach. In Chapter 3 we therefore characterized methods for identify allosteric coupling in proteins. After studying methods for detecting allosteric coupling in proteins, we proceed to address the consequences of allosteric coupling on mechanical organelles comprised of filamentous actin (F-actin) and bundling proteins (Chap. 4). From the structure of F-actin bundles we next discuss the mechanics of F-actin itself. In Chapter 5 we address the implications of recently identified F-actin polymorphisms by computationally deforming the different F-actin forms and characterizing their apparent flexibilities. Next, in Chapter 6, we investigate the mechanism by which actin binding proteins redistribute F-actin's strain energy upon binding. Lastly, we provide an outlook for future work in computational molecular biophysics (Chap. 7).

Chapter 2

Evolutionary dynamics of the cadherin superfamily

Abstract

Cadherins are a superfamily of cell-cell adhesion proteins that mediate cell sorting in animal tissue. Theory suggests that the sorting of cells expressing different cadherin paralogs is a manifestation of modest (i.e. 1 kcal/mol) differences in the binding affinities of homophilic and heterophilic dimers, with homophilic interactions the more favorable. Currently there exists no structural basis to explain the small binding affinity differences between homophilic and heterophilic dimers. To uncover the principles underlying binding specificity we characterized the amino acids sequences of the cadherin superfamily by identifying sequence positions in the putative binding domain that are conserved or variable, distinguish the metazoan paralogs, or are constrained by natural selection to evolve concertedly. This chapter details the inference of the important sequence positions using tools from information theory. We identified sets of conserved residues comprising the core of the binding domain, residues on the binding interface with a statistically significant specificity signal, and small and sparse network of coevolving residues suggestive of allosteric coupling. The results of the analysis provide an experimentally testable model to further develop the theory of cadherin-mediated cell sorting.

2.1 Introduction

The genomic advances required for the evolution of multicellular lifeforms from primitive unicellular ancestors are not fully understood. Presumably the required genetic machinery included genes that regulate differentiation, cell-cell communication, and cell adhesion [144]. Differentiation and cell-cell adhesion are linked, in a sense, because multicellular organisms are comprised of tissues with a tightly regulated spatial distribution of distinct cell types. This chapter concerns the mechanism by which different cell types establish and maintain order in tissue.

Seminal work by Steinberg demonstrated that the ordered arrangement of cells in tissue is based in part on the tendency of cells with the same phenotype to preferentially adhere to one another instead of adhering to cells with a distinct phenotype [159, 158, 160]. Based on his experimental observations, Steinberg formulated what he called the differential adhesion hypothesis. The hypothesis states that the ordered arrangement of cells in tissue is due to surface tension, and that the surface tension is a consequence of differences in adhesion between the different cell types in the tissue. The cell surface molecule or molecules responsible for differential adhesion were not known at the time of Steinberg's first publications on the subject, and it was not until some 20 years after his initial observations that a superfamily of genes called cadherins were proposed as responsible for cell sorting in tissue [163, 123, 122].

Cadherin-mediated cell aggregation and sorting is well documented in the literature, both by *in vitro* [163, 123, 122, 150, 151, 82, 20, 89, 162, 176] and *in vivo* [123, 131, 140] assays. A common *in vitro* assays entails cloning a cadherin gene into an animal cell type that displays a low endogenous level of surface cadherins (e.g. L-fibroblasts, or Chinese hamster ovary (CHO) cells), suspending the cells in media, and then agitating the suspension for a few hours to promote mixing. After a few hours of mixing the sample is removed from its container and examined with a microscope to score the degree of cell aggregation. Cells expressing cadherins tend to aggregate in the presence of calcium, hence the name, but do not aggregate appreciably without calcium.

A variation on the *in vitro* mixing experiment described above is used to measure the degree two expressed cadherin genes can make cells adhere homophilicly or heterophilicly, i.e. the assay measures differential adhesion. Cells expressing different cadherin genes are labeled with different colored dyes so that the type of cadherin expressed in each cell population can be distinguished visually. After mixing the two cell populations as described before, a microscope is used to visualize the aggregates. Figure 2-1 shows the results of cadherin-mediated aggregation and sorting by three genes, cdh1 (E-cadherin), cdh2 (N-cadherin), and cdh6 (Cadherin-6b) [91]. Homophilic adhesion (Fig. 2-1 A-C) induces complete mixing while heterophilic adhesion induces either incomplete mixing (Fig. 2-1 D) or complete segregation (Fig. 2-1

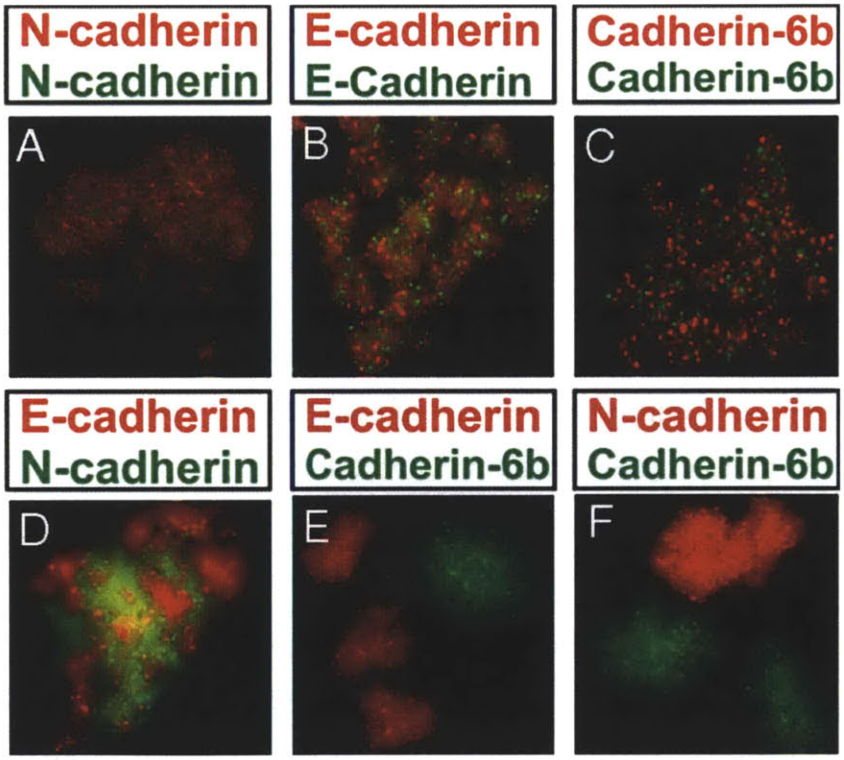

Figure 2-1: Cadherins mediate cell sorting *in vitro*. An *in vitro* cell aggregation assay from [91] demonstrates cadherin-mediated cell aggregation and sorting. Two CHO cell lines expressing genetically identical cadherin genes—N-cadherin, E-cadherin, or cadherin-6b—form interspersed mixtures (A-C). A mixture of cells expressing closely-related paralogous Type I cadherins form homotypic aggregates that adhere to each other (D). A mixture that is equal parts cells expressing a Type I cadherin and cells expressing a Type II cadherin form non-contacting homotypic aggregates (E-F)

.

E-F).

Sorting assays are also conducted *in vivo.* In an *in vivo* assay cadherin-expressing or control cells are injected into a heterogeneous tissue inside an animal. The injected cells tend to partition to the part of the tissue containing cells expressing the same or a functionally similar cadherin [123]. An alternative *in vivo* model used by [140, 131] involved localized ectopic expression of cadherin genes in an animal model. The cells expressing the additional cadherin genes fail to segregate, thereby lending further support to a mechanism of cell sorting controlled by cadherin expression.

Before high-throughput genome sequencing rapidly increased the rate and apparent ease of new gene discovery, almost all of the newly discovered cadherin genes of the 1980s and 1990s were carefully tested by a cell adhesion or cell sorting assay. Table 2.1 summarizes the results of cadherin-mediated cell sorting assays reported in the literature. Importantly, the data show that cadherins favor homophilic adhesions in general and heterophilic adhesions in just a few cases. Moreover, the cadherin pairs with greatest sequence similarity tend to mix while those with less sequence similarity tend to segregate.

With recent advances in genome sequencing, hundreds of cadherin genes have been discovered through comparative genomics. All cadherins found in modern metazoa are the descendants of a pre-metazoan gene family—perhaps resembling the cadherins found in the choanoflagellate *M. brevicollis* [6]—and consequently share structural and functional characteristics. The best studied cadherins, which we focus our attention on here, are from the so-called Cadherin Major Branch (CMB), and in particular the C-1 subbranch [79]. CMB cadherins comprise four or more extracellular (EC) domains separated by conserved calcium-binding regions, typically a single-pass transmembrane domain, and a cytoplasmic domain that can interact with catenins [79]. Some cadherins also contain an amino-terminal pro-domain, although the adhesive form of the molecule that is expressed on the cell surface has had the pro-domain enzymatically removed, thus enabling adhesion via EC homodimerization [129, 69, 68]. Although the number of EC domains varies between cadherin paralogs, their sequences and tertiary structures are conserved [137]. The prominent conserva-

cdh	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19	20	22	24	26
1	2	1	2	0		0	0	0	0	0	0	0		0		2	0		0			
2	1	2		2		0													0			
3	2		2	0										0								
4	0	2	0	2																		
5																0						
6	0	0				2	1	0	2	1	0	0		0			1		2			
7	0					1	2	0	1	0	0	1					2					
8	0					0	0	2	0	0	2	0					0					
9	0					2	1	0	2	2	0	0					1					
10	0					1	0	0	2	2	0	0					0					
11	0					0	0	2	0	0	2	0					0					
12	0					0	1	0	0	0	0	2					1					
13													2									
15	0		0			0								2			0					
16																						
17	2				0											2						
18	0					1	2	0	1	0	0	1		0			2					
19																						
20	0	0				2													2			
22																				2		
24																					2	
26																						

Table 2.1: Summary of known cadherin interactions assembled from the literature. All interactions were measured by a cell sorting assay except for the interactions between cdh1 or cdh5 and cdh17, which was measured by atomic force microscopy. The degree of cell sorting was scored on a scale from 0 to 2. A score of 0 means the cells segregated into non-contacting aggregates (Fig. 2-1 E-F) or bound non-specifically by AFM. A score of 1 means the cells formed contacting homotypic aggregates (Fig. 2-1 C). A score of 2 means the cells intermixed (Fig. 2-1 A-B) or bound specifically by AFM. The data was curated from [150, 151, 131, 82, 20, 89, 162, 176, 18]

tion of the EC domains has led to an apparent paradox: If the EC domains engage in cadherin-cadherin interactions and are structurally very similar, how can the binding interactions be sufficiently specific to give rise to differential adhesion at a cellular level?

A theoretical analysis by Chen *et al.* provides insight into how even subtle differences in cadherin structures can manifest as cell-level differential adhesion that is capable of driving cell sorting [31]. At equilibrium, the expected concentration of cadherin dimers of type i and j, C_{ij}, at a junction between two cells follows the Boltzmann distribution,

$$C_{ij} = C_i C_j e^{-\frac{\Delta G}{RT}} \tag{2.1}$$

where C_i and C_j are the concentrations of monomers on the respective cell surfaces (10,000-200,000 monomers/cell [45, 53]), $\Delta G < 0$ is the free energy change of cadherin binding, and RT is the thermal energy scale. Assuming the free energy change for a homophilic cadherin bond is about -4 kcal/mol [68, 117] and that a heterophilic bond is slightly less favorable at -3 kcal/mol, according to Eqn. 2.1, there ought to be $\sim$ 5 homophilic bounds for each heterophilic bond. Homophilic bonds therefore significantly outnumber heterophilic bonds, and consequently homophilic adhesion would be the dominant cell-cell interaction. Under such conditions different cell types aggregate according to Steinberg's differential adhesion hypothesis (Fig. 2-1E-F). If, on the other hand, the homophilic binding affinity was -11 kcal/mol while heterophilic binding affinity was -10 kcal/mol, the number of homophilic and heterophilic dimers per junction would each be $\sim$ 1000. In this case of ubiquitous strong adhesion, neither homophilic nor heterophilic interactions could dominate and therefore the different cell types would intermix (Fig. 2-1 A-C). Because cadherins bind by a weak strand-swapping interaction (Fig. 2-2) with a ΔG of just a few kcal/mol [69, 68, 117, 91], Chen *et al.* theorized that subtle differences in cadherin structures, and the associated small differences between homophilic and heterophilic binding affinities, can cause cell sorting.

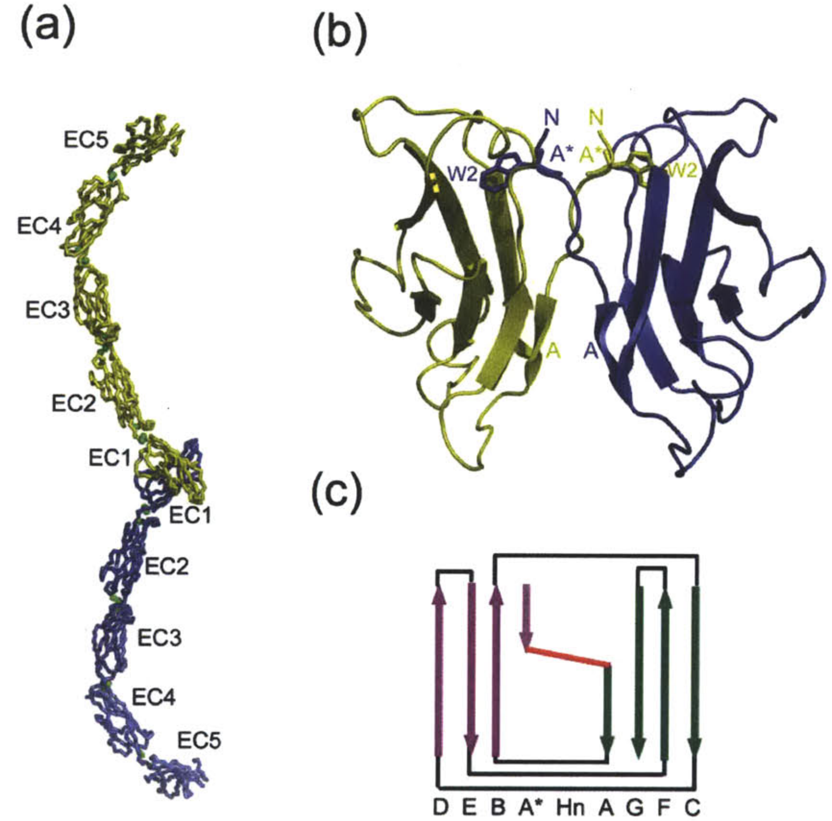

Figure 2-2: Structural models of a cadherin adhesive dimer [137]. Trans dimer interaction of five amino terminal EC domains [22] (a). Details of the A* strand-swapping interaction wherein the tryptophan at the second sequence position docks into the hydrophobic pocket of its cadherin binding partner (b). Schematic representation of the cadherin "Greek-key" secondary structure with labels of the β-strands (c).

Past studies attempted to identify the biophysical determinants of cadherin-mediated cell sorting based on the bulk measurements of cells. Niessen and Gumbiner measured the shear force required to detach cadherin-expressing cells from substrates coated with purified cadherins [121]. Although their measurements detected no adhesion specificity, the cadherins they tested could nevertheless mediate cell sorting when expressed in cells. The authors did not provide an alternative mechanism to explain their results. We suspect that their assay's inability to precisely control for cadherin expression levels and also the use of cadherins with high sequence similarity may have resulted in no discernible binding specificity in the adhesion experiments and expression-level-dominated cell sorting in the sorting experiments. In fact, expression level has been shown to mediate cell sorting. Duguay and Steinberg varied the cadherin expression level in cell lines and measured sorting ability. They concluded that the number of cadherins expressed on the cell surface, C_i and C_j in Eqn. 2.1, as well as the dimerization affinity, ΔG, together control cell sorting [45]. Foty and Steinberg went on to show that the surface tension of a cellular aggregate, modeled as a drop of liquid, is a linear function of cadherin expression level [53]. Therefore the hypothesis that binding affinity contributes to cell sorting is still defensible despite the paper by Niessen and Gumbiner claiming otherwise.

Single molecules biophysics is a natural approach for quantifying the strength of cadhern-mediated adhesion without the confounding effects from varying cell surface expression levels. Panorchan *et al.* measured the rupture force of cadherin homodimers formed between cells using a molecular force probe and found that cdh1 (E-cadherin) bonds are about two to four times stronger than cdh2 (N-cadherin) bonds at two different loading rates [130]. The authors did not measure hetero-dimer rupture forces. Prakasam *et al.* used surface force measurements to compare homophilic and heterophilic cadherin adhesion [138]. Although their instrument could resolve differences in bond energies, there was not a significant difference between homophilic and heterophilic bond energies even though the cadherins expressed in cells mediated cell sorting. Like Niessen's work, the sorting they observed may have been confounded by surface expression levels.

A study by Katsamba *et al.* was the first to verify experimentally Chen *et al.*'s molecular explanation for cell sorting [91]. Their protocol utilized surface plasmon resonance to precisely quantify the dissociation constant of cdh1 (E-cadherin), cdh2 (N-cadherin) and cdh6 homo- and heterodimers. They found the bond strength could be ordered qualitatively as cdh6:cdh6>>cdh2:cdh2>cdh1:cdh2>cdh1:cdh1, and that cdh6 does not bind specifically to either cdh1 or cdh2. The results of their sorting assay (Fig. 2-1) supported the theory presented in Chen *et al.* [31] and Steinberg's differential adhesion hypothesis [159, 158, 160]. In showing that the theory of differential adhesion is supported experimentally, Katsamba *et al.*'s work permits more focused questioning. In particular, their work begs the question as to what features of the cadherin binding interface determine the specificity of homophilic and heterophilic interactions.

A few studies attempted to identify the parts of cadherins that are responsible for the subtle differences between homophilic and heterophilic binding that drive cell-sorting observed *in vitro*. Both Nose *et al.* [122] and Patel *et al.* [131] confirmed by domain swapping and a cell sorting assay that the amino-terminal EC domain, EC1, contains the specificity determining binding site. Nose *et al.* investigated further by attempting to make cdh1 (E-cadherin) mutants that bind specifically to cdh3 (P-cadherin) using site directed mutagenesis. Their mixing experiments on nine distinct mutants found one dual mutation, S78G-S83E to cdh1, that only marginally enhanced the mixing of cdh1 and cdh3 expressing cells [122]—the other eight cdh1 mutants were still specific only for cdh1. Beyond the two studies mentioned so far, to our knowledge, there has not been any attempted mutagenesis studies aimed at identifying a molecular basis for the differences in homophilic and heterophilic binding. The focused set of testable mutations we provide in this work may motivate further studies into this important matter.

In a series of studies we ask three questions about the structure of cadherins to gain some physical insight into the structure-specificity relationship. We base our analysis on the evolutionary record represented in the genomes of a broad phylogenetic sample of metazoan cadherin sequences. First we asked which residues are conserved and which are variable. Next, we identify coevolution between sequence positions to infer important phyisochemical residue-residue interactions. Finally, we ask which sequence positions distinguish the largest cadherin clades from one another, assuming the predicted sequence positions correspond to specificity-determining residues.

2.2 Methods

Data acquisition and preparation

Cadherin amino acid sequences were collected from the Ensembl databank, which is suitable because it contains a comprehensive set of metazoan protein sequences [51]. For each species in the Ensemble databank we acquired the amino acid sequence of every protein with PFAM's extracellular-cadherin (EC) domain identifier, PF00028 [49]. To reduce the dataset to sequences from the Cadherin Major Branch (CMB) [79] we performed a local BLASTp search [8] using the Ensembl sequences as the database and a small set of annotated cadherin sequences from mouse (*M. musculus*) and human (*H. sapiens*) as the queries. The cadherin genes in the reference set were cdh1, cdh2, cdh3, cdh4, cdh5, cdh6, cdh7, cdh8, cdh9, cdh10, cdh11, cdh12, cdh13, cdh15, cdh16, cdh17, cdh18, cdh19, cdh20, cdh22, cdh24, and cdh26. The set of cadherin genes includes all cadherin from the CMB except for desmocollins and desmogleins. We excluded desmocollins and desmogleins for lack of experimental evidence demonstrating that they function in cell sorting. The BLASTp search identified unannotated Ensembl sequences that were orthologous to the genes in the annotated reference set. Ensemble sequences that did not match any of the reference sequences were removed from the data set. Our procedure resulted in 460 cadherin sequences from the CMB.

We chose to focus our analyses on the putative extracellular cadherin binding (EC) domain. All of the cadherins in the CMB have five EC domains except for the 7-domain (7D) family, cdh16 and cd17, which have seven EC domains. Based on the position of a conserved tryptophan residue Hulpiau *et al.* proposed that the 7D cadherin EC3 domain, (numbering from the amino-terminal EC domain) is the ancestor of EC1 found in the 5-domain cadherins, and that EC1 and EC2 in the 7D family were the result of domain duplication [79]. We therefore assumed that EC3 is the binding domain of the 7D cadherins, while EC1 is the binding domains of all of the other cadherins.

We generated 22 multiple sequence alignments (MSA), one for each of the 22 cadherin genes identified by the BLASTp search described above. From the alignments we isolated the extracellular-cadherin binding domain (ECB) by visually searching for the conserved tryptophan at sequence position 2 (isoleucine in cdh13) and the first calcium binding site motif DXXDX. We excised the ECB from all the sequences alignments, pooled the fragments, and the re-aligned all of the ECB domains. All of the sequence alignments were computed with the `MAFFT-G-INI-i` algorithm [90] which is suitable for sequences with conserved starts and ends.

Sequence conservation analysis

We used Shannon entropy to quantify the variation of sequence positions in the ECB domain. In equation A.1 x_i is one of the twenty natural amino acids or a gap introduced by the sequence alignment algorithm and $p(x_i)$ is the observed frequency of amino acid x_i in column i of the multiple sequence alignment. We denote this entropy with H_{21} because it utilizes a 21-letter amino acid alphabet. We also calculated a 7- and 8-letter entropy from physiochemical amino acid alphabets reviewed in [173]. The 7-letter alphabet is AVLIMC, FWYH, STNQ, KR, DE, GP, and a gap character. The 8-letter alphabet is LIVMFYWA, DENQ, KRH, ST, P, C, G, and a gap character. The physiochemical amino acids alphabets served as a qualitative measure of robustness for our information-theoretic calculations.

Inference of coevolution between sequence positions

Our starting dataset was the ECB sequence alignment used for sequence conservation analysis. Coevolution analysis requires sequence diversity, so we removed redundant sequences from the alignment using the EMBOSS program `skipredundant` with a pairwise sequence identity threshold of 62%. The final sequence alignment comprised

154 sequences with a median sequence identity of 33% and no two sequences with greater than 62% identity.

We employed an approach for inferring coevolution described originally by Atchley [11], with further modifications described by Buslje *et al.* [28]. We estimate the covariance of sequence positions i and j using mutual information (Eqn. A.2). Conservation due to phylogeny biases the coevolution signal, so we followed the suggestion of [28] and disregarded columns where $H(X_i) < 0.3\log(Q)$, where Q is the size of the amino acid alphabet. When constructing the contingency table for columns i and j, N_{ij}, we disregarded sequences with a gap in either column. For some column pairs this exclusion condition lead to too few sequences to estimate the mutual information accurately according to a heuristic criteria for predicting contacting residues from coevolution [111]. We therefore required at least 125 sequences per contingency table, otherwise we defined the mutual information of the column pair as zero.

From the contingency table, N_{ij}, with $\sum_{i,j} N_{ij} = N$, we estimated the mutual information from the pairwise frequencies, f_{ij}, with pseudocounts to account for unobserved amino acids [28].

$$\begin{aligned} p(x_i, x_j) &\approx f(x_i, x_j) \\ &= \frac{1}{\lambda Q + N}\left[\frac{\lambda}{Q} + N_{ij}\right] \\ p(x_i) &\approx f(x_i) \\ &= \sum_{i=1}^{Q} p(x_i, x_j) \end{aligned}$$

Note that the two limiting case behave as expected

$$\begin{aligned} \lim_{\lambda \to 0} f(x_i, x_j) &= \frac{N_{ij}}{N} \\ \lim_{\lambda \to \infty} f(x_i, x_j) &= \frac{1}{Q^2} \end{aligned}$$

The frequency definition was used in [178] and [28], and the later found that $\lambda/Q = 0.05$ is optimal for a 20-letter alphabet that excludes gaps. The mutual information between columns i and j according to Eqn. A.2, with frequency replacing probability, is

$$I_{ij} = \sum_{x_i \in X_i} \sum_{x_j \in X_j} f(x_i, x_j) \log \frac{f(x_i, x_j)}{f(x_i) f(x_j)}$$

To correct for bias introduced by phylogeny we apply the average product correction (APC) [46], defined for column pairs as

$$APC_{ij} = \frac{\left(\sum_{i=1}^{M} I_{ij}\right)\left(\sum_{j=1}^{M} I_{ij}\right)}{\sum_{i=1}^{M} \sum_{j=1}^{M} I_{ij}}$$

where M is the number of columns in the sequence alignment. The phylogeny-corrected mutual information is then $I_{ij}^{(p)} = I_{ij} - APC_{ij}$.

To asses the statistical significance of $I_{ij}^{(p)}$ we generate 100 randomly shuffled sequence alignments and defined a Z-score for $I_{ij}^{(p)}$ in the standard way

$$Z_{ij} = \frac{I_{ij}^{(p)} - \left\langle I_{ij}^{(p)} \right\rangle}{\sigma\left(I_{ij}^{(p)}\right)}$$

where $\left\langle I_{ij}^{(p)} \right\rangle$ and $\sigma\left(I_{ij}^{(p)}\right)$ are the mean and standard deviation, respectively, of $I_{ij}^{(p)}$ computed from shuffled alignments.

Paralog specificity analysis

We generated a phylogenetic tree from the original ECB MSA with the computer program PHYLIP [47]. We added an outgroup to the ECB MSA by aligning the complete amino acid sequence of BS-cadherin from the uchordate *B. schlosseri* [105] using `MAFFT-L-INS-i` [90]. An outgroup serves as a monophyletic reference sequence to compare against all of the other sequences in the alignment. We next generated 100 bootstrap samples of the sequence alignment from which we generated 100 corresponding distance matrices using the JTT matrix [86]. The distance matrices were inputs for the Neighbor-joining (NJ) clustering algorithm which produced 100 phylogenetic trees. From the resulting 100 NJ trees we computed a majority-rule consensus tree and reassigned branch length via maximum likelihood, again with the JTT substitution matrix.

From the phylogenetic tree we identified the first five clades following the divergence of pre-metazoan and metazoan cadherins. A clade defines a point in genetic history at which an ancestral gene duplicated and its offspring diverged. At the first clade the CMB splits into two gene groups, and at the second clade one of the branches splits again to make three cadherin gene groups, and so on. In our analysis we consider just the first five clades, meaning we conducted one analysis of specificity with the sequences divided into two, three, four, five, or six cadherin subtypes. For each analysis we computed the mutual information between the subtypes the sequences belong to (x_i in Eqn. A.2) and the amino acid character in a particular column (y_i in Eqn. A.2).

The mutual information of the columns of the sequence alignment is not useful without a comparison to an expected value. We computed an expected mutual information using Protocol I described in [118]. Briefly, we first shuffled the order of amino acids in the columns of the sequence alignment 5000 times, each time computing a randomized mutual information, I_i^{sh}, to generate a distribution of mutual information, $P(I_i^{sh})$. From the distribution of the shuffled mutual information we estimated an expected mutual information, $I_i^{exp} = \alpha I_i^{sh} + \beta$. We obtained the constants α and β by linear regression of I_i versus I_i^{sh}. From the linear equation we computed the mean and standard deviation of the expected mutual information,

$$\begin{aligned} \langle I_i^{exp} \rangle &= \alpha \left\langle I_i^{sh} \right\rangle + \beta \\ \sigma(I_i^{exp}) &= \alpha\sigma(I_i^{sh}) \end{aligned}$$

Statistical significance is assigned by computing a Z-score and its corresponding p-value

$$Z_i = \frac{I_i - \langle I_i^{exp} \rangle}{\sigma(I_i^{exp})}$$

$$P(Z_i \geq z_i) \;=\; 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-\frac{z_i^2}{2}} dz$$

Because we assessed the statistical significance of each column in the sequence alignment as though the columns are independent, we applied the Bonferonni correction for multiple hypothesis testing to reduce the number of potential false positives. The Bonferroni correction involves dividing the p-value required for statistical significance by the number of statistical tests performed. Here we divided by the number of columns in the alignment.

2.3 Results and discussion

The analysis in this chapter assumes that the ECB structure is conserved, and therefore specificity depends primarily on sequence variation and not structure variation. An example protein family with specificity encoded primarily by sequence is the bZIP coiled-coils [52]. Unfortunately, as of this writing the number of solved ECB structures is much less than the number of amino acid sequences [79], so we must estimate structural conservation indirectly. The absence of extended gaped regions in the sequence alignment (Fig. 2-3), with exception of loops opposite the putative strand-swap interface, supports the assumption that the ECB structure is conserved. Moreover, a structure-based superposition of 22 EC domains showed that the 3D structure is conserved, with a root mean square displacement less than 5 Å. Perhaps as more ECB crystal structures are solved and reported, we may relax our assumption by directly incorporating geometry into our sequence analysis.

Sequence conservation

Although the 20 amino acids can be partitioned into a number of physiochemical subgroups for the purpose of calculating information theoretic quantities [173], the choice of amino acid alphabet does not affect the computed structure of the ECB sequence dataset. We compared three different alphabets and found that the sequence conservation profile are qualitatively similar (Fig. 2-4). Therefore for the remainder

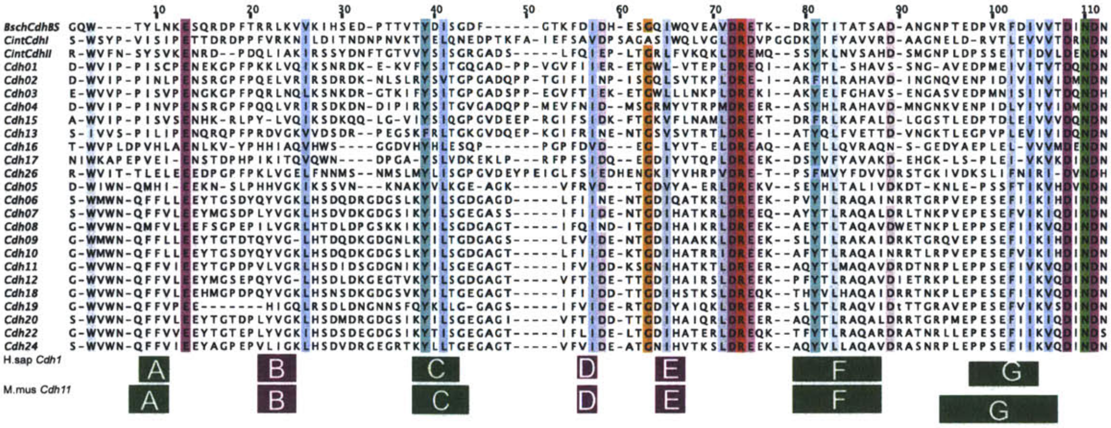

Figure 2-3: Consensus protein sequence alignment of the ECB domains. Highlighted columns indicate conserved sequence positions. Each cadherin sequence identified with an Arabic number is the majority rule consensus of all of the orthologs in the complete MSA. Uchordate cadherins include BS-cadherin from *B. schlosseri* and cdhI and cdhII from *C. intestinalis.* Columns corresponding to positions with established functions include 3 and 39, which participate in strand swapping, and 13, 72-74, and 108-112, which coordinate Ca^{2+}. The coordinates of the beta-strands (see Fig. 2-2 for label convention) of *H. sapiens* cdh1 and *M. musculus* cdh11 are shown below the alignment.

of this chapter we highlight the results computed with the 21-letter alphabet.

The jaggedness of the raw sequence conservation profile, as well as the fact that the ECB is entirely beta-strands, suggests that nature preferentially conserves residue exposed to one particular environment. We partitioned ECB residues into buried or exposed subtypes and found that the buried residues are preferentially conserved over solvent-exposed residues (Fig. 2-5). We hypothesize that the buried and conserved residues are either part of the hydrophobic core of the protein, and therefore required for proper folding, or are part of the hydrophobic binding pocket involved in strand-swapping.

To interpret sequence conservation using the ECB structure, we mapped the sequence positions from the alignment onto two reference proteins for which there are quality structural models in the Protein Databank (Fig. 2-6). For our purpose we define the variable (red) or conserved (blue) positions as those with Shannon entropy at least one standard deviation above or below the mean entropy, respectively. Conserved residues reside in the hydrophobic core, binding pocket, or coordinate

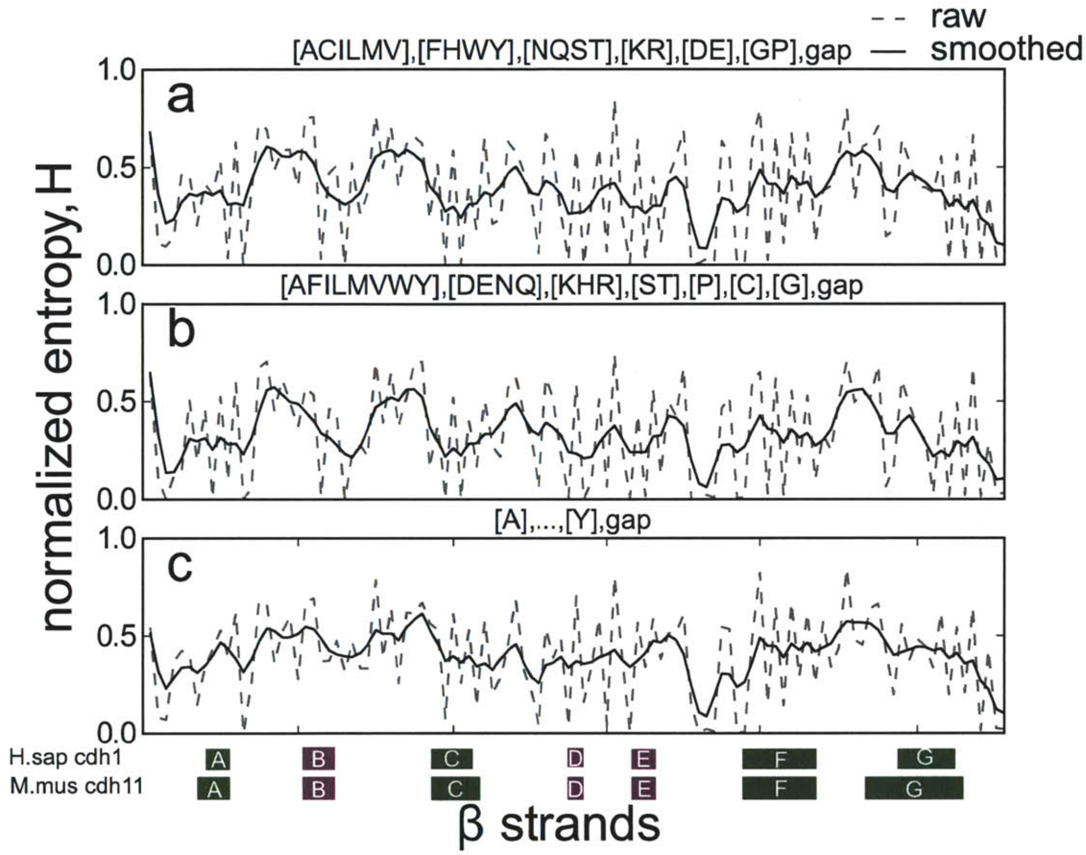

Figure 2-4: ECB sequence conservation profile. The entropy profiles are based on 7 (a), 8 (b), or 21 (c) letter alphabets, listed as titles of the panels. Both raw and smoothed profiles are shown, with the smoothed profile the result of a 7-wide Bartlett window function. The coordinates of beta-strands are indicated.

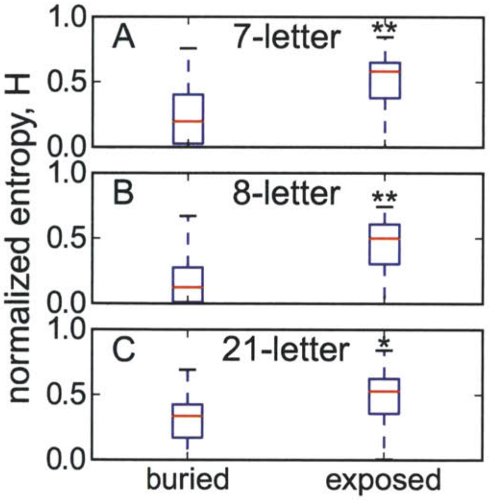

Figure 2-5: Buried residues are more conserved than solvent-exposed residues. Residues were classified as buried if their solvent accessible surface area in PDB ID 2O72 or 2A4E was less than 20% of the area within a G-X-G tripeptide, otherwise the residue was defined as solvent-exposed. The buried residues are more conserved than the exposed residues according to all 3 amino acid alphabets (A-C). $*p < 10^{-7}$; $**p < 10^{-10}$,

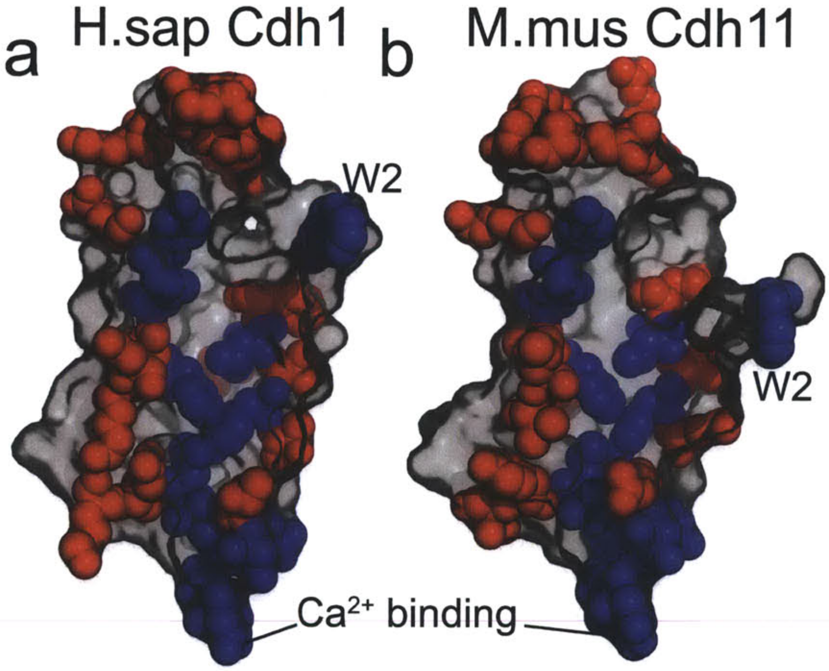

Figure 2-6: Spatial distribution of conserved and variable residues in ECB. The Type I cadherin is *H. sapiens* cdh1 (H.sap Cdh1, PDB ID 2O72) and the Type II cadherin is *M. musculus* cdh11 (M.mus Cdh11, PDB ID 2A4E).

Ca^{2+}. We conjecture that the conserved residues in the hydrophobic core contribute to the domain's stability and are necessary for proper folding, while the those in the binding pocket form favorable interactions with the conserved tryptophan involved in strand-swapping. Because both the hydrophobic pocket and tryptophan are common to most ECB domains, we may attribute the promiscuous binding observed experimentally (Tab. 2.1) to this conserved binding interaction.

Interestingly, residues that participate in the strand-swapping interaction are not necessarily conserved. Type I/II cadherins dimers form a salt bridge between E89/E87 and the N-terminus, yet that sequence position is not conserved (Tab. 2.2). The sequence positions corresponding to *M. musculus* cdh11 residues Y13, V19, and L20 are also variable yet they contribute to the Type II binding interface, while the corresponding sequence positions in *H. sapiens* cdh1 do not participate in strand-swapping. The variable positions that participate in strand-swapping may confer

Variable sequence positions			**Conserved sequence positions**		
Hsap cdh1	Mmus cdh11	H_{21}	Hsap cdh1	Mmus cdh11	H_{21}
E13	Y13‡	0.681	W2†	W2‡	0.070
K14	T14	0.660	E11^{*B}	E11^{*B}	0.000
K19	V19‡	0.673	Y36†	Y37‡	0.072
N20	L20‡	0.693	G40	G41	0.137
K28	I28	0.786	F51^{B}	F58^{B}	0.035
K30	S30	0.642	I53^{B}	I50^{B}	0.160
K33	N34	0.671	G55^{B}	G55	0.005
T45	T46	0.683	L63^{B}	L63	0.144
I52	V49	0.704	E64*	D64	0.005
E56	K53	0.802	R65^{B}	R65*	0.020
T75	T72	0.822	E66*	E66	0.010
F77	M74	0.645	A72	A69	0.005
H79	Q76	0.669	T73	Q70	0.005
G85	N83	0.844	L76^{B}	L73^{B}	0.155
V88	L86	0.640	S78^{B†}	A75^{B‡}	0.143
E89†	E87‡	0.664	N102*	N100*	0.014
Q101*	I99*	0.643	N104*	N102*	0.024

Table 2.2: Conserved and variable ECB domain sequence positions. * Residues that participate in strand-swapping interactions. § Residue positions that participate in Type II but not Type I cadherin protomer-protomer interface. † Residues in the ECB domain that that bind to calcium ions. ^{B} Buried residues, i.e. less than 20% their G-X-G surface area exposed.

specificity to cadherin binding interactions.

Inference of coevolution

A small sparse network of ECB residue pairs demonstrate coevolution. The distribution of z-scores contains a gap where $1000 < Z < 1400$ (Fig. 2-7). After the gap the histogram contains eight residues pairs listed in Table 2.3 that are the most likely candidates for coevolving residue pairs.

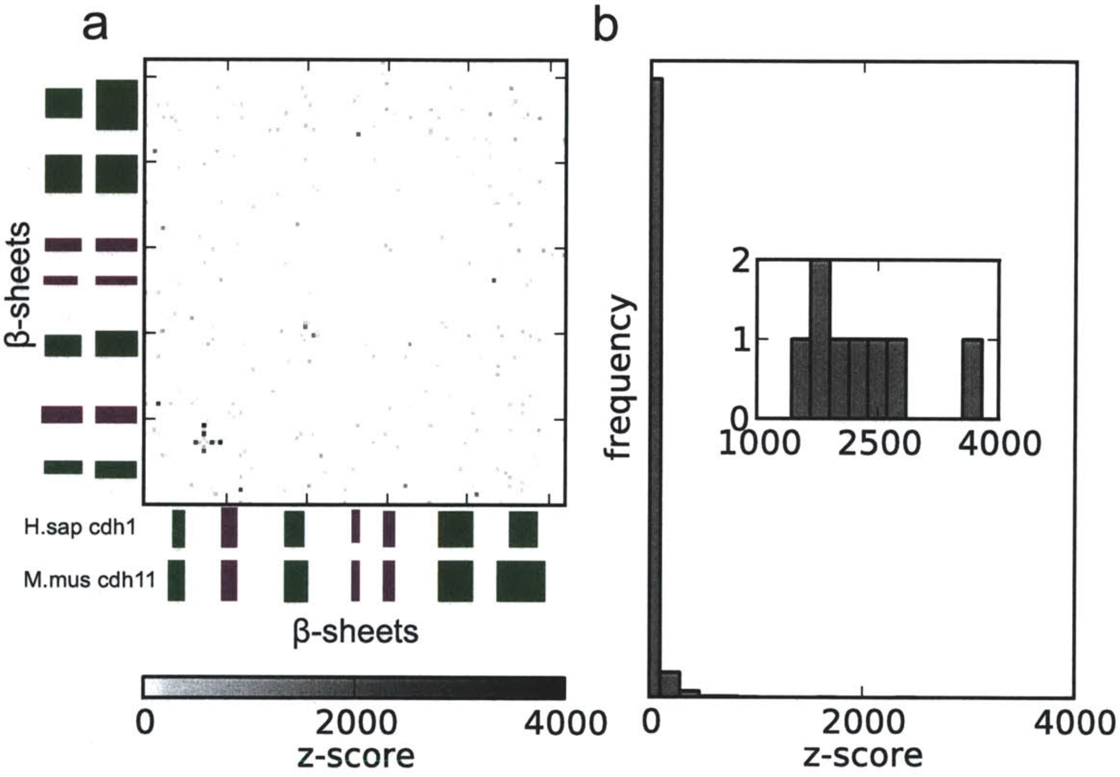

Figure 2-7: ECB contains eight putative coevolving residues pairs. A heatmap of z-scores shows the location of statistically significant coevolving pairs as a function of sequence position (a). The distribution of z-scores has a long and sparse tail (b). Eight residue pairs have z-scores in the long tail, i.e. $z > 1000$ (b inset).

The set of candidate coevolving residues are enriched with residues that participate in strand-swapping in either Type I homodimers or Type II homodimers or both (Tab. 2.3). We hypothesize that nature permitted coordinated evolution to maintain the sorting ability of the cadherin gene family.

Hsap cdh1		Mmus cdh11		Z-score	$\min\{H_i, H_j\}$
K19	G15	†V19	G15	3792	0.644
F17	G15	D17	G15	2767	0.639
*I24	I4	*L24	‡W4	2458	0.545
G15	E13	G15	‡Y13	2349	0.644
A87	I53	P85	I50	1929	0.433
G42	G40	G43	G41	1678	0.448
S83	*V3	D81	*V3	1666	0.615
P65	†P5	T62	N5	1430	0.626

Table 2.3: Eight coevolving sequence position pairs with the greatest z-scores. * Residues involved in Type I and Type II cadherin strand-swapping. † Residues involved in Type I cadherin strand-swapping. ‡ Residues involved in Type II strand swapping.

For further insight into the possible function of the coevolving residues, we mapped sequence positions to the structural models described before (Fig. 2-8). Type I cadherins comprise five coevolving intramolecular contacts. The same sequence position pairs in Type II cadherins are either non-contacts, intermolecular contacts, or also intramolecular contacts. The lone intermolecular contact that is specific to Type II cadherins, L24-W4, may contribute to binding specificity.

Paralog specificity analyis

The topology of the phylogenetic tree we computed details the order in which cadherin genes diverged from a common ancestor (Fig. 2-9). From a common ancestor cadherin, the first gene duplication event generated the 7D family (clade I), named so because cdh16 and cdh17 have seven EC domains rather than five. Type I and II cadherins diverge at the next duplication event (clade II). The following three duplication events diversify the Type I cadherin subfamily by creating cdh13, cdh15, and cdh26 (clade III-V, respectively). At each clade we asked which sequence positions distinguish the cadherin subtypes from each other.

Paralog specificity analysis identified a set of putative specificity determining sequence positions of the five clades we analyzed (Fig. 2-10). The specificity determinants are concentrated near β-strands A, B, G, and at the loop between strands C and D. The specificity profiles differ little among the analyzed clades. For brevity

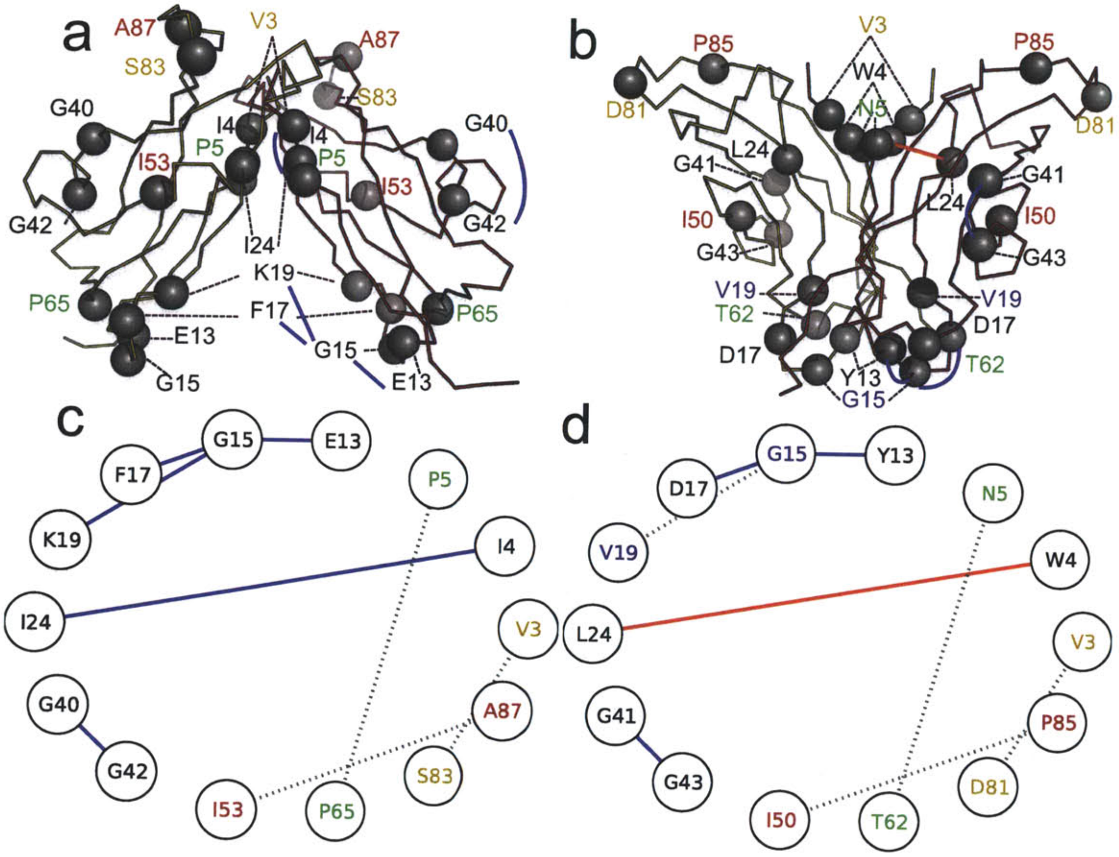

Figure 2-8: Location of coevolving residues pairs in Type I (a, c) and Type II (b, d) ECB models. The blue lines represent intramolecular contacts. The red lines represent intermolecular contacts found in the strand swapping model. The dashed gray lines are coevolving pairs that are not in contact.

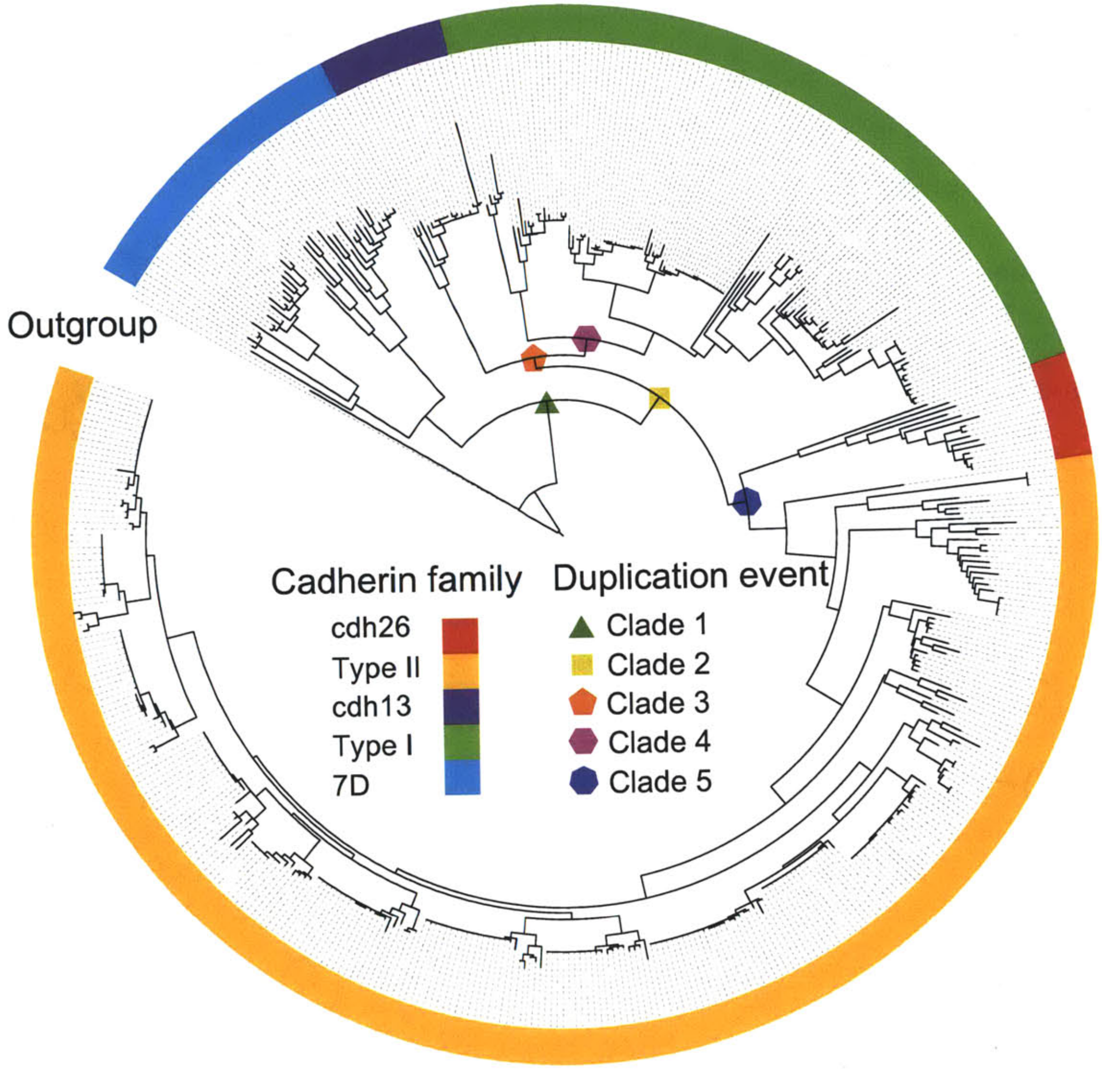

Figure 2-9: Phylogenetic tree of metazoan ECB. The first clade (green triangle) marks the duplication event that lead to the 7D genes diverging from the other cadherins in the major branch. At the second clade (yellow square) genes similar to Type I or II cadherins diverged. At the next divergences point (orange pentagram) cdh13 splits from other Type I cadherins, followed by cdh15 splitting from the remaining Type I cadherins (magenta hexagon). Finally, the cdh26 and Type II cadherins diverge (blue heptagon).

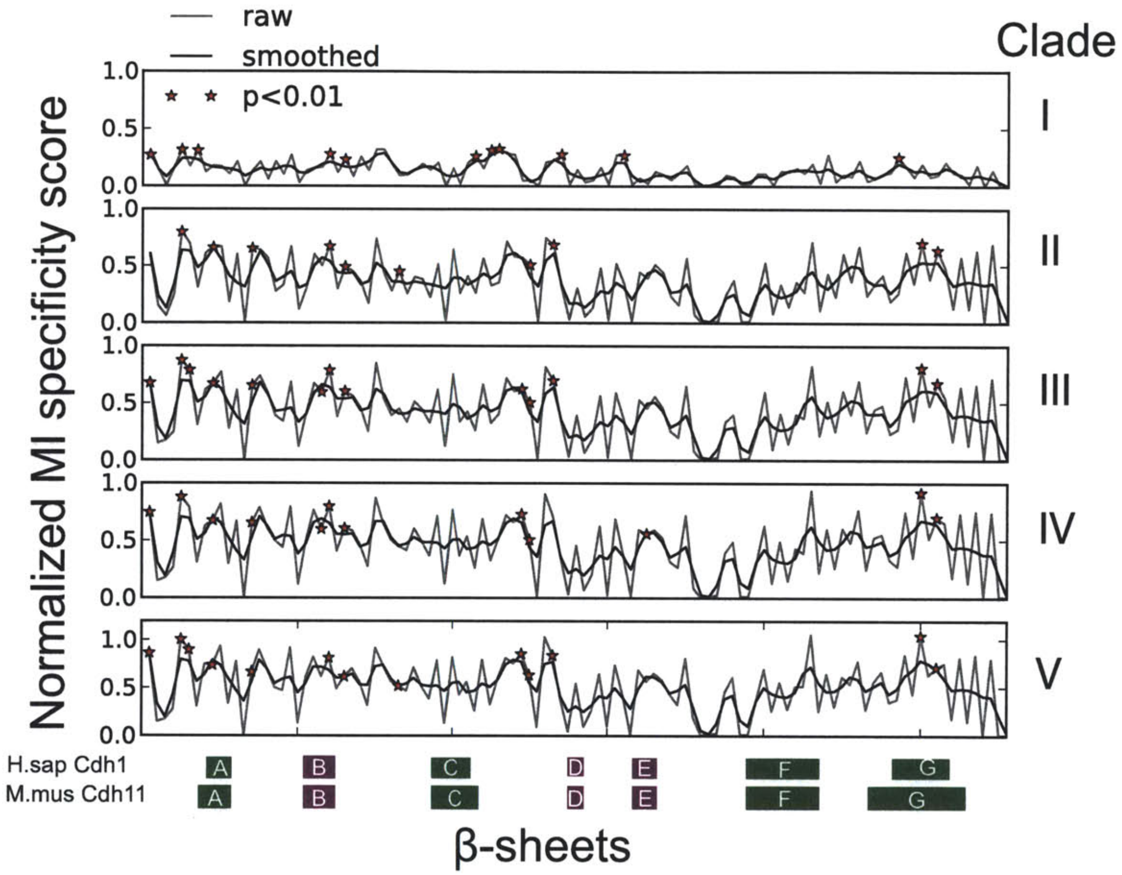

Figure 2-10: Specificity scores of the five clades. The paralog specificity is shown as a function of sequence positions. Statistically significant specificity determinants are marked with red stars.

we focus on the specificity of clade V, but report a summary of the other clades in Appendix A.

Every statistically significant specificity determinant has a well-established function in binding (Tab. 2.4). Every residues except for N12/E12, which coordinates Ca^{2+}, participates in either Type I or Type II strand-swapping. The specificity of the calcium-coordinating residue is intriguing, as calcium is thought to stabilize the putative transition state of the strand-swapping reaction [157]. We therefore propose that binding kinetics may also be specific among the cadherin subtypes.

Hsap cdh1	Mmus cdh11	p-value
D1[†]	G1[‡]	5×10^{-9}
I4	W4[‡]	4×10^{-11}
P5[†]	N5	8×10^{-7}
I7	F7[‡]	7×10^{-9}
N12	E12[*]	3×10^{-14}
V22	G22[‡]	1×10^{-8}
I24[†]	L24[‡]	6×10^{-15}
M92[†]	S90[‡]	2×10^{-13}
I94	F92[‡]	1×10^{-11}

Table 2.4: Predicted specificity-determining residues corresponding to clade V. Residues with established functions include those that participate in Type I cadherin strand-swapping, [†], Type II cadherin strand-swapping, [‡], or residues that that coordinate Ca^{2+}, [*].

To validate our predicted specificity determinants we mapped the mutual information Z-scores onto the cadherin structural models from before. The binding interfaces of both Type I and Type II cadherin dimers are enriched with putative specificity determining residues (Fig.2-11). The putative sequence positions therefore define a space with which one can, in principle, engineer novel cadherin specificity via mutagenesis.

2.4 Concluding remarks

Based on our results, we propose a few strategies for engineering cadherin specificity for engineering biology. The conserved residues lining the hydrophobic pocket into

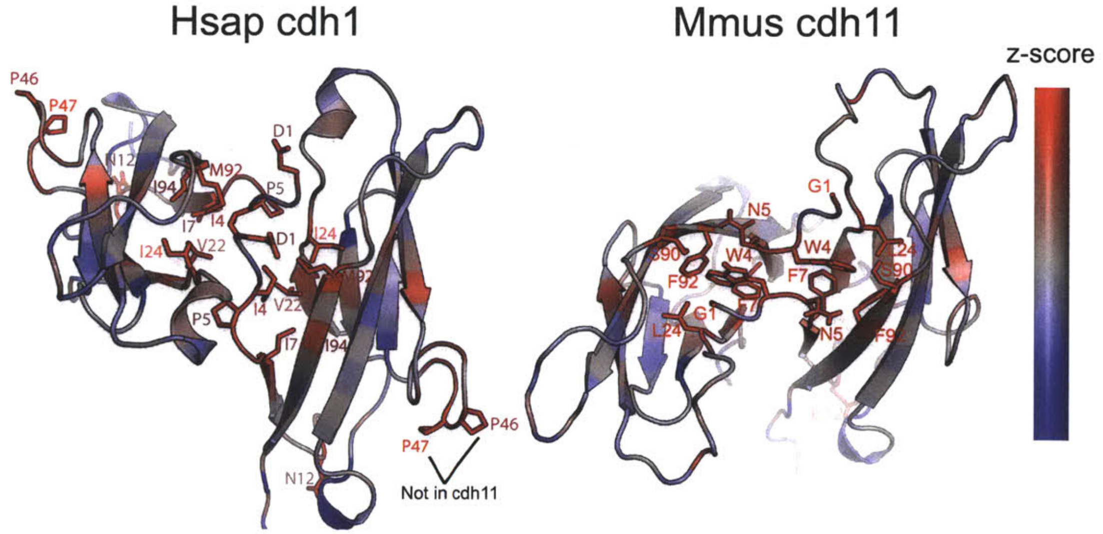

Figure 2-11: Specificity-determining residues are located on the ECB-ECB strand-swap interface. Sequence positions corresponding to $p < 10^{-5}$ are shown as sticks.

which the conserved tryptophan docks may confer promiscuity. Therefore mutating either W2 or Y36 (Type I indexing) or both may produce an orthogonal cadherin system that does not interact with native cadherins. Coevolution analysis suggests that the W4-L24 interaction that is specific to Type II cadherins may confer specificity. Mutating one or bother residues may modulate the specificity of Type II cadherins, which is a large subfamily of cadherins with known promiscuity (Tab. 2.1). Finally, the binding site residues predicted to confer specificity provide a extensive space of residues that can be mutated either experimentally or computationally as part of a screen for novel cadherin sequences with programmed specificity or promiscuity. The number of novel cadherin designs appears extensive.

Chapter 3

Inferring allosteric coupling in proteins from conformational dynamics

Abstract

Allostery is the dynamical coupling of functional sites within biological molecules. How to robustly identifying networks of amino acids comprising an allosteric network is an important question in protein science. Popular approaches include molecular dynamics, graph-theoretic analysis of protein crystal structures, and amino acid coevolution analysis. This chapter details two methods for identifying correlated networks from protein conformational dynamics. One method is based on clustering and the other is based on a community detection algorithm. We apply the approaches to canonical allosteric proteins and cadherins and find that the approaches produce putative correlated networks with distinct topologies. The clustering approach identifies networks that are spatially disconnected while the community detection approach produces networks that are spatially compact. With some basic insight in hand, we propose scaling the two methods for a database-wide study of proteins dynamics. Further, understanding of allosteric coupling may help inform the design of proteins and protein inhibitors for applied biology.

3.1 Introduction

In Chapter 2 we inferred physiochemical interactions of the residues in proteins based on the covariation of amino acid sequence positions that are guided by evolutionary forces. Because our input data was only amino acid sequences, we merely assumed the inferred coupling between amino acids reflected the actual conformational dynamics coupling in the molecule. The lack of actual 3D dynamical data to inform the results of coevolution analysis drove us to ask questions about the actual dynamics of cadherins and other proteins of interest. Specifically, we asked which residues in the protein comprise a correlated network.

The objective of this study was to develop unsupervised tools capable of identifying a correlated network of amino acids in proteins from conformational dynamics. To our knowledge this problem has not been formally addressed in the literature, although there are some well-cited attempts. For example, del Sol *et al.* applied graph theoretic principles to protein structure models to identify residues that are important for efficient allosteric communication between known functional sites [39]. Other work has attempted to define networks of residues that mediate signaling. Statistical coupling analysis (SCA) infers an allosteric network by identifying covarying amino acid sequence positions [109, 161]. Finally, in other work del Sol *et al.* applied graph theory and modularity maximization [143] to isolate putative networks of allosterically coupled residues from protein crystal structures [156]. Importantly, none of the methods described so far incorporated dynamical information in the predictions of correlated networks. Our analysis therefore differs because we construct our putative correlated networks from dynamical data derived from a molecular mechanics model.

This chapter describes two approaches for identifying putative correlated networks in proteins. We first present a mathematical definition of a correlated network then describe the two methods of identifying the network with a computer and conformational dynamics data. Next, we apply the approach to a few canonical examples of allosteric proteins. We conclude by highlighting the relevant topological differences of the networks computed by the two approaches.

3.2 Methods

Normal mode analysis

We derive our predictions of allostery from the conformational dynamics of protein structural models. Correlations in amino acid fluctuations are typically derived from either molecular dynamics (MD) simulations or normal mode analysis (NMA). In this work we use NMA for several reasons. First, with NMA we can verify the convergence of the atomic fluctuations by inspecting the computed eigenvalues and extrapolating their spectra to estimate the truncation error due to the linear approximation of the energy landscape. In contrast, MD simulations must be analyzed statistically to estimate the convergence of any quantity of interest, including atomic fluctuations. The statistical nature of MD also means several independent simulations are required to evaluate the convergence of a quantity, whereas NMA requires only one calculation.

Normal mode analysis is a linearized analysis of Newton's equation of motion near a stationary point on the energy landscape. The general form of Newton's second law for a system of points is

$$\mathbf{M}\ddot{\mathbf{r}} = -\nabla_{\mathbf{r}} U\left(\mathbf{r}\right)$$

where $\mathbf{M}$ is the mass matrix, $\mathbf{r}$ is position vector of the points, and U is the potential energy of the system. If we assume the energy of the system is near a minima we can approximate the energy gradient (i.e. force) by a Taylor series expansion $-\nabla_{\mathbf{r}} \approx -\mathbf{K}\mathbf{u}$, where $\mathbf{K}$ is the so-called stiffness matrix and $\mathbf{u} = \Delta\mathbf{r}$ is the displacement vector. The assumption that the system is near equilibrium leads to a linear representation of Netwon's second law

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{0} \tag{3.1}$$

The equation of motion admits a harmonic solution of the form

$$\mathbf{u}(t) = \mathbf{x}\cos\left(\omega t + \delta\right) \tag{3.2}$$

where $\mathbf{x}$ is the fluctuation vector, ω is the vibrational frequency, and δ is an arbitrary phase lag. Substituting Eqn. 3.2 into Eqn. 3.1, one arrives at a generalized eigenvalue problem

$$\left(\mathbf{M}\omega^2 - \mathbf{K}\right)\mathbf{x} = 0$$

which admits $3N - 6$ nontrivial solutions that describes the position of the particles as a function of time. In practice, one can covert the generalized eigenvalue problem to a standard eigenvalue problem by mass-normalizing the displacement vectors

$$\left(\mathbf{I}\omega^2 - \tilde{\mathbf{K}}\right)\mathbf{y} = 0 \tag{3.3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{y} = \mathbf{M}^{1/2}\mathbf{x}$. The eigenvectors $\{\mathbf{x}_1, ...\mathbf{x}_{3N}\}$ are mass-orthonormal, i.e. $\mathbf{x}_i^T\mathbf{M}\ \mathbf{x}_j = \mathbf{y}_i^T\mathbf{x}_j = \delta_{ij}$. The set of eigenvectors and eigenvalues comprise the dynamical data required to describe conformational dynamics of a protein.

We performed normal mode analysis on proteins with the molecular mechanics computer program CHARMM [25]. We minimized the energy of the structure with successive rounds of Steepest descent and Adaptive-basis Newton-Raphson energy minimization with harmonic restraints on the α-carbons to prevent the structure from deviating significantly from the experimentally determined crystal structure. The stiffness of a restraints were defined on a per atom basis, with the initial restraint stiffnesses inversely proportional to the temperature factor. After each round of minimization the stiffnesses of the restraints were reduced by 10% until to total energy from the harmonic restraints was less that 0.1 kcal/mol. Prior to computing the eigenvalues and eigenvectors we ran one more minimization to ensure that the root mean square energy gradient was less than 1×10^{-4} kcal/mol. To solve the eigenvalue problem describing the minimized structure we used the coarse-grained rotation-translation block (RTB) normal mode method [164, 106]. RTB assumes that the residues are rigid bodies with three rotational and three translational degrees of freedom. This approximation prohibits access to the highest frequency modes, however the low frequency modes are of interest since they determine most of the global motion. Importantly, although becoming less so with advances in computer technology, the RTB approximation drastically reduces the computational requirement for solving the eigenvalue problem. Instead of solving a $O(3N)^2$ eigenvalue problem the computer solves a $O(6R)^2$ eigenvalue problem, where N is the number of atoms in the protein and R is the number of residues. From the energy minimized configuration we computed $M = 6 + 3n_{ions} + n_\alpha$ normal modes using CHARMM, where n_{ions} is the number of ions in the system and n_α is the number of α-carbon atoms. The first term accounts for the rigid-body rotation and translation modes. The second accounts for the ions which, in the RTB model, do not have the rotational degrees of freedom. The last term ensures that, regardless of the size of the system, the frequencies will cover long and short timescale motions.

Quantification of correlated motion

From the solution to Eqn. 3.3 we obtain the equilibrium fluctuations of the atoms, $\mathbf{x}_i(t)$, or, in terms of normal mode indices, $\mathbf{x}_{ik}$. From the fluctuations we computed correlations between the α-carbon atoms of the proteins. The typical correlation metric is the Pearson correlation coefficient.

$$r_P\left[\mathbf{x}_i, \mathbf{x}_j\right] = \frac{\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle}{\sqrt{\langle \mathbf{x}_i^2 \rangle \langle \mathbf{x}_j^2 \rangle}} \tag{3.4}$$

where the angled brackets denote thermal average [26].

$$\langle \mathbf{x}_i \rangle = k_B T \sum_{k=7}^{3N} \frac{\mathbf{x}_{ik}}{\omega_k}$$

A significant limitation of the Pearson correlation coefficient is that it only captures colinear correlated motions (App. B). To avoid this limitation we use the generalized correlation coefficient [102], which is based on the mutual information between the α-carbon fluctuations of α-carbons i and j

$$\begin{aligned} I\left[\mathbf{x}_i, \mathbf{x}_j\right] &= H[\mathbf{x}_i] + H[\mathbf{x}_j] - H[\mathbf{x}_i, \mathbf{x}_j] & (3.5) \\ r_I\left[\mathbf{x}_i, \mathbf{x}_j\right] &= \sqrt{1 - \exp\left(-\frac{2}{d} I\left[\mathbf{x}_i, \mathbf{x}_j\right]\right)} & (3.6) \end{aligned}$$

Assuming that the fluctuations follow a Gaussian distribution, i.e. that the energy landscape is locally harmonic, the joint distribution of $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$P\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{1}{\left|2\pi\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_j\right) \qquad (3.7)$$

In this limiting case the entropy is analytic [7].

$$H_{Gaussian}[\mathbf{x}_i] = \frac{d}{2}\left[1 + \ln 2\pi + \ln|\Sigma|\right] \qquad (3.8)$$

When the joint distribution is a Gaussian we denote the generalized correlation coefficient r_{LMI}.

Computing the generalized mutual information between two atoms requires the joint and marginal covariance matrices. The covariance between atoms i and j, Σ_{ij}, is 6-by-6 block matrix

$$\Sigma = \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix} \qquad (3.9)$$

where Σ_{ij} is the 3-by-3 covariance matrix corresponding to atoms i and j. We estimate the covariance matrix from the fluctuation vectors $\mathbf{x}_{ik}$ derived from normal mode analysis [26].

$$\begin{aligned} \Sigma_{ij} &\approx \left\langle \mathbf{x}_i \mathbf{x}_j^T \right\rangle / \left(\left\langle \mathbf{x}_i^2 \right\rangle \left\langle \mathbf{x}_j^2 \right\rangle\right)^{1/2} \\ &\approx k_B T \sum_{k=7}^{3N} \frac{\mathbf{x}_{ik}\mathbf{x}_{jk}^T}{\omega_k^2} \end{aligned}$$

where ω_k is the frequency of mode k in radians per time.

The correlation between α-carbons i and j quantifies the degree their fluctuations are coupled. The objective of the next section is to identify networks containing atoms that are more correlated with themselves than with atoms outside the network. We refer to such networks as either clusters, communities, or modules. In the next two sections we describe procedures that automatically identify modules from pairwise correlation metrics.

From the dynamical correlations we derive networks of correlated residues. Two approaches are mentioned in the literature for identifying networks of correlated amino acids in proteins, the clustering approach and the community detection approach. The clustering approach was applied to sequence data in work from R. Ranganathan's laboratory [109, 161]. In contrast to the clustering approach, the community detection approach uses modularity maximization to directly compute an optimal partition of the data. An example of applying the community detection approach is del Sol *et al.*'s work [156]. In the following section we will describe the application of the clustering approach and then the community detection approach to dynamical data computed via normal mode analysis.

Identifying correlated networks by clustering

The clustering approach compiles residues from a protein into a network by running a clustering algorithm on an α-carbon correlation matrix. In general a clustering algorithm outputs one or more ways to partition data points based on a distance metric that quantifies how different the objects are from each other. For our purposes, we cluster using a distance metric based on mutual information, $d_{ij} = 1 - r_{LMI}[\mathbf{x}_i, \mathbf{x}_j]$, which is zero for atoms that are perfectly correlated and unity for atoms that are uncorrelated.

Countless algorithms operate on a distance metric and output a clustering solution. To identify correlated dynamical networks in proteins we apply agglomerative (bottom-up) clustering and K-means (dispersion minimizing) clustering. Agglomerative clustering algorithms includes single-link, complete-link, average-link (UPGMA), or weighted-link (WUPGMA). Of course, different algorithms produce different results. Single-link and complete-link are essentially opposites, while UPGMA and WUPGMA represent a compromise between the single and complete linkage extremes. Single-link performs poorly because it violates the compactness expectation of a clustering solution, meaning clusters from single-linkage analysis will contain observations that are far apart according to the distance metric. Conversely, complete-link tends to violate the closeness expectation, meaning members of a cluster can be more similar to members of another cluster than to members of their own. Mathematically, the mean distance between two clusters partitioned by single or complete link goes to 0 or infinity, respectively, as the number of samples $N \to \infty$ [67]. Average-link and WUPGMA represent a compromise between single- and complete-link. Depending on the distribution of pairwise distances, average-link or weighted clustering may more closely resemble either single or complete-linkage results. Lastly, k-means clustering attempts to minimize the distance between the center of a cluster and all of the points that belong to the cluster. K-means tends to generate clusters with members roughly equidistant from their center.

All of the clustering algorithms we implement require a user-specified parameter that sets the number of clusters. Ideally one chooses the parameter to optimize one or more functions that quantify the quality of the clustering solution. To distinguish the quality of the clustering solutions an objective function is evaluated, thereby permitting selection of the best solution from those that are available. Handl *et al.* review many internal validation metrics useful for determine the appropriate number of clusters [66]. We optimize two opposing internal validation measures: the intracluster variance and the connectivity (Fig. 3-1).

The intracluster variance is a measure of how far the data within a cluster are from the cluster's centroid. Mathematically, the intracluster variance of a partition of a dataset, $V(P)$, is defined

$$V(P) = \sqrt{\frac{1}{N} \sum_{P_k \in P} \sum_{i \in P_k} (\mathbf{x}_i - \langle \mathbf{x}_k \rangle)^2} \qquad (3.10)$$

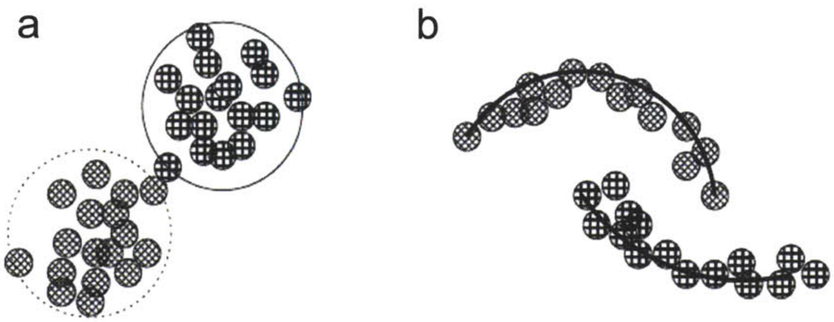

Figure 3-1: Example datasets that are compact (a) or connected (b). Adapted from [66]

where $\mathbf{x}_i$ is the coordinate vector of data point i. (We discuss the mathematics of converting a correlation matrix into coordinates later in the chapter.) The intracluster variance is positive and should be minimized.

The connectivity of a clustering solution is a measure of overlap between separate clusters. The connectivity of a partition of a dataset $Conn(P)$ is defined

$$Conn\,(P) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{L} w_{i,nn_{i(j)}} \tag{3.11}$$

where

$$w_{i,nn_{i(j)}} = \left\{ \begin{array}{ll} \frac{1}{j} & \text{if } P_k : i \in P_k \wedge nn_{i(j)} \in P_k \\ 0 & \text{otherwise} \end{array} \right\}$$

In words, the connectivity is a penalty that accumulates whenever L^{th} nearest neighbors do not belong to the same cluster. The connectivity is positive should be minimized.

The variance metric requires that the data points to be clustered have coordinates from which the distance to the centroid can be computed. However, the correlations we use to perform the clustering are a direct measure of distance, $d_{ij} = 1 - r_{LMI}[\mathbf{x}_i, \mathbf{x}_j]$, and so the variance as given in Eqn. 3.10 can not be computed directly. We therefore borrow an operation from spectral clustering to convert distances into coordinates [43]. From the correlations matrix $[\mathbf{C}]_{ij} = r_{LMI}[\mathbf{x}_i, \mathbf{x}_j]$ we compute the normalized

Laplacian matrix, $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{C}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{D}$ is a diagonal matrix containing the sum of the rows (or columns) of the correlation matrix $\mathbf{C}$, i.e. $[\mathbf{D}]_{ij} = \left(\sum_k C_{ik}\right)\delta_{ij}$. The eigenvectors of $\mathbf{L}$ constitute a basis set that serves as coordinates of the sample points in $\mathbb{R}^k$. For every clustering solution P we compute the variance $V(P)$ from the first $k = 3$ eigenvectors of $\mathbf{L}$ and identify solutions that minimizes both the variance and connectivity. Note that in all of the cases we studied the first three eigenvalues accounted for more than 95% of the total eigenspectrum.

As validation for our implementation Fig. B-1 demonstrates the use of internal cluster validation metrics on Golub's cancer cell transcription profiles [66]. As the number of clusters increases, the variance decreases while the connectivity increases. The 3rd data point corresponds to a partition of the data into three clusters, which matches the number of cancer types in the data set.

In general one cannot decrease the variance without increasing the connectivity. This poses the problem of how to choose the appropriate number of clusters. To deal with this problem we identify a set of solutions that are Pareto optimal in the variance-connectivity space. A partition P^* is Pareto optimal if there is not another partition P such that $V(P) \leq V(P^*)$ and $C(P) \leq C(P^*)$, and either $V(P) < V(P^*)$, $C(P) < C(P^*)$, or both. In our analysis we identify all clustering solutions that are Pareto optimal.

The clustering approach we described is not designed to give a good solution in general. Rather, it is a means of choosing the best solution from the set of solutions that are available. If all of the clustering solutions are poor, the Pareto optimal solution will be poor. An alternative approach for identifying allosteric coupling was therefore sought.

Identifying correlated networks by community detection

Community detection algorithms directly optimizes an intuitive description of a correlated network. Informally, a community is a collection of objects with dense intra-community connections and sparse intercommunity. Similarly, we define a network of correlated residues in a protein as α-carbons with strongly correlated motion that is distinct from other parts of the protein. One represents a community by a weighted and undirected graph, where the members of the communities are called nodes, the connections between members are edges, and the community to which a member belongs is a node label. For our purposes the nodes are the α-carbons in a protein and the edges are the pairwise correlations between the atomic fluctuations, $r_{LMI}[\mathbf{x}_i, \mathbf{x}_j]$. We apply a community detection algorithm to a graphical representation of a protein's motion, thereby assigning the α-carbons in the protein to correlated networks.

Mathematically, one identifies communities by minimizing an energy function

$$H\left(\{\sigma\}\right) = -\sum_{i \neq j} J_{ij}\delta\left(\sigma_i, \sigma_j\right) \tag{3.12}$$

where J_{ij} is energy associated with assigning nodes i and j to the same community, σ_i is the community label of node i, and $\delta(x, y)$ is unity when $x = y$ and otherwise 0. In modularity maximization $J_{ij} = w_{ij} - \gamma \left\langle w \right\rangle_{ij}$, where w_{ij} is the affinity of node i for node j, $\left\langle w \right\rangle_{ij}$ is the affinity under an appropriate null model, and γ is a free parameter. The configurational model is a common null model for community detection [120, 143].

$$\left\langle w \right\rangle_{ij} = \frac{(\sum_i w_{ij})(\sum_j w_{ij})}{2\left(\sum_i \sum_j w_{ij}\right)}$$

The configuration model has the same form as the average-product correction used to remove the effect of phylogeny in coevolution analysis (Chap. 2). A clever clustering algorithm dubbed Superparametric Clustering (SPC) [21] uses the same form of Hamiltonian as Eqn. 3.12, except the authors define the interaction energy as

$$J_{ij} = J_{ji} = \frac{1}{Z}\exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right)$$

where Z is a normalization constant and σ is a length scale. Note that null model in SPC is $\langle w_{ij} \rangle = 0$, meaning nodes are not expected to have connections outside of their communities.

Most published work on community detection has focused on identifying communities in prototypical networks, such as the Internet, World Wide Web, social networking websites, journal citations, protein-protein interactions, or synthetic networks [99]. Such networks are canonically large, sparse, sometimes directed, and often scale free [14]. In contrast, the networks obtained from fluctuations of α-carbons in proteins are small, dense, undirected and weighted. To our knowledge, relatively little work focuses on dense weighted networks, with the exception of Heimo *et al.* who used a Potts model of community detection to identify structure in a stock market dataset [70].

We applied Heimo *et al.*'s weighted network modularity optimization formulation to protein dynamics data. For every protein we generate a graph where the nodes are α-carbons and the weighted edges are the r_{LMI} between pairs of α-carbons. To reduce the computational burden we make the network sparse by searching for communities in the graph's Maximum Spanning Tree (MST).

To minimize the Hamiltonian we use a greed stochastic search algorithm that follows from the description in [143]. Briefly, we loop through the nodes in the graph in random sequential order. For each node we identify the neighbor node from a different community (if any) that decreases the Hamiltonian the most, and then move the node to the new community. If we define the affinity matrix $w_{ij} = r_{LMI}[\mathbf{x}_i, \mathbf{x}_j]$, then the update of the Hamiltonian for the move of node l from community ϕ to community α is

$$\begin{aligned} \Delta H &= \sum_{i \neq l} w_{lj} \delta(\phi, \sigma_j) - \gamma \frac{s_l}{2S} (S_\phi - s_l) \\ &- \sum_{i \neq l} w_{lj} \delta(\alpha, \sigma_j) + \gamma \frac{s_l}{2S} S_\alpha \end{aligned}$$

where s_l is sum of the edge weights incident upon node l, S_ϕ is the sum of s_l for all nodes in community ϕ, and S is the sum of all the edge weights in the graph. If $\Delta H > 0$ for a test node the algorithm leaves the node in its current community, otherwise it moves the node to the new community. After looping through all of the nodes, the algorithm then loops through all of the communities and computes the greatest energy decrease that would result from merging two communities. If the largest decrease is positive we do not merge the communities. We repeat looping through the nodes and then communities until we can no longer decrease the energy of the Hamiltonian. Although efficient, this procedure is not guaranteed to find a global minimum. We therefore run the procedure three times and report the lowest energy community. To examine correlated networks at different resolutions we vary the resolution-controlling parameter $\gamma > 0$ and map the resultant community onto the 3D protein structures.

We validated our implementation of the community detection algorithm using a simple dense weighted graph from the original authors [70]. The graph comprises N_b blocks each with N_c nodes. All of the nodes are connected by weighted edges. The edge weight between nodes within a block is $w_i = 1$ while the edge weight between nodes in different blocks is $w_b = 0.1$. By varying the resolution parameter, γ, different community structures emerge. For this network we find one community when $\gamma = 0.3$ and four communities when $\gamma = 1.5$, (Fig. B-2). Our results agree with the reported results [70].

3.3 Results and discussion

The algorithms just described can be applied to any protein, but before attempting to learn something new, it is valuable to test the approaches on a few well-studied example proteins. We therefore implemented both approaches to characterize the conformational dynamics of two proteins that show evidence for allosteric coupling, fascin and rhodopsin, with results for hemoglobin and β-trypsin provided in the Appendix B). We also applied the approaches to cadherin dimers to gain further insight into the structure-function relationship we studied in Chapter 2.

Fascin

Fascin is a 55 kDa globular protein that bundles filamentous actin (F-actin) [180]. The structure of *H. sapiens* fascin-1 comprises four beta-trefoil domains, F1-F4, that constitute two lobes, F1-F2 and F3-F4 [148]. The C-terminal half of fascin contains as least one F-actin binding site [126], while another F-actin binding site is thought to exist near the N-terminal half [126, 148]. Conformation dynamics suggests that the subdomains F1 and F3, although not in contact, are allosterically coupled [148]. We applied both the clustering and community detection approaches to fascin conformational dynamics computed from the crystal structure (PDB ID 1DFC) as a means of validating our algorithms and comparing the two approaches.

Fascin's dynamical correlation map has a block diagonal structure (Fig. 3-2A), showing that the residues within subdomains are more tightly coupled to each other than to residues in different subdomains. Although the correlation map shows coupling between subdomains F1 and F3, it is not clear from the map alone what the strength of the coupling is. Hierarchical clustering by average linkage and WPGMA clustering produce Pareto optimal clustering solutions (Fig. 3-2B) that elucidate the structure and strength of the allosteric coupling. Mapping the clustering solutions onto the protein structural model using colors to distinguish the different communities reveals F1 and F3 are allosterically coupled (Fig. 3-2C), which agrees with the published results [148] and thereby validates our implementation.

Community detection reveals correlated networks that are distinct from the clustering method solutions in several ways (Fig. 3-3). First, hierarchical clustering produces one solution for a specified number of clusters while community detection can identify multiple solutions with the same number of clusters. In this way community detection provides richer insight into the structure of correlated protein motion. Second, the networks derived from clustering do not correspond to the networks derived from community detection. For example, the Pareto optimal solution assembled by average-link clustering consistently places subdomains F1-F3 within the same cluster (not shown). Community detection, in contrast, finds a 2-cluster partition with F1

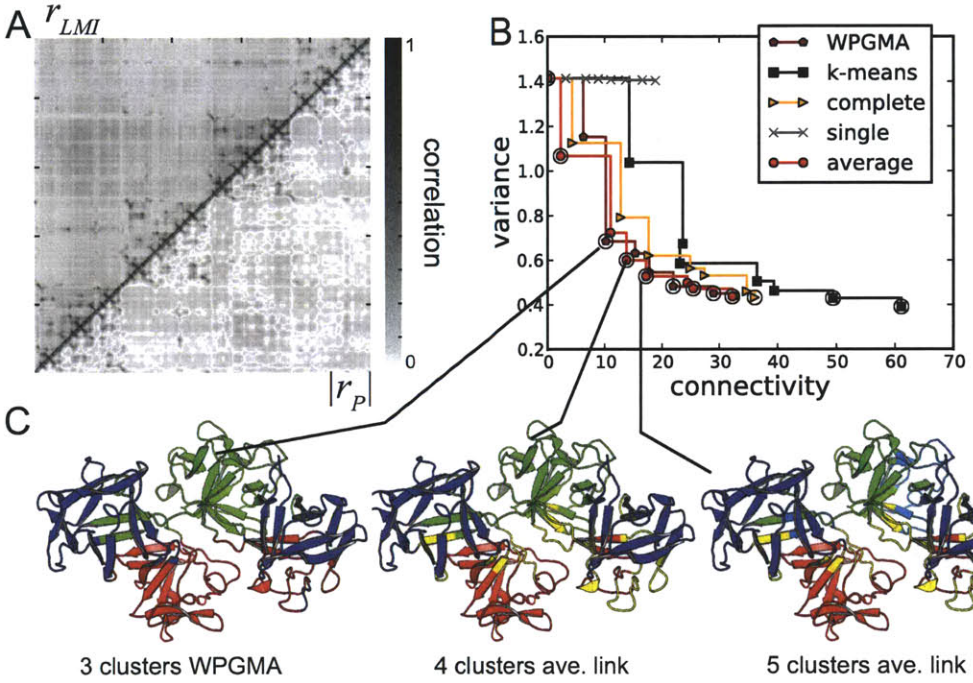

Figure 3-2: Validation of cluster analysis procedure on the protein fascin. The upper triangle of the pairwise correlation plot equals r_{LMI} and lower triangle is the magnitude of the Pearson correlation coefficient $|r_P|$. The Pareto optimal clustering procedure identifies allosteric coupling between subdomains F1 and F3. In the structure regions with the same color belong to the same cluster.

and F2 (i.e. lobe 1) in a community and F3 and F4 (lobe 2) in another community. Also unlike clustering, community detection resolves a 4-cluster solution with F1 and F3 in separate domains, whereas the Pareto optimal 4-cluster solution from the clustering approach puts F1 and F3 in the same cluster.

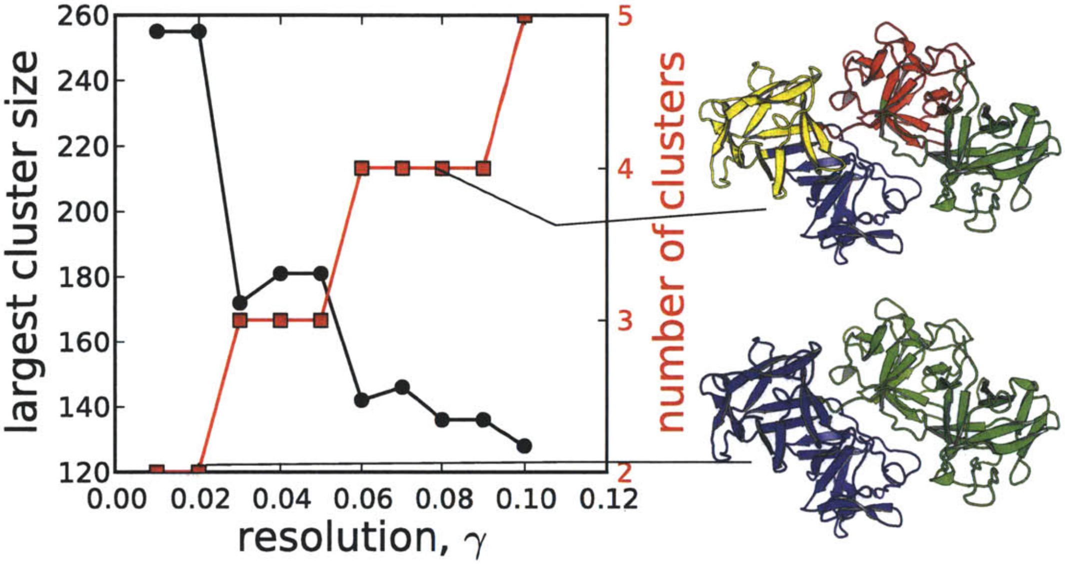

Figure 3-3: Correlated dynamical networks in fascin computed by community detection. The largest cluster size (black circles) and number of cluster (red squares) are plotted at resolution ranging from 0.01 to 0.1. The structures depict which communities the α-carbons belong to. At $\gamma = 0.02$ subdomains F1 and F2 are in the green network while F3 and F4 are in the blue network. At $\gamma = 0.08$ the subdomains are their own networks.

Bovine rhodopsin

G-protein coupled receptors (GPCRs) are a large gene family of transmembrane signaling proteins. Upon stimulation GPCRs transmit signals across a membrane by switching between quiescent and signaling states. Rhodopsin (PDB ID 1HZX) signals via a conformational change that converts a photon into a biochemical signal in rod cells of the retina. The conformational change that rhodopsin and other GPCRs undergo to transmits signals is of interest for both developing therapeutics and understanding the five senses.

Clustering analysis reveals a membrane-spanning correlated α-carbon network

(Fig. 3-4). Helix VIII, which run parallel to the cytoplasmic side of the cell membrane and orthogonal to the central helix bundle, correlates with the extracellular portions of helices I and II. The placement of the disconnected helix fragments within the same cluster exists in several solutions, suggesting the long-range coupling contributes significantly to the global motion of the protein. To our knowledge, this clustering solution provides the first evidence of long-range allosteric coupling between the inside and outside of cells through GPCRs.

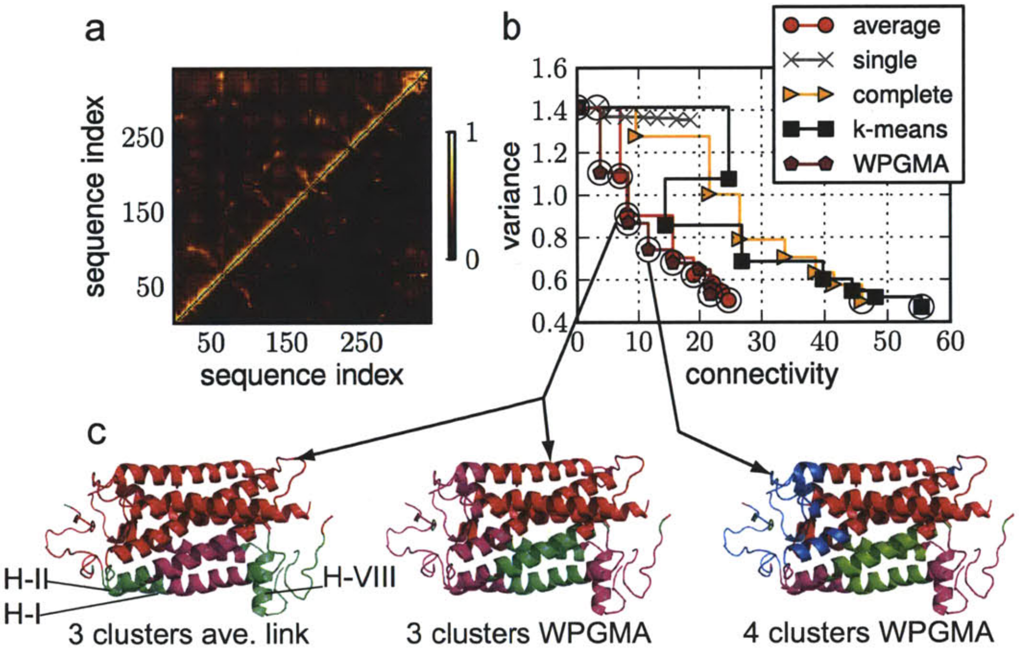

Figure 3-4: Pareto-optimal dynamics-based allosteric network in Bovine rhodopsin. The α-carbon fluctuation correlation matrix is shown in (a). Clustering solutions were plotted on a space representing cluster separation and compactness (b). Three solutions on the Pareto front (c).

Our greedy search algorithm converges to two stable solutions at resolutions between $\gamma = 0.01$ and $\gamma = 0.05$ (Fig. 3-5a) that are inconsistent with the clustering solutions. Unlike the clustering solutions, none of the α-helices partition into more than one network (Fig. 3-5b-c), suggesting that the fluctuations of the atoms within an α-helix are highly correlated. The clustering solution also suggests that cytoplasmic helix VIII is coupled to the extracellular portion by long-range communication with helices I and II (Fig. 3-4c), while community detection identifies cytoplasmic-extracellular coupling via coordination of helix VIII with helices V, VI, and VII (blue and brown in Fig. 3-5c).

The rhodopsin example demonstrates that community detection generates sets of solutions that hierarchical clustering can not, i.e. solutions that are not hierarchical. At low resolution helix V is coupled with helices VI-VIII (Fig.3-5b), while at a higher resolution V is coupled with helices III and IV (Fig.3-5c). There is no agglomerative operation that can generate the low resolution solution from the high resolution solution, therefore the communities are not hierarchical.

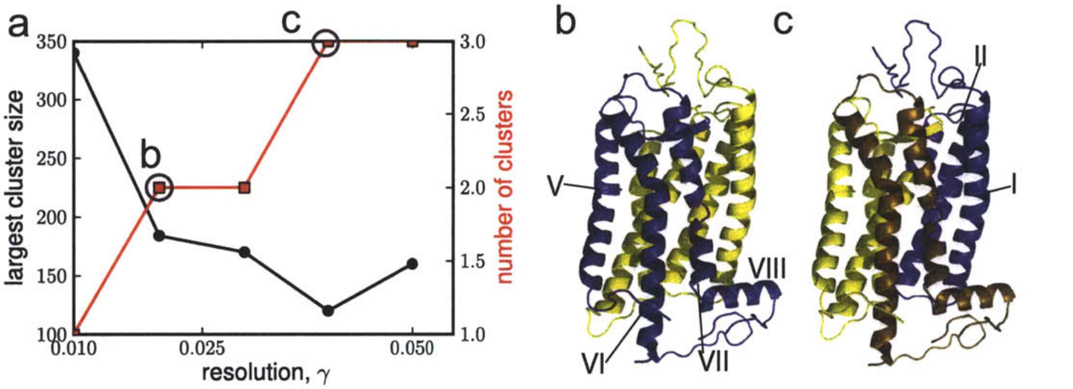

Figure 3-5: Correlated dynamical networks in Bovine rhodopsin inferred by community detection. The number of communities increases with the resolution parameter γ (a). The circled points in (a) correspond to the community structure shown in (b, c).

Interestingly, the putative coupling mechanisms we identify have not been reported in the literature, at least not to our knowledge. Signal transduction in rhodopsin is thought to occur through the coordinated relaxation of the helix triad comprising helices III, VI, and VII, and is triggered by a perturbation to residue 296 in helix VII [38]. In other GPCRs, the N-terminal loop covers the helix bundle like a lid, and may modulate the quiescent-signaling transition in some GPCR classes. For instance, mutations in the N-terminus of opioid receptors enhance spontaneous signaling activity [38], and that signaling terminates at the transducin binding site on the C-terminus [170]. We therefore hypothesize that either the N-terminus communicates with the C-terminus via long-range coupling (Fig. 3-4) or via coordinated movements of helices V, VI, and VII (Fig.3-5). Testing this hypothesis would require site-directed mutagenesis experiments, and may only apply to a fraction of the ~600 GPCRs.

Cadherin strand-swap dimers

In Chapter 2 we inferred a set of coevolving amino acid pairs in the cadherin ECB domain. About half of the residue pairs we identified were curiously not in contact, suggesting that the coevolving pairs may be allosterically coupled. Here we applied the tools developed in this chapter to gain insight into the conformation dynamics and potential allosteric coupling in cadherin strand-swap dimers.

Figure 3-6 shows 3-cluster Pareto optimal solutions for Type I and II cadherin strand-swap dimers. Average-link hierarchical clustering generated the Pareto front, so just those results are shown. The Type I cadherin dimer optimally partitions into three clusters wherein the two Ca^{2+} binding sites are in separate clusters and the EC1 and EC2 domains belong to the same cluster. In contrast to the Type I cadherin clustering solution, the Type II cadherin dimer clustering solution demonstrates coupling between the calcium binding domains and the amino-terminal portion of EC1. The analysis suggests that the amino-terminal portion of EC1 is more tightly coupled to the calcium binding regions in Type II strand-swap dimers than in Type I strand-swap dimers. The differences in coupling between Type I and II dimers may manifest as differences in rates of calcium-induced activation and binding of the Type I and Type II EC domains [157].

We applied the community detection approach to Type I and Type II cadherin strand-swap dimer conformational dynamics and identified two stable solutions and a transition solution within the resolution range $0.01 \leq \gamma \leq 0.09$ (Fig. 3-7a and e). The lowest resolution solution identified two correlated networks; one network corresponds to each cadherin protomer (Fig. 3-7b and f). Three correlated networks were identified at the intermediate resolution (Fig. 3-7c and g). One of the intermediate resolution networks is a combination of the EC1 domains from the two protomers, while the other two are the EC2 domains from the two protomers. At the highest resolution each EC domain is a network and the calcium binding domains are shared between

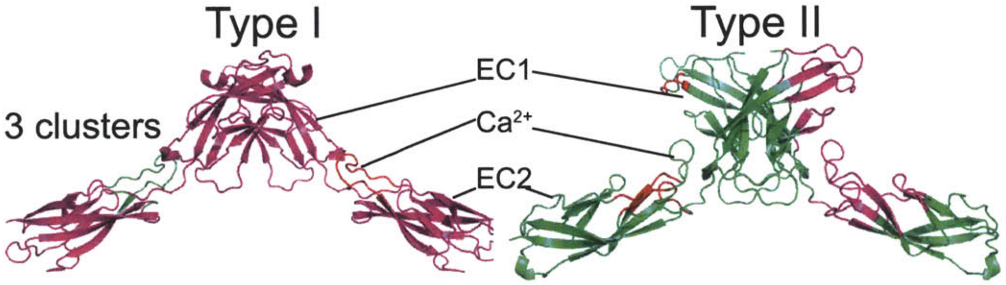

Figure 3-6: Pareto-optimal dynamics-based allosteric network in cadherin strand-swap dimers. Two- (top row) and three-cluster (bottom row) solutions are shown for Type I cadherin *H. sapiens* cdh1 (PDB ID 2O72) and *M. musculus* cdh11 (PDB ID 2A4E). The color coding in red, green, or magenta represents membership in predicted correlated networks.

EC1 and EC2 of the protomer to which they belong (Fig. 3-7d and h).

The networks computed by community detection at all resolutions, γ, are spatially compact, i.e. the networks comprise protomers, interacting EC1 domains, or domains within protomers. This result indicates three points about cadherin strand-swap dimers: coupling within a protomer is greater than between protomers; coupling between binding site residues in EC1 is greater than between binding site residues in the rest of EC1 or any part of EC2; and that coupling within EC domains is greater than coupling between EC domains.

Interestingly, the calcium binding EC1-EC2 linker domain clustered with EC2 at all resolutions and in both cadherin models. Experiments show that the calcium binding region between EC1 and EC2 destabilizes an EC2 fragment, but that the EC2 domain with the linker is partially rescued with the addition of calcium [139]. That the linker region affects the stability of EC2 supports our prediction that EC2 is dynamically coupled to the linker.

The stable Type I and Type II cadherin networks solved by the community detection algorithm are topologically similar at all resolution levels. On one hand, the similarity is reassuring given that the molecules are structurally similar and our correlations come from low-energy global motion predicted by normal mode analysis. On the other hand, the similarity also suggests the community detection approach may not be sensitive enough to distinguish different types of allosteric communication in

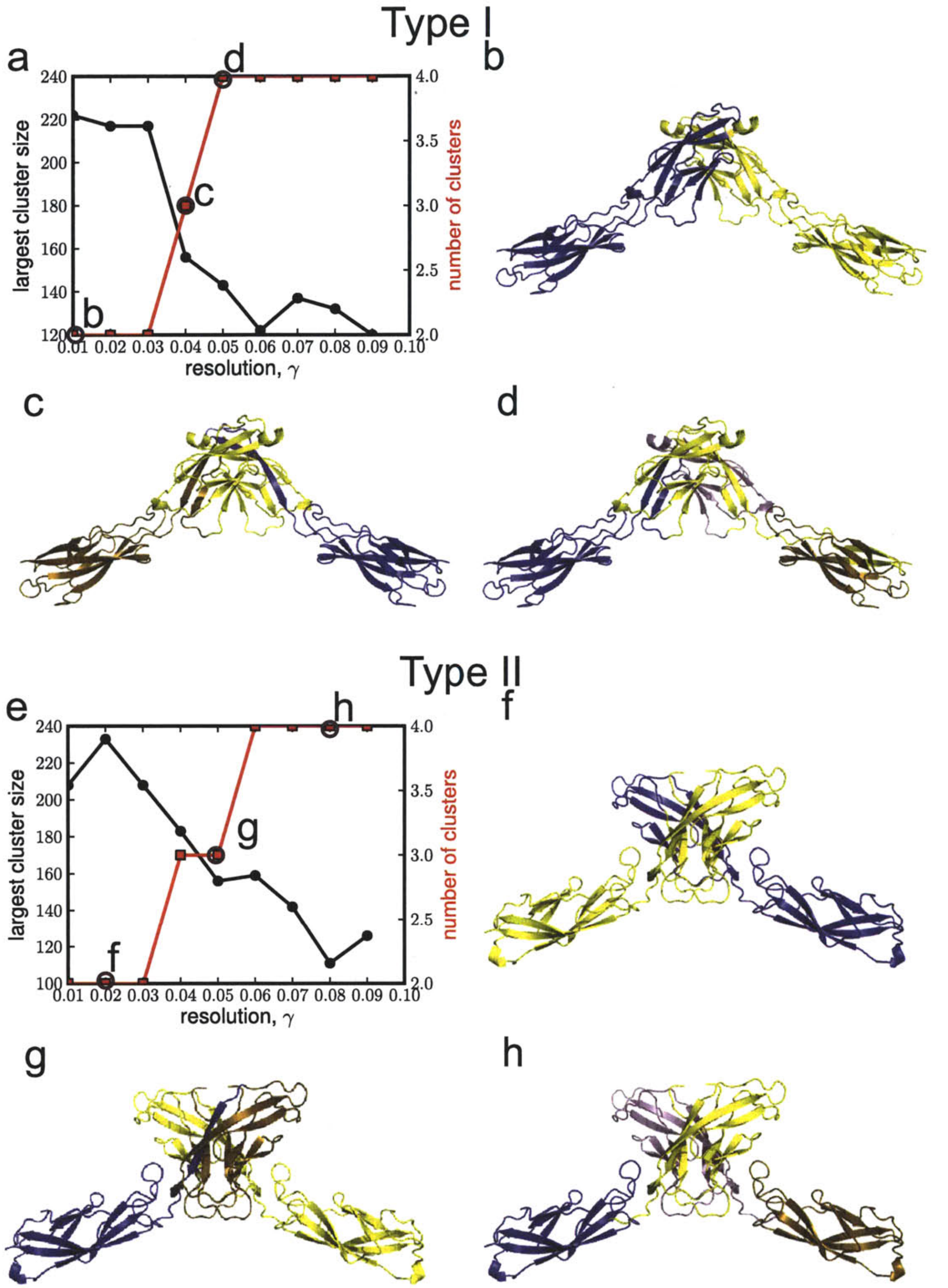

Figure 3-7: Correlated dynamical networks in cadherin strand-swap dimers detected by community detection. Three stable solutions occur in the range $0.01 \leq \gamma \leq 0.09$ (a, e): communities of 2 (b, f), 3 (c, g) or 4 (d, h) are shown on structures of cadherin strand-swap dimers. The color coding in yellow, blue, brown, or mauve represents membership in predicted correlated networks.

similar molecules. One could formally characterize the sensitivity by comparing the networks of paralogs, orthologs, mutants, or conformers, and perhaps optimize the algorithm or data collection procedure to improve the sensitivity.

3.4 Concluding remarks

For all of the proteins we examined for correlated networks, the Pareto optimal clustering solutions do not resemble the solutions generated by community detection. While clustering produces spatially-disconnected correlated clusters, community detection produces spatially compact clusters. It is therefore reasonable to ask which method is superior. The clustering approach attempts to find the most compact clusters from a set of precomputed solutions. The algorithms that produce the clustering solutions optimize different characteristics of the data structure that depend on the dissimilarity measure, and no single approach is guaranteed to find a solution that minimizes both objective functions $V(P)$ and $Conn(P)$. In contrast, the community detection approach has a well-defined and intuitive objective function to optimize. Moreover, the clustering approach can only generate a finite and discrete set of solutions, while the community detection approach has the power to generate a solution spectrum. We therefore recommend the community detection algorithm as the more principled method for detecting correlated networks. A quantitative comparison is necessary to fully justify one approach over the other.

We foresee several applications of protein correlated network detection. One application is the rational design of enzyme inhibitors. By predicting sites on an enzyme that are dynamically coupled to the active site one can, in principle, design molecules that allosterically inhibit ligand binding, catalysis, or both. A second application is in the novel design of biological macromolecules. The community detection approach suggests that proteins can be partitioned into modules of dynamically correlated atoms, a notion explored previously for protein crystal structures [156]. The modules may serve as building blocks for engineering novel chimeric proteins with designed function. The final proposed application is for the basic understanding of protein dynamics. With a database of structures, such as the Protein Databank, the conformation dynamics can be computed via normal mode analysis or molecular dynamics and subsequently interpreted with the clustering or community detections approaches. Conveniently, conformational dynamics databases already exist for both proteins [116] and protein complexes [92], so implementing a large-scale survey only requires porting data from the databases into a format compatible with clustering or community detection computer programs.

Chapter 4

Consequences of allosteric coupling between tandem binding domains in F-actin bundling proteins

Abstract

Eukaryotic cells construct mechanical organelles from ordered bundles of F-actin and an assortment of actin bundling proteins (ABPs). A complete understanding of the structure and behavior of bundles requires physical insight into the ABPs that cross-link bundled F-actin. Our current understanding of ABPs includes the hypothesis that the tandem F-actin binding sites bind to F-actin cooperatively. Unfortunately, the means of testing for cooperativity in a protein are still somewhat primitive (Chap. 3). We therefore ask what are the consequences of allosteric coupling on the structure of F-actin bundles, a quality that can be observed and quantified experimentally. Using a mathematical model of ABPs binding to transversely ordered F-actin, we study the dependence of the bundle's structure on strength of cooperative cross-linking. Our analysis shows that coopertivity provides competitive advantage that favors cross-links occupying F-actin binding sites instead of ABPs that bind without cross-linking. We interpret our result with a new hypothesis that nature ought to select for cooperativity in ABPs, and therefore that cooperativity is a general feature

of ABPs.

4.1 Introduction

Filamentous actin (F-actin) is a biological polymer capable of forming a variety of complex structures in cells, from tangled networks to ordered bundles. F-actin bundles in particular serve as the building-blocks for a variety of organelles, including but not limited to filopodia, microvilli, the contractile ring in dividing cells, stress fibers, and structures inside neurosensory cells that detect pressure waves, gravity, or other mechanical stimuli (Reviewed in [42]). F-actin bundles also constitute an integral part of the contractile machinery in smooth and striated muscle. Because of the ubiquitous nature of F-actin bundles in biology and physiology, the physics underlying regulation of bundle structure is of considerable interest.

Many ligands can drive F-actin bundle formation, including cations and basic peptides [165], “inert” molecules like poly-ethylene-glycol (PEG) [78], and so-called actin bundling proteins (ABPs, reviewed in [15]). At sufficiently high F-actin concentrations entropy can drive bundling as well (reviewed in [71]). The bundling agents mediates the structure and mechanics of both F-actin and the composite bundle [23, 10, 9, 152, 153, 17, 35, 36]. Of particular interest here are the ABPs that generate transverse hexagonal F-actin bundles (Fig. 4-1). How the conformational dynamics of the bundling protein might influence the structure of the F-actin bundle is an open question we attempt to address.

Structural data and conformational dynamics analysis provide evidence that actin bundling proteins exhibit allosteric coupling between their tandem actin binding domains. One example is the protein fascin, studied in Chap. 3. Fascin contains two lobes. The C-terminal lobe has a known F-actin binding site while the N-terminal lobe has a putative F-actin binding site [126]. Conformational dynamics analysis identified intriguing coupling between the subdomains containing the binding site, suggesting that binding to F-actin may be cooperative [148]. Fimbrin/plastin is another ABP with tandem F-actin binding domains. The evidence for cooperative

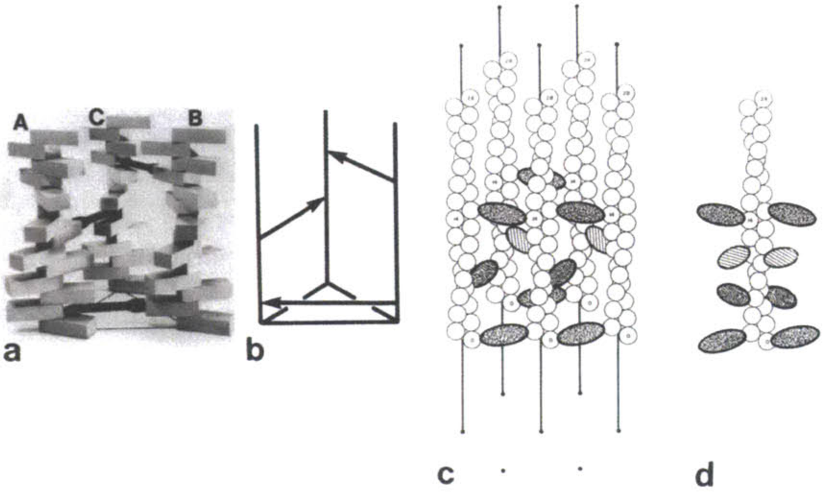

Figure 4-1: Actin bundling proteins organizes F-actin into transverse hexagonal bundles. Balsa wood model of a hexagonal F-actin bundle with ABP cross-bridges (a). Schematic of ABP connectivity (b). Schematic of five F-actins cross-linked by ABPs (c). Single F-actin from (c) with ABPs attached at cross-linking positions (d). Figure adapted from [42].

binding by fimbrin is the polymorhpic nature of the calponin homology domain 2 CH2 [95], which contains one of the F-actin binding sites. What affect, if any, cooperative cross-linking has the the structure of F-actin bundles is an open question. By understanding the consequences of cooperative cross-linking, we may design experiments that test whether binding to F-actin is cooperative or not for any given ABP, including fascin and fimbrin.

The new contribution of this chapter is attention to the consequences of cooperative cross-linking on the structure of the bundle. Although a number of studies have examined cooperative adsorption of proteins to F-actin [74, 75, 155], or F-actin bundles [152, 153] the nature of the cooperativity differed from the present study: the previous work modeled cooperativity between binding sites on F-actin, while this study addresses cooperativity between ABP binding sites.

This chapter describes the predictions of a simple mathematical model that incorporates the physics of allosteric communication between the F-actin binding domains in an F-actin bundling protein. We formulated two versions of the model, one more parsimonious than the other, and compared their predictions. Using both models we examined the effect of cooperative cross-linking on the F-actin bundle's structure. We conclude with application of the model for identifying allostery in ABPs F-actin pelleting assays.

4.2 Methods

We utilize a simple model of adsorption to a 1-dimensional lattice to represent ABPs binding to parallel F-actins (Fig. 4-2). The lattice consists N rows each containing a pair of adjacent F-actin binding sites. In this simple formulation, we incorporate neither cooperation among F-actin binding sites nor deformation of F-actin do to cross-linking. We therefore need not consider the transverse spacing of the binding site, which so happens to be irregular [42]. The rows of the lattice can occupy any of five states. The energies of the states are

$$\begin{aligned}
E_U &= 0 \\
E_L &= \epsilon - \mu \\
E_R &= \epsilon - \mu \\
E_B &= 2\epsilon - 2\mu \\
E_C &= 2\epsilon - \mu + \sigma
\end{aligned}$$

where every term is normalized to thermal energy. A row is in its reference state E_U when occupied by solvent. If one of the two adjacent F-actin binding sites is occupied by the 'left' domain of an ABP, the energy, E_L, is the sum of an enthalpic term ϵ and an entropic penalty μ from removing the ABP from solution and immobilizing it on F-actin. We assume the binding sites for the 'right' domain of the ABP contributes the same energy when occupied, E_R, and therefore the singly-occupied state is degenerate. The assumption that the actin binding domains exhibit the same affinity for F-actin is not strictly correct. For example the apparent dissociation constants of fimbrin's actin binding domains 1 and 2 are 0.34 ± 0.04 and 2.6 ± 0.3 μM, respectively [104]. A more general model could trivially incorporate distinct binding affinities. The third state, with energy E_B, is a doubly-occupied state. In the doubly-occupied state both of the F-actin binding sites in a row are occupied by a domain of two different proteins. We do not know if the doubly-bound state is ever realized *in vitro* or *in vivo*, so we define a 4-state model without the doubly-bound state and a 5-state model with it in an effort to understand the consequences of the different assumptions. The final state is the cross-link state, where one ABP binds both of the adjacent F-actin sites in a row to form a cross-link. In the cross-linked state both binding sites are occupied, which contributes an enthalpic term, 2ϵ, and an entropic penalty, μ. An additional term, σ, accounts for potential allosteric communication between ABP domains, manifest in this model as cooperative cross-linking.

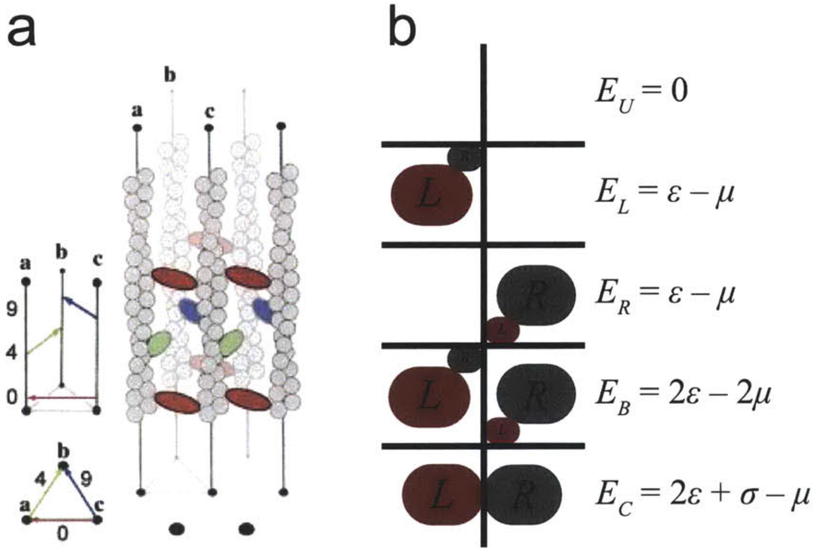
a
b
a
b
c
a
b
c
9
4
0
b
4
9
a
c
0
L
R
$E_U = 0$
$E_L = \varepsilon - \mu$
$E_R = \varepsilon - \mu$
$E_B = 2\varepsilon - 2\mu$
$E_C = 2\varepsilon + \sigma - \mu$

Figure 4-2: Schematic of a mathematical model representing parallel actin filaments with adjacent binding sites for ABPs. Geometry of transverse F-factin bundles adapted from [177] (a). The lattice is represented as N rows of adjacent F-actin binding sites (b). The cells on the left correspond to one F-actin while those on the right correspond to the adjacent F-actin. An F-actin bundling protein with two binding domains is represented as two pill-shapes, where each binding domain is labeled a different color and designated with and L for left and R for right binding domain, respectively. The left domain binds on the left of side of the array while the right domain binds to the right side of the array. The N rows in the lattice can exist in one of five states, enumerated from top to bottom: unoccupied, bound by the left ABP domain, bound by the right ABP domain, bound by two different ABP, each contributing a domain, or cross-linked. The corresponding energies of each a state are shown to the right of the cartoon.

We are interested in two quantities that characterize the structure of the F-actin bundle: the number of F-actin binding rows that are blocked from acquiring a cross-link, θ, and the fraction of sites that are cross-linked, ρ. A blocked row has energy E_L, E_R, or E_B (Fig. 4-2) and cannot acquire a cross-link. The cross-link density is of interest from a biophysical perspective because cross-links function as glue that holds the filaments together and bare load [17, 36]. A row cannot both be blocked and cross-linked, therefore the cross-linking state must compete for rows to bind with all of the other states. Because of the physical importance of the cross-link density, we characterize the structure of the lattice by reporting the blocked row density, θ, and the cross-link density, ρ, as a function of ABP bulk concentration.

We solved the model analytically and validated it numerically using Markov Chain Monte Carlo integration with the Python module pymc. Although more complicated models that include nearest-neighbor cooperativity between F-actin binding sites in the N rows can also be solved analytically using the transfer matrix method [41, 24, 98, 135, 166, 167, 169, 168], we still chose to implement a numerical solution to simplify the task of extending the model to include long-range coupling between the F-actin binding sites. Such a complicated model has not been implemented as of yet.

4.3 Results and discussion

Analytic solution

When the rows of the lattice are independent of each other, the model admits a particularly simple solution that we explore here. The partition function for a row of the lattice is a summation of Boltzmann weight functions.

$$\begin{aligned} q_4 &= 1 + 2e^{-\epsilon+\mu} + e^{-2\epsilon-\sigma+\mu} \\ q_5 &= 1 + 2e^{-\epsilon+\mu} + e^{-2\epsilon-\sigma+\mu} + e^{-2\epsilon+2\mu} \end{aligned}$$

where q_4 and q_5 correspond to the 4- and 5-state models, respectively. Since the row partition functions are independent the composite partition function is simply $Q_s = q_s^N, s = 4, 5$. The fraction of blocked rows, θ_s, or cross-linked rows, ρ_s, is

$$\begin{aligned}
\theta_4 &= \frac{2e^{-\epsilon+\mu}}{1+2e^{-\epsilon+\mu}+e^{-2\epsilon-\sigma+\mu}} \\
\rho_4 &= \frac{e^{-2\epsilon-\sigma+\mu}}{1+2e^{-\epsilon+\mu}+e^{-2\epsilon-\sigma+\mu}} \\
\theta_5 &= \frac{2e^{-\epsilon+\mu}+e^{-2\epsilon+2\mu}}{1+2e^{-\epsilon+\mu}+e^{-2\epsilon-\sigma+\mu}+e^{-2\epsilon+2\mu}} \\
\rho_5 &= \frac{e^{-2\epsilon-\sigma+\mu}}{1+2e^{-\epsilon+\mu}+e^{-2\epsilon-\sigma+\mu}+e^{-2\epsilon+2\mu}}
\end{aligned}$$

By expressing the (dimensionless) ABP chemical potential as a function bulk concentration, $\mu = \mu_0 + \ln(C/C_0)$, we can express the structure of the F-actin binding sites as a function of the concentration of cross-linker in the surrounding environment relative to a reference concentration, C_0.

Comparison 4-state and 5-state titration curves

The key difference between the 4-state and 5-state model is that the 5-state model permits the doubly-bound state wherein two ABPs bind to adjacent F-actin binding sites at the same row of the bundle. The ABP titration curve demonstrates the consequence of permitting the doubly-bound state on the fraction of cross-linked F-actin sites (Fig. 4-3). At low effective fugacity, $z = e^{\mu} \to 0$, the blocked and cross-linked fractions both tend toward zero. As the bulk concentration of ABP increase ($\mu \to \infty$), the 4-state and 5-state cross-link fractions, ρ_4 and ρ_5, respectively, diverge. The cross-link density in the 4-state model asymptotes to non-zero value

$$\lim_{\mu\to\infty} \rho_4 = \frac{e^{-\epsilon-\sigma}}{2+e^{-\epsilon-\sigma}}$$

while in the 5-state model the cross-link density vanishes

$$\lim_{\mu \to \infty} \rho_5 = 0$$

and the fraction of blocked sites approaches unity

$$\lim_{\mu \to \infty} \theta_5 = 1$$

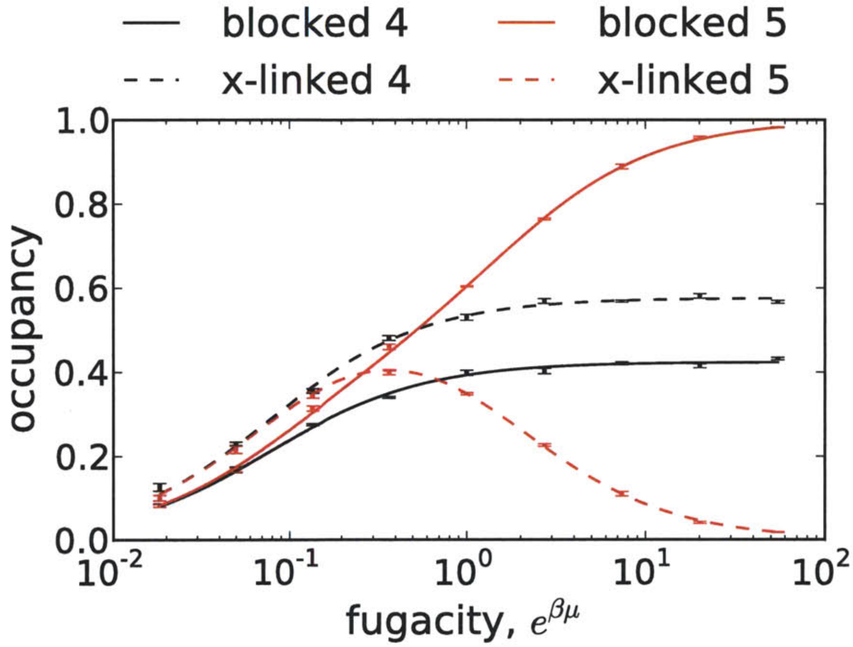

Figure 4-3: Titration curves demonstrating the difference between the 4-state and 5-state models. The fraction of blocked (solid lines) or cross-linked rows (dashed lines) are plotted against the effective fugacity, $z = e^{\mu}$ using both the 4-state (black lines) and 5-state (red lines) models. The lines are analytic while the errorbars are from simulation.

The behaviors of the 4- and 5- state models at high ABP concentrations diverge, so it is worthwhile considering the physical interpretations wherein one or the other model might match experimental data. Consider the characteristic spacing between adjacent F-actin sites in a row of the bundle, d. If the effective radius of a hypothetical ABP is r, then the ratio $\delta = 4r/d$ determines whether the 4- or 5-state model is more appropriate. If $\delta \sim 1$, then no more than one ABP could fit snuggly into the space between adjacent F-actin binding sites. This scenario represents the 4-state model. In contrast, if $\delta << 1$, two ABPs could simultaneously bind adjacent F-actins, a scenario represented by the 5-state model. The radii of gyrations of fascin (PDB ID 1DFC), fimbrin (PDB ID 1PXY), and alpha-actinin (PDB ID 1SJJ) are 2.5, 2.4, and 11 nm, respectively. The observed spacing between filaments in a hexagonal bundle is $d \approx 13$ nm [42], and the radius of gyration of F-actin is $R \approx 2.37$ nm [124]. Therefore $d \approx 8.3$ nm, which to just wide enough to fit a ABP like fascin or fimbrin. The effective length of alpha-actinin in bundles was measured to be about 35 nm [115], which is less than twice its radius of gyration. An alpha-actinin bundle could contain the doubly-bound state if the adjacent proteins avoid each other. Therefore both the 4- or 5-state models are appropriate depending on the effective size and flexibility of the ABP.

Adsorption of an ABP with allosteric binding sites

Allosteric coupling between ABP binding sites provides a means of increasing the maximal cross-link density (Fig. 4-4). In both the 4-state and 5-state models allosteric coupling decreases the fraction of blocked sites (Fig. 4-4a, c) while increasing the fraction of cross-links (Fig. 4-4b, d). Thus, allosteric coupling between ABP domains is one of nature's options for optimizing the cross-link density in F-actin bundles.

4.4 Concluding remarks

We developed a simple model for binding of ABPs to transverse F-actin bundles. Although our objective was to understand the consequence of allostery in ABPs on the structure of bundles, the model is general and can be used for other questions. One extension could incorporate the effects of intrafilament cooperativity among F-actin binding sites. Recent work by Galkin *et al.*, for example, proposes G-actins coordinate their states over a 17 protomer length scale [60]. As second extension one could include the ABP-induced F-actin deformations [152, 153] to create a rich mechanochemical

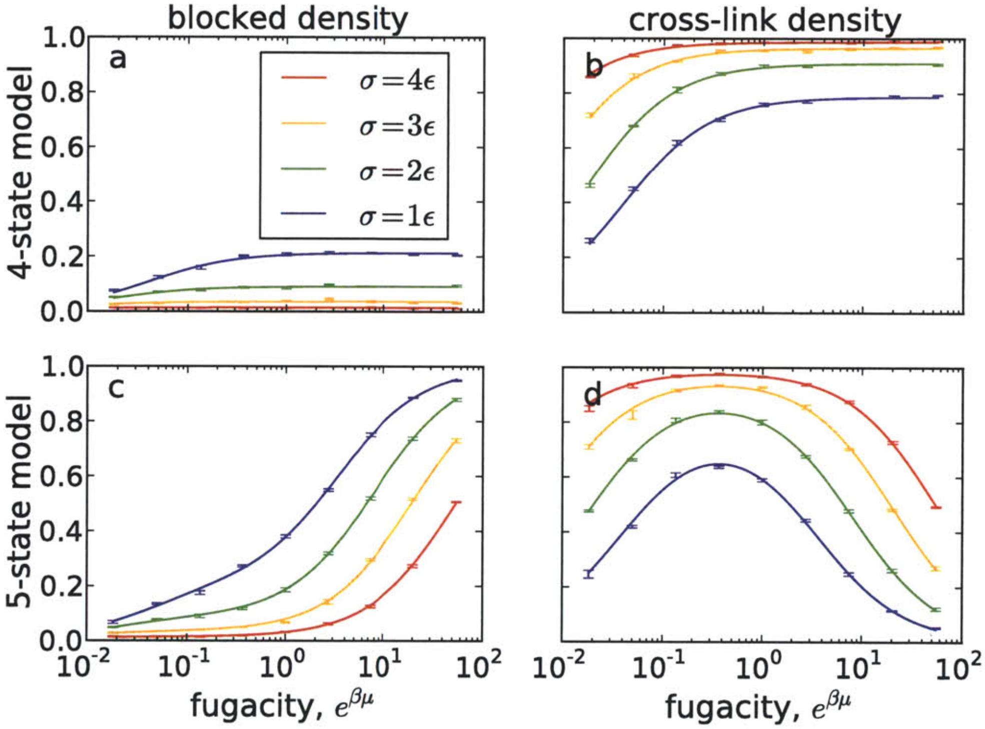

Figure 4-4: Allosteric coupling between tandem ABP binding sites drives cross-linking. The lines are analytic while the errorbars are from simulation. In all cases $\epsilon = -1$

model of F-actin bundles. Lastly, one could use the model to interpret experimental data. So-called pelleting assays are a means of measuring the bundling proclivity of ABPs. The assay has been performed for several bundling proteins, including annexin [81], epsin [30], fascin [180], alpha-actinin [115], fimbrin [64], and villin [84]. Interpreting pelleting assay data with our simple model may assist in the formulation and testing of hypotheses about the structure and stability of bundles. It is important to note, however, the technical issues with the assay, such as its inability to detecting whether or not an ABP is in any of the states from our model. What the assay actually determines is the total amount of ABP per actin. For that metric, one must compute the concentration of ABP in the bundle,

$$\begin{aligned} c_4 &= \frac{2e^{-\epsilon+\mu} + e^{-2\epsilon-\sigma+\mu}}{1 + 2e^{-\epsilon+\mu} + e^{-2\epsilon-\sigma+\mu}} \\ c_5 &= \frac{2e^{-\epsilon+\mu} + e^{-2\epsilon-\sigma+\mu} + 2e^{-2\epsilon+2\mu}}{1 + 2e^{-\epsilon+\mu} + e^{-2\epsilon-\sigma+\mu} + e^{-2\epsilon+2\mu}} \end{aligned}$$

which one can then normalize to the actin concentration. To test for ABP cooperativity one could make two defective ABPs, one with each F-actin binding site inhibited, and measure the titration curves relative to the wild-type ABP. One could then use the c_4 and c_5 to interpret the curves.

Chapter 5

Mediation of F-actin mechanics by actin subdomain 2

Abstract

Filamentous actin (F-actin) is a ubiquitous eukaryotic macromolecule that serves a range of important biological functions, including cell migration, division, adhesion and force sensing. Due to its centrality in cellular biomechanics, the mechanical properties of F-actin are of great interest *per se* and for constructing models of large biomechanical structures, like the networks and bundles found in organelles. Although numerous studies have developed useful mechanical models of F-actin, recent structural data now demonstrate that F-actin adopts an ensemble of states, each which we hypothesize exhibits unique mechanical characteristics. To address the mechanical consequences of the varied structural states we implemented a structure-based computer model to characterize the stiffnesses of the models. We find that our modeling predictions agrees well with the available experimental evidence. In addition we demonstrate that F-actin's mechanical behavior in general deviates significantly from standard assumption required by so-called "wire" F-actin models, by demonstrating significant coupling between stretching, twisting, and bending. We also show that mechanical properties of F-actin are sensitive to the structure of actin subdomain 2. Finally, we described an intuitive model for how the structure of S2 can mediate

the flexibility of F-actin. Overall, this work provides novel quantitative information on the ensemble of F-actin flexibilities that can be further transferred to large scale models for faithful simulation of F-actin within organelle or cellular contexts.

5.1 Introduction

Filamentous actin (F-actin) serves important biomechanical functions in a variety of cellular processes including migration, division, adhesion, and mechanosensation [136, 42]. As a force-bearing and -generating component of the cytoskeleton, the mechanical behavior of F-actin has received considerable attention [136, 50]. Experimental [83, 97, 108, 62, 182, 172, 146] and computational [19, 16, 37, 33, 133] studies provide detailed characterizations of the mechanics of the actin filament, which is typically coarse-grain modeled as a homogeneous and isotropic rod with a characteristic stretching, bending, and twisting stiffness, and more recently twist-bend coupling stiffness [37]. Such models of F-actin mechanics are interesting in and of themselves, but also permit the construction of larger scale mechanical models, such as F-actin networks [94, 61] and bundles [16, 36, 72], which are important for understanding processes occurring and organelle or cellular levels.

Numerous studies have demonstrated that F-actin's mechanical properties depend on the varied structural states that it samples, with preferences for particular states mediated by interactions with a diverse set of actin-binding proteins, small peptides, cations, and nucleotides. For example, while the peptide phalloidin, smooth and skeletal muscle tropomyosin, and the unphosphorylated actin-binding fragment of caldesmon, H32K, all increase the bending stiffness of F-actin [83, 62, 65], cofilin markedly decreases both its bending [113] and torsional stiffness [141]. In addition to proteins and peptides, cations, proteolysis, and chemical cross-linking of actin protomers also mediate filament flexibility [83, 127, 128, 134] by altering the 3D structure (reviewed in [73, 125]).

The atomic-level structure of the bare filament was recently re-examined in detail by several independent research groups [76, 124, 60, 54] . The newer F-actin models consistently demonstrate alignment of the major domains of G-actin that enclose the nucleotide binding cleft, while in the first F-actin model [77] the major domains are twisted as in numerous G-actin crystal structures (reviewed in [125]). Unique to Galkin *et al.*'s work is evidence that actin protomers within the filament adopt an ensemble of structural states. Five canonical structures, denoted modes 1-5, each comprise a synchronized and contiguous blocks of about 17 actin protomers in F-actin. Of G-actin's four subdomains, S1-S4 [87], S2 and its constituent DNase I binding loop (D-loop) principally differentiate the five modes by adopting distinct structural states in each mode. As noted by the authors, the structural polymorphisms of S2 are interesting from a mechanical perspective in light of direct evidence demonstrating that the structure of S2 in F-actin mediates filament flexibility [127]. To the best our knowledge, this experimental observation has not been reconciled with the five canonical modes.

The principal structural difference observed in G-actin that distinguish filament modes 1-5 lies in subdomain S2 (residues 33-69) (Fig. 5-1a). In modes 1-3 the D-loop (residues 38-52) adopts an ordered structure; a loop, helix, or a helix rotated 18° away from the axis of the filament (5-1b). The Oda and Fujii model D-loops adopt the extended state observed in mode 1, while the recent Holmes model D-loop is a helix like in mode 2 and 3 . In contrast to the modern F-actin models and modes 1-3, the D-loop is disordered in mode 4, while the entirety of S2 is disordered in mode 5 (Fig. 5-1c). Aside from S2, the rest of the actin protomer is structurally conserved, with a root mean squared displacement between S1, S3 and S4 less than 2.26 Å for all protomer pairs (Tab. 5.1). Moreover, the helical symmetry, described by the axial rise and rotation per monomer, differs by less than 0.01° and 0.1Å between modes, respectively, demonstrating that the quaternary structure of the five F-actin modes is also conserved.

Structure-based modeling provides a useful tool to test the hypothesis that the conformational state of the S2 sub-domain of G-actin mediates mechanical flexibility of the filament, as well as to examine in detail the structural origin of the observed changes therein. While previous coarse-grained and molecular dynamics simulations

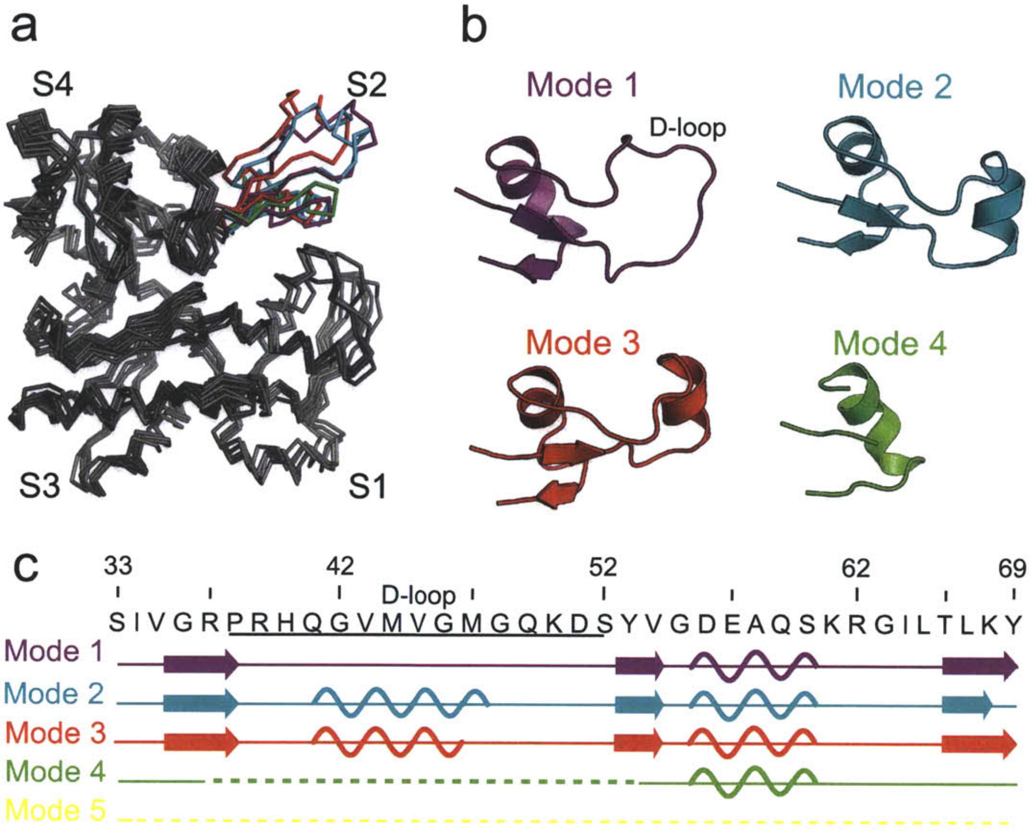

Figure 5-1: F-actin subdomain 2 and the DNase I binding loop is polymorphic. Subdomain 2 of F-actin is polymorphic. (a) A structure-based superposition of 5 G-actin modes shows that S2 is polymorphic (colored segments) while S1, S3 and S4 are structurally conserved (mean squared displacement, M.S.D. ¡ 2.3Å. Subdomains 1, 3, and 4 are colored gray, and S2 is colored purple, cyan, red, green, and yellow in modes 1-5, respectively. (b) Cartoon representations of the structure of S2 (residues 33-69), including the DNase I binding loop (residues 38-52). (c) In modes 1-4 the DNase I binding loop is a loop, helix, shifted helix, or disordered (dashed line); in mode 5 S2 is disordered and thus absent from the reconstruction. Secondary structure assignments are from the computer program DSSP [88] and rendering was done in PyMOL [40].

Mode	1	2	3	4
2	1.47			
3	1.57	1.81		
4	1.38	1.48	1.79	
5	2.26	2.00	1.99	2.17

Table 5.1: Subdomain 1, 3, 4 are structurally conserved across the five F-actin modes. Root mean square displacement between aligned optimal pairwise alignments of subdomains 1, 3, and 4, measured in Å.

confirm the experimental observation that F-actin flexibility does depend on the structure of the D-loop [19, 33, 133] this work focuses on the precise relationship between geometry and mechanics. For example, while a local reduction in filament mass along the backbone of the filament may locally lead to an increase in flexibility, it is unclear how this local change affects the overall mechanics of the filament on larger length-scales. Toward this end we constructed structure-based computational models of Galkin *et al.*'s five F-actin modes and subjected them to simple deformation or free vibration to quantitatively investigate the geometric role of S2 on the mechanical properties of F-actin.

5.2 Methods

Construction of molecular models

V. Galkin kindly provided the atomic G-actin models of F-actin modes 1-5 from [60] as well as the helical symmetry operations required to construct filaments. The symmetry operations expressed as an axial rise, z, and axial twist, θ, per monomer (z, θ) follow: mode 1 $(166.64^\circ, 27.53\text{Å})$; modes 2-4 $(166.60^\circ, 27.60\mathring{A})$; and mode 5 $(166.67^\circ, 27.60\mathring{A})$. We constructed left-handed 52-monomer filaments with computer program CHARMM [25, 27] version c35b1.

Key modeling assumptions

A common approach for modeling disordered peptide segments that is applicable here is to insert the peptide sequence, in this case the D-loop or S2 (Fig. 5-1c), into the filament model and use energy minimization or Monte Carlo simulation to generate one or more candidate conformations. Each conformation of the disordered segment then represents one of many possible structural states in the model. Here we adopted a parsimonious approach by assuming that the disordered segments of the filament, with their associated diffuse electron density, do not bear mechanical load when deformed. To implement this assumption we simple exclude the disordered segments from the structure when constructing the model. This approximation is equivalent to the assumption that mechanical stiffness of the filament is dominated by enthalpic contributions that require specific atomic-level interactions in well-defined conformational states, as opposed to entropic contributions locally induced by changes in the set of accessible chain conformational states.

The Finite Element representation of our filament models assumes the filaments behave as a homogeneous, isotropic, linear elastic solids with a constant Young's modulus, E, Poisson ratio, ν , and mass density, ρ. Although simplistic, this continuum representation of protein mechanics has been shown remarkably adept at predicting global motion of proteins [16]. We assigned all of the models the same material properties to compare purely geometrical differences between the modes 1-5. Specifically, we chose $E = 450$ MPa to obtain a mode 1 constrained bending stiffness of 7.3×10^4 pN·nm reported for rhodamine-phalloidin labeled F-actin [62]. Choosing other measurements for tuning gave similar estimates: with κ_E from [97] $E = 511$ MPa while tuning with κ_T from [182] $E = 355$ MPa. We assigned a constant mass density $\rho = 1.3$ g/cm^3 from calculating the volume of the finite element meshes and the mass of the atomic models. The ratio $c = (E/\rho)^{1/2}$ is the wave speed in a material, which along with the characteristic length of the model determines its vibration frequencies and persistence length, L_p. With E and ρ so chosen we obtained a reasonable mode 1 persistence length estimate of 9.7μm.

Mesh generation

To mathematically define the shape of the models we created triangular surface meshes from the atomic filaments model using GAMEr, a component of Finite Element Toolkit (FETK) version 1.3 [183] with default parameters. GAMEr constructs a surface by defining a level set of a Gaussian kernel density function, thus approximating the electron density of a protein. This approach differs from molecular surface mesh generation, such as MSMS [145] which approximate the shape of the protein exposed to a solvent environment. Both meshing strategies give volumes corresponding to the correct mass density for proteins, but GAMEr meshes are better quality. The surface meshes generated with GAMEr were minimally filtered using MeshLab version 1.3 [34] to remove non-manifold facets, intersecting facets, and to patch holes. Tetrahedral volumetric meshes were subsequently generated by constrained Delaunay tetrahedralization implemented in TetGen version 1.4.3 [154] with flag "-q1.333Fi" and additional inserted vertices at the coordinates of all of the heavy atoms.

Mechanical characterization

We subjected the filament models to simple deformations shown in Fig. 5-3b: extension, bending, and torsion. Dirichlet boundary conditions are defined as follows. At the atomic center of mass of G-actin number 1 and 52 we defined so-called master nodes, one for each end. All of the nodes within a 1 Å thick z-slice of the master node were assigned slave node a status. The slaves nodes were rigidly linked to the master node such that their distance and orientation was maintained when the master nodes was displaced or rotated about an axis. With these constraints we applied the following displacements to the master nodes to generate constrained extension, bending, and torsion:

$$u_z(0) = -1/2; \quad u_z(L) = 1/2; \quad \text{other degrees of freedom fixed} \tag{5.1}$$

$$\theta_r(0) = -1/2; \quad \theta_r(L) = 1/2; \quad \text{other degrees of freedom fixed} \tag{5.2}$$

$$\phi_z(0) = -1/2; \quad \phi_z(L) = 1/2; \quad \text{other degrees of freedom fixed} \tag{5.3}$$

$$\tag{5.4}$$

The reaction force, F_z, moment, M_r , and torque, T_z, at the master nodes equal the stiffness of the filament per unit length,

$$\begin{bmatrix} F_z \\ M_r \\ T_z \end{bmatrix} = \frac{1}{L} \begin{bmatrix} \kappa_E & \kappa_{EB} & \kappa_{ET} \\ \kappa_{BE} & \kappa_{BB} & \kappa_{BT} \\ \kappa_{TE} & \kappa_{TB} & \kappa_{TT} \end{bmatrix} \begin{bmatrix} u_z \\ \theta_r \\ \phi_z \end{bmatrix} \tag{5.5}$$

Each of the three boundary conditions in Eqn. 5.4 generates a column in the matrix of Eqn. 5.5. Finally, because of the finite length and helicity of the filaments the bending stiffness appears anisotropic. We therefore report an averaged bending stiffness by varying the orientation bending plane over a range $0 \leq \phi < \pi$ (Fig. 5-3a).

Finite element calculations were performed using the commercially available computer program ADINA version 8.7 (ADINA R&D, Inc., Watertown, MA). Static and normal mode calculations were conducted using eight 2.5 GHz CPUS with a total of 32 or 64 GB and RAM, respectively. Unconstrained normal modes were computed using the Subspace Iteration Method [147].

5.3 Results and discussion

Simulation of thermal fluctuations

The persistence length scale, L_p , provides a gross measure of the stiffness of a polymer that is readily comparable with experimental measurements of thermal fluctuations. Physically, L_p is defined as the length-scale over which tangent-tangent correlations decay along the contour length of the filament, l [100]. For bending-dominated fluctuations of a worm-like chain in three dimensions, L_p is related simply to the bending stiffness via $L_p k_B T = \kappa_B$ where k_B is Boltzmann's constant and T is temperature. In reality thermal fluctuations engage a complex mixture of bending, twisting, and stretching deformations. To capture such complex motion we performed unconstrained normal mode analysis (see methods) to obtain the shape and frequency of the free vibrations (Fig. 5-2a) The vibration frequency of the lowest non-degenerate harmonic bending mode, ω_1, is related to the persistence length of an equivalent homogeneous elastic rod by

$$L_p = \left(\frac{m\omega_1^2}{k_B T \beta_1^4} \right) l \tag{5.6}$$

where m is the mass of the filament and $\beta_1 l = 4.730$ [112]. The persistence lengths of modes 1-4 are similar ($8.8 \pm 1.1\mu m$) while of mode 5 is $23 \pm 3\%$ of the mean L_p of modes 1-4 (Fig. 5-2b). The apparent 23% reduction in the persistence length corresponding to a disordered S2 agrees with work by Orlova *et al.*, who identified a

25% reduction in persistence length when S2 is disordered [127].

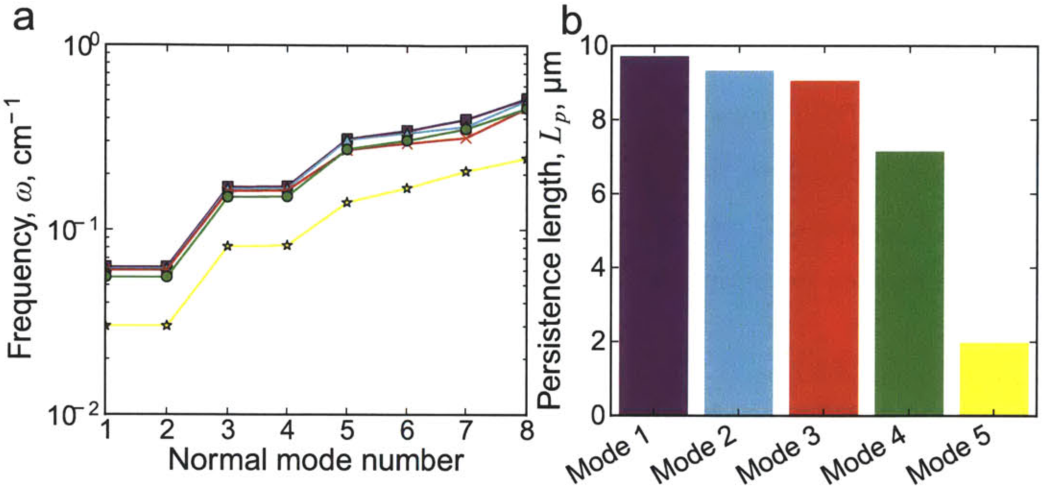

Figure 5-2: F-actin's free vibration flexibility is sensitive to the structure of S2. (a) Frequency distribution of the first eight normal modes. The first harmonic (normal modes 1 and 2) and second harmonic (modes 3 and 4) are degenerate bending modes. (b) The persistence length derived from the first normal mode frequency. Mode 5 is on average $23 \pm 13\%$ as stiff as modes the average of modes 1-4.

Computation of F-actin flexibility

By applying the appropriate boundary conditions we uncover the contributions of the stretching, κ_E, bending, κ_B , and twisting, κ_T, flexibility to the gross flexibility of F-actin (Fig. 5-3a). The five F-actin modes demonstrate variable stiffness in each of these three principal deformation modes (Fig. 5-3b). Mode 5 is the most flexible in all deformations. Averaging the flexibilities of mode 1-4 to serve as a stiff reference filament, mode 5 is then 53±4%, 29±4%, or 34±7% as stiff in extension, bending, and torsion, respectively. The increased flexibilities of mode 5 falls well outside the error range observed in the controlled experiments [97],[62],[182] (Fig. 5-3b). Moreover, the differences are larger than observed between independent studies observed in the literature. For example, Lui *et al.* [108] and Kojima *et al.* [97] measured a comparable κ_E, with the former $79 \pm 15\%$ of the later. Similarly, Yasuda *et al.* [182] and Tsuda *et al.* [172] both measured the torsional stiffness of F-actin and the ratio of Tsuda to Ysuda was $94 \pm 29\%$. The computed ranges are comparable to variation observed when F-actin is exposed to different ligands. Bathing F-actin with different ions can cut the bending stiffness to 32% of the larger stiffness [127] and the torsional stiffness to 33 ± 19% [182]. The fact that our purely structural model is sufficient to capture such dramatic differences underscores the importance of geometry in determining the stiffness of F-actin.

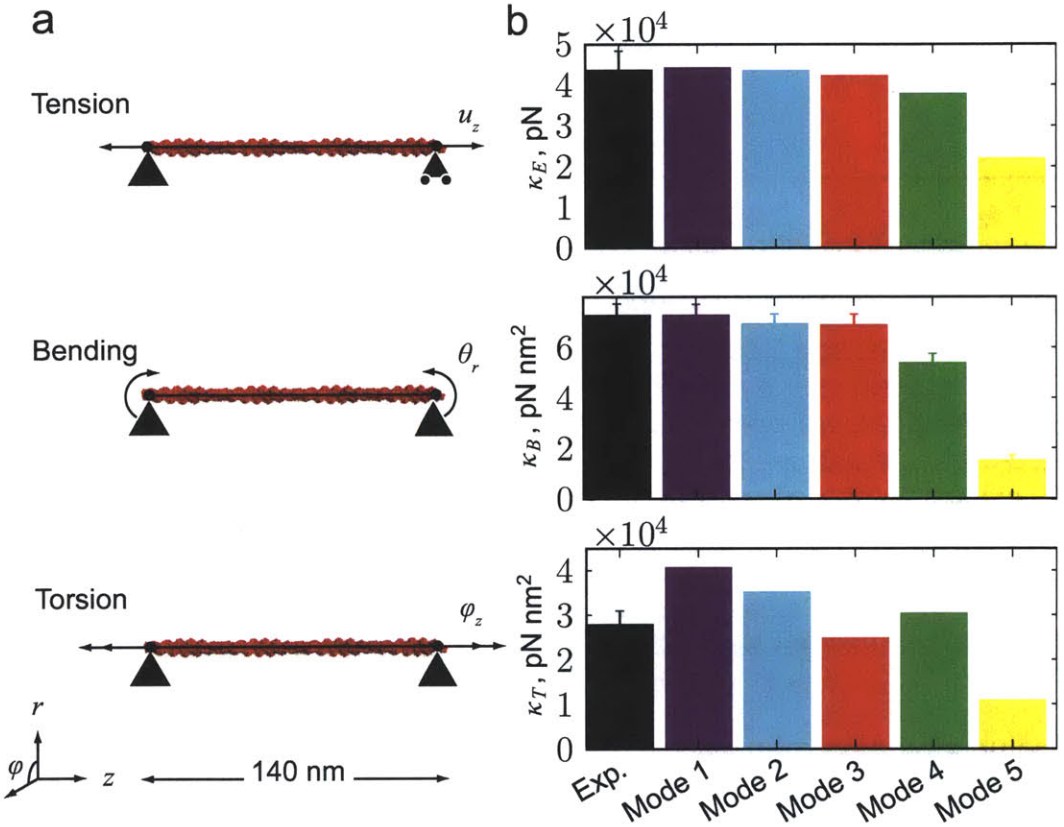

Figure 5-3: (a) Schematic of the deformations applied to F-actin: extension parallel to the long axis (z-direction); bending orthogonal to the radial axis (r-direction); and torsion parallel to the long axis (z-direction). In practice we vary bending axis over the azimuthal angle $0 < \phi < \pi$ radians to obtain an average bending stiffness that approximates the isotropic bending stiffness of a much longer filament. In bending and torsion the ends are free to rotate but may not translate in r- or z-direction. (b) Experimental and computed flexibilities. The three principal stiffnesses are extensional, κ_E, flexural, κ_B, and torsional, κ_T. Experimental results reported are from [97], [62], and [182], with error bars reflecting the experimentally reported standard errors. The error bars on the computed bending stiffnesses are standard deviations from the 20 azimuthally-distributed bending directions $0 < \phi < \pi$.

Coupling between deformation modes

Mode	$\kappa_{BE}^{\dagger} \times 10^4$ pN nm	$\kappa_{ET} \times 10^4$ pN nm	$\kappa_{BT}^{\dagger} \times 10^4$ pN nm^2
1	6.19	0.54	0.99
2	6.23	0.64	1.18
3	6.27	1.12	2.02
4	5.38	1.33	0.61
5	3.23	0.29	0.50

Table 5.2: F-actin demonstrates significant and varied stretch-bend, stretch-twist and twist-bend coupling. † Because the bending stiffness is anisotropic at 140 nm length scale (Fig. 5-3 b), we apply extension or torsion to measure well-defined stretch-bend or twist-bend coupling, respectively.

In addition to the principal rigidities so far characterized, the boundary conditions applied in Fig. 5-3b also generates forces and moments in proportion to extension-bending, κ_{EB}, extension-torsion, κ_{ET} , and bending-torsion, κ_{BT}, mechanical coupling coefficients (Tab. 5.2). As also shown by De La Cruz *et al.* [37], the coupling terms can be as large respect as the principal stiffnesses. Here we see, for example, that κ_T and κ_{BT} are both about 2 pN nm^2 in the mode 3 model, demonstrating that when F-actin adopts the mode 3 state it tends to untwist when bent more so than in any other mode. Moreover, the coupling terms vary significantly between models.

Geometric interpretation

Although our results so far were computed from continuum models of the filaments, the geometry of the continuum model is defined by atomic resolution structures. We therefore argue for an atomic interpretation of the mediation of F-actin mechanics by S2. Figure 5-4 depicts what we hypothesize is the geometrical determinant of F-actin flexibility, namely, the extent of ordered contacts between adjacent actin protomers in F-actin. Note that the longitudinal contact surface, defined as interface between protomers n and $n + 2$, decreases monotonically with mode number (Fig. 5-4a-b) similar to the extensional and bending stiffnesses and persistence length (Fig. 5-3 b, Fig. 5-2b). The correlation between κ_E, κ_B , or L_p and the buried longitudinal surface area, SES_{n+2}, is large (0.91, 0.93 and 0.93, respectively) while the correlation with the torsional stiffness κ_T is less (0.79) (Tab. 5.3), indicating that the longitudinal contact area is more predictive of the extensional and bending stiffnesses and the persistence length than the torsional stiffness. This relationship is rather intuitive given that the stress in both extension and bending is directed along the axis of the filament, and therefore longitudinal contacts bear the brunt of the axial load and consequently mediate the resistance of the filament to axial forces. In contrast to the longitudinal contacts, the extent of the lateral contacts in F-actin (Fig.5-4c-d), defined as interface between protomers n and $n+1$, is not monotonic with mode number, but rather mirrors the behavior of the torsional stiffnesses (Fig. 5-3b). The correlation between κ_T and the lateral contact buried surface area, SES_{n+1}, is 0.88, while the correlation between κ_E, κ_B, or L_p and SES_{n+1} is much lower, 0.60, 0.54, and 0.58, respectively (Tab. 5.3). The dependence of torsional stiffness on lateral contact interface follows since the lateral contact interface is positioned to support a circumferential shear stress in torsion.

	κ_E	κ_B	κ_T	κ_{ET}	κ_{BE}	κ_{BT}	L_p
SES_{n+2}	0.91	0.93	0.79	0.66	0.91	0.68	0.93
SES_{n+1}	0.60	0.54	0.88	-0.28	0.52	-0.26	0.58

Table 5.3: Linear dependence of mechanical measures on geometrical measures. The Pearson correlation quantifies the linear dependence between geometrical properties, the longitudinal, SES_{n+2}, and lateral, SES_{n+1}, buried surfaces areas Å^2 and measured mechanical properties, κ_{ij} and L_p.

5.4 Concluding remarks

The key contribution of this work is toward developing high-order models of F-actin bundles and networks. Based on the frequency of the five modes observed experimentally [60] and the mechanical characteristics of the modes (this work), detailed models are possible. Moreover, using or general framework one can incorporate the affects of actin binding protein decoration on the mechanics of actin filaments, bundles, and networks (Chap. 6).

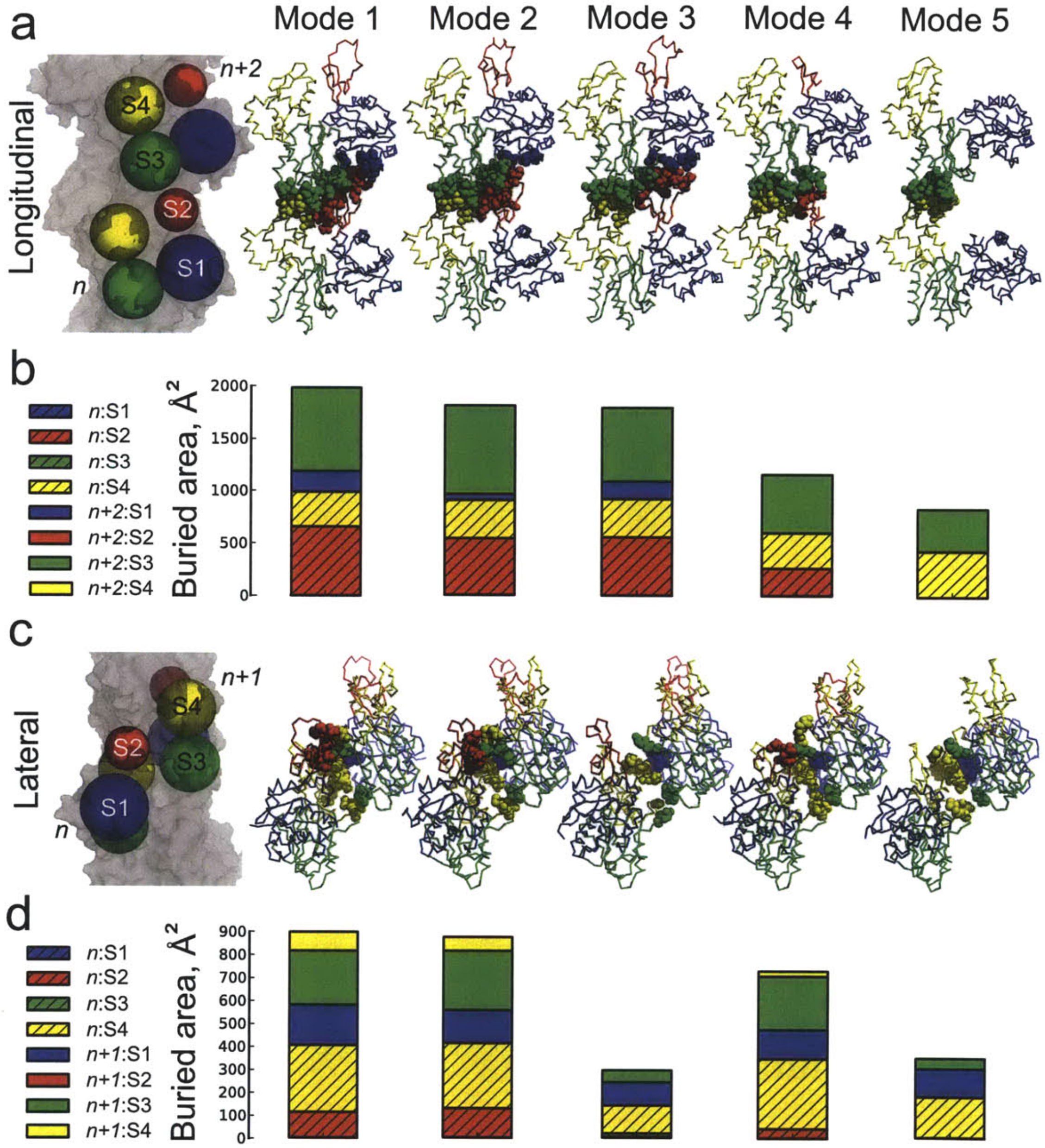

Figure 5-4: Actin S2 polymorphisms mediate the lateral and longitudinal interface. (a) Longitudinal contacts between protomers n and $n + 2$. Actin subdomains are color coded: S1 (blue), S2 (red), S3 (green), and S4 (yellow). Interprotomer contacts, defined as residue pairs with heavy atoms at most 5 Å, are rendered as Van der Waals spheres. (b) Buried surface area of each subdomain in the longitudinal contact. (c) Lateral contacts between protomers n and $n + 1$. (d) Buried surface area of each subdomain in the lateral contact.

Chapter 6

Mediation of F-actin stability by actin binding proteins

Abstract

Filamentous actin (F-actin) binds with a host of proteins that regulates its mechanics. Whereas some actin binding proteins (ABPs) increase F-actin's flexibility and decrease its stability, others do the exact opposite. Structural studies of decorated F-actin lead to the conjecture that the calponin-homology domains of alpha-actin and fimbrin stabilize F-actin by stapling together adjacent actins within the same protofilament. To test this hypothesis we used structure-based computer modeling of bare and decorated F-actin to measure the flexibility and stress concentration in the filaments when subjected to loads. We find that both ABPs increase the gross stiffness while relaxing the strain at actin-actin interfaces. These results are consistent with experimental observations and provide additional unique insight into the mediation of F-actin mechanics and stability by ABPs.

6.1 Introduction

Globular (G-actin) and filamentous actin (F-actin) interact with over 160 actin binding proteins (ABPs) to form a complex system that serves a broad range of functions important for regulating the structure of the cytoskeleton [44]. Bundling and cross-linking ABPs [179] organize F-actin into structures that are required for a variety of cellular processes, including muscle contraction [142], cytokenesis [132], intracellular transport [110], and cell migration [136]. A thorough understanding of such cellular processes that includes quantitative mathematical modeling requires a detailed understanding of the mechanical properties of the bare and decorated F-actin [61, 72, 94, 36, 16].

The mechanical properties of both bare and decorated of F-actin are well studied. Innovative experimental techniques provide estimates of F-actin flexibilities in simple deformations like extension, κ_E, [97, 108], torsion, κ_T, [172, 182], or freely fluctuating in a bending dominated motion, κ_B, [172, 182, 83]. In addition, F-actin's helical geometry engages coupled deformation modes such as twisting coupled with bending [16, 19, 37] . Importantly, F-actin's mechanical properties are not fixed but vary in response to seemingly subtle structural changes, particularly in the DNase I binding loop (D-loop) in subdomain 2 (S2) [127, 33, 133] and the various actin ligands which can modulate the flexibility of F-actin over several fold [172, 65, 146, 73]. In fact, a variety of binding factors mediate the flexibility of F-actin (reviewed in [73]), including divalent cations [127], peptides like phalloidin, and ABPs. While some ABPs that decorate F-actin make the filament more flexibile, others make it stiffer. For example, cofilin increases both the flexural [113] and torsional [141] flexibility by shifting the D-loop and hydrophobic plug away from the C-terminus [146]. In contrast, unphosphorylated caldesemon fragment H32K as well as smooth and skeletal muscle tropomyosin/troponin increase the bending rigidity 1.5-2 fold [172, 65] . How any given ABP mediates the flexibility of F-actin remains an open question.

A second question closely related to F-actin flexibility concerns the stability of F-actin when decorated with ABPs. Like their effects on flexibility, ABPs demonstrate opposing effects on filament stability. For example cofilin promotes disassembly [103] while alpha-actinin [29, 104], fimbrin [104, 32], and coronin [56] inhibit disassembly. Based on electron microscopy data Galkin *et al.* [57] propose that fimbrin's actin-binding domain 2 (ABD2) stabilizes F-actin by stapling adjacent protomers within the same protofilament, thereby forming a bridge that stabilizes a "crack" between actin subunits and prevents it from growing until the filament severs. The calponin-homology domain of alpha-actinin CH3 [59] and coronin-1A bind similarly [56], suggesting that the bridging mechanism might be conserved. Determining whether or not the bridging-stabilization hypothesis is correct requires a structure-based mechanical model, which to our knowledge has not been implemented for this purpose.

This paper presents a description of how calponin homology domains from the ABPs fimbrin and alpha-actinin decorate F-actin to mediate it flexibility and stability. Using structure-based computer modeling we first show that ABP decoration decreases the flexibility of the F-actin in extension, bending, and torsion. We next demonstrate the validity of the bridging-stability hypothesis using the suggestion from [119] that proposes regions in a protein with elevated strain energy are the most susceptible to unfolding and fracture. We show that the strain energy at protomer-protomer interfaces within F-actin are stabilized by ABP decoration.

6.2 Methods

Additional details are provided in Chapter 5.

Molecular models

We constructed four molecular models: F-actin decorated with fimbrin/L-plastin actin binding domain ABD2, which comprises calponin-homology domains CH1 and CH2 [59]; F-actin decorated with alpha-actin CH1 [57]; and the two undecorated models derived from the aforementioned structures by deleting the ABPs. We constructed the models with 52 G-actin subunits by applying helical symmetry operations 51 times to a seed subunit. The rise and twist operations are 166.5°/27.3Å (fimbrin) and 167.2°/26.6Å (alpha-actinin). The final lengths of the filaments are both $\approx$140 nm. We attempted this protocol with a model of coronin-decorated F-actin [56] but hand to abandon it because the strands in undecorated model are disconnected.

Finite element analysis

The distribution of strain energy density (SED) throughout the filament provides a 3D map of the parts of the filament that are bearing load. The SED, $W(\mathbf{x})$, is scalar function of position, $\mathbf{x}$, and is defined as $dW = \sigma d\epsilon$, where $\sigma(\mathbf{x})$ is the stress tensor and $\epsilon(\mathbf{x})$ is the strain tensor [55]. From the displacement field we computed the strain, stress, and SED, which we then normalized to the work done on the filament. Normalization permits a direct comparison of the spatial distribution of the strain energy density within the filament with and without ABP decoration. Physically, F-actin regions with reduced SED when decorated are stabilized by ABP while regions with increased SED are destabilized [119].

6.3 Results and discussion

ABPs decoration decreases F-actin's flexibility

Applying a unit extension, u_z, rotation, θ_r, or twist, ϕ_z, to the ends of the four filament models (Fig. 6-1 a) generates reaction forces and moments proportional to the extensional, κ_E, bending, κ_B , and torsional, κ_T, flexibilities. All three principal stiffnesses increase with ABP decoration (Fig. 6-1 b). The ABPs increase the extensional and torsional stiffness by about the same amount (28% alpha-actin vs 26% fimbrin). In contrast, fimbrin increases the bending stiffness nearly twice as much as alpha-actinin (63% fimbrin vs 34% alpha-actinin). The ABPs make the largest impact on torsion, where alpha-actin and fimbrin increase κ_T 96% and 81%, respectively.

In addition to the principal stiffnesses, application of the boundary conditions shown in Figure 6-1a provide the coupling between deformation modes (Tab. 6.1). The undecorated filaments coupling coefficients are significant in comparison to the principal stiffnesses. Like the principal stiffnesses, decoration increases the coupling stiffnesses. Fimbrin increases the extension-torsion, κ_{ET}, and bending-torsion, κ_{BT}, stiffnesses 264% and 205%, respectively, while alpha-actin increases the same coupling coefficients 176% and 148%, respectively. Both ABPs increase extension-bending

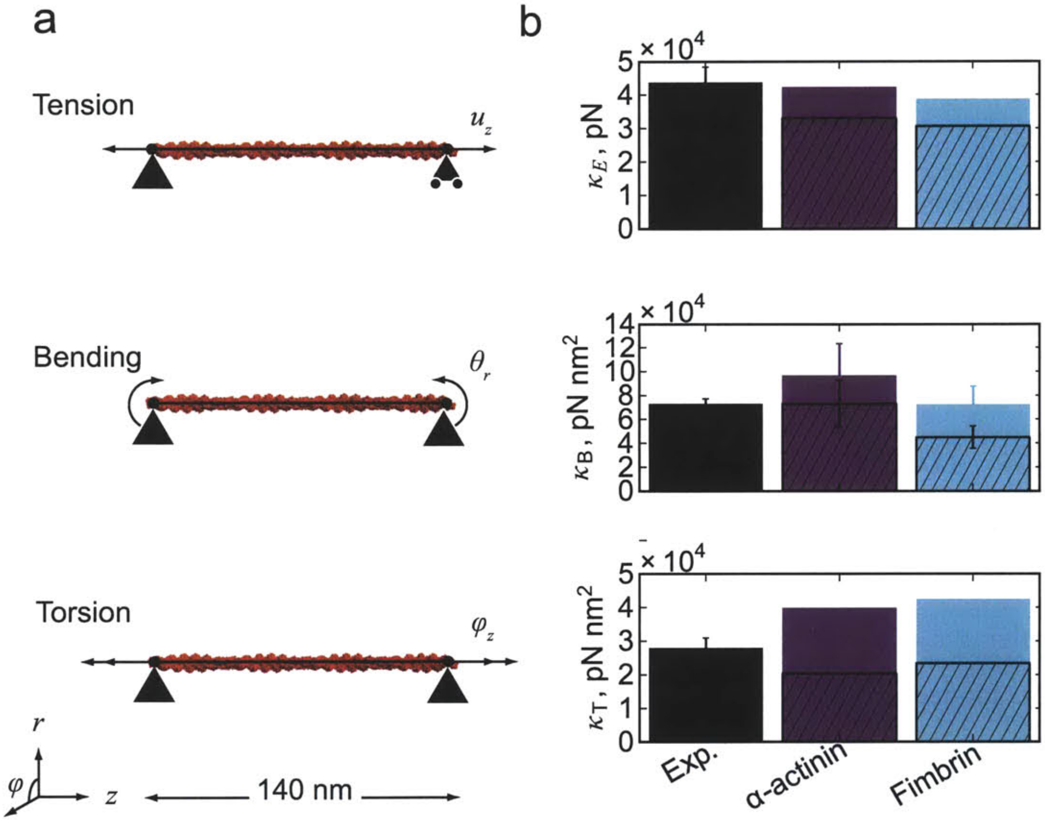

Figure 6-1: ABPs decrease the flexibility of F-actin. (a) Schematic of the deformations applied to the F-actin models: extension parallel to the long axis, u_z; bending parallel to the radial axis, θ_r; and torsion parallel to the long axis, ϕ_z. (b) Comparison of decorated and undecorated filament stiffnesses with experimental reference values κ_E [97], κ_B [62], κ_T[182]. The solid bars correspond to the decorated filament while the hashed bars correspond to bare F-actin in the decorated conformation. The error bars on the computed bending flexibilities are standard deviations of the distribution of κ_B derived by varying the bending axis r over polar angle $0 \leq \phi < \pi$ radians.

coupling, κ_{EB}, only modestly, about 25%. To our knowledge, these data are the first complete characterization of the mechanical properties of F-actin and its sensitivity to calponin-homology domain decoration.

Model	$\kappa_B E^{\dagger} \times 10^4$ pN nm	$\kappa_E T \times 10^4$ pN nm	$\kappa_B T^{\dagger} \times 10^4$ pN nm^2
alpha-actinin	8.90	1.79	3.33
Δ alpha-actinin	7.01	0.65	1.34
fimbrin	7.66	1.30	2.18
Δfimbrin	6.16	0.36	0.71

Table 6.1: Coupling between deformation modes in F-actin with and without (Δ) ABP decoration.

ABPs redistribute F-actin strain energy density

ABP decoration differentially affects SED in and around the actin subdomains, S1-S4. We computed the mean SED difference between the decorated and undecorated filaments by averaging W per α-carbon over protomers in the 52-protomer filament. While the SED of subdomain S1 increases with ABP decoration (alpha-actinin or fimbrin), the SED of subdomain S2-S4 decreases (Fig. 6-2). The observation is invariant to the both ABP decorating F-actin and mode of deformation. The correspondence between SED difference and the four actin subdomain indicates that the strain energy redistribution is function of the 3D structure of F-actin and the mechanism of CH domain binding.

The SED change due to ABP decoration relaxes the strain energy at the actin-actin interfaces while increasing the strain energy at the actin-ABP interfaces (Fig. 6-3). Here we define the interfaces between subunits as the heavy atoms within 5 Å of another subunit. Atoms in regions within the top 95% or bottom 5% of the SED range cluster at the actin-ABP or actin-actin interfaces, respectively. Because this observation is independent of the ABP and mode of deformation, we hypothesize the observed strain redistribution is a general mechanism of CH domain decoration, which is consistent with the conserved stapling mechanism of alpha-actinin, fimbrin, and coronin, which all stabilize F-actin [32, 104, 29, 56].

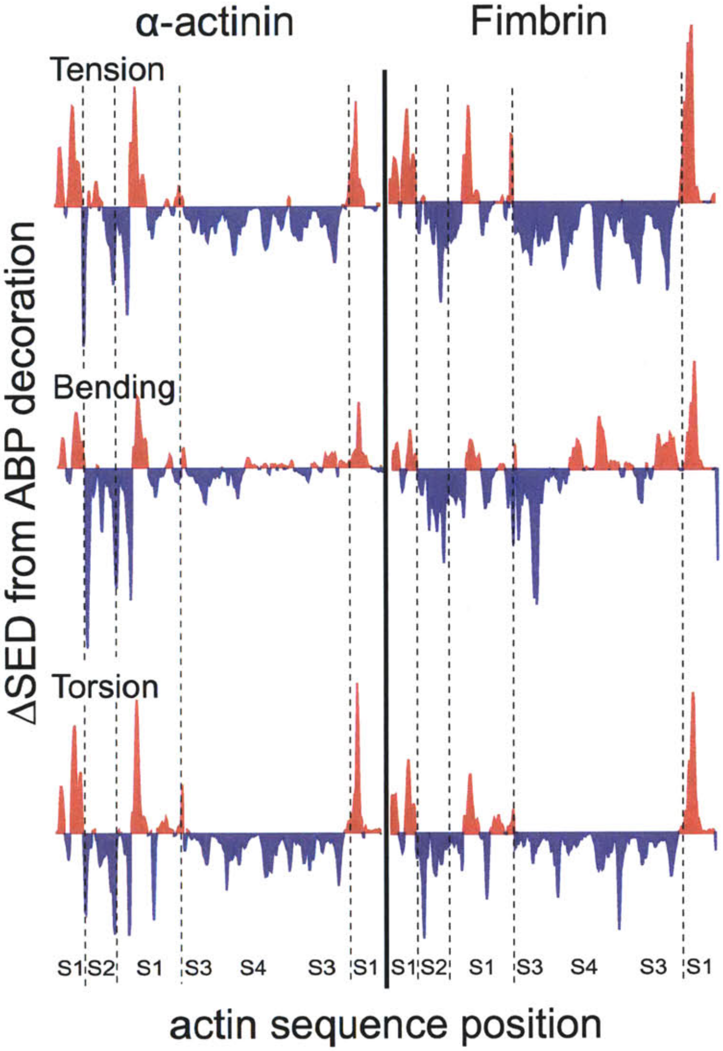

Figure 6-2: Actin binding proteins destabilizes actin subdomain 1 while stabilizing subdomains 2, 3, and 4. Each plot shows the change in strain energy density per unit work due to ABP decoration as a function of actin sequence position. Destabilized regions are shown in red while stabilized regions are shown in blue. The location of actin subdomains S1, S2, S3, and S4 [87], is shown for reference.

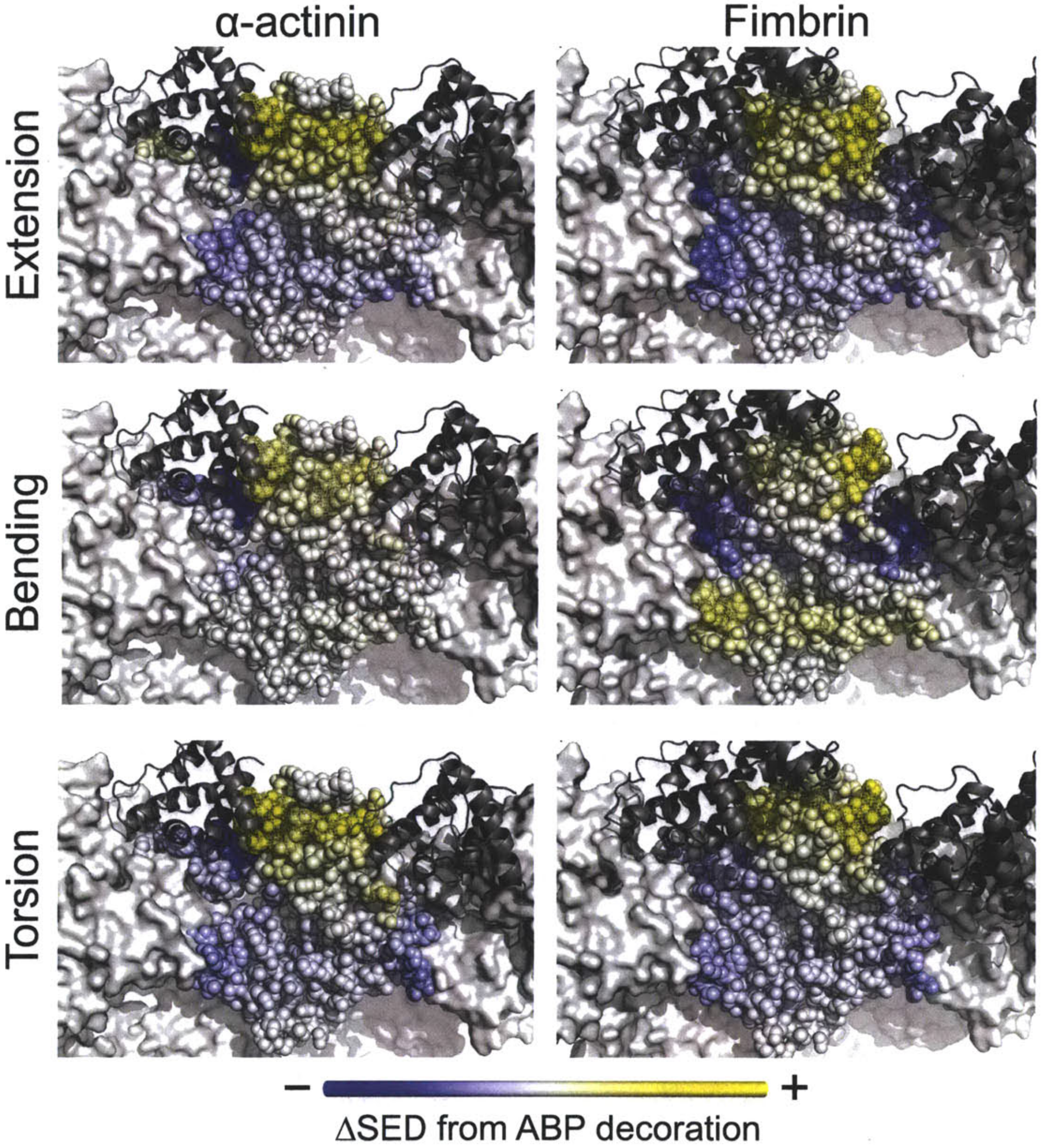

Figure 6-3: Actin binding proteins stabilize the actin-actin interface at the expense of the actin-ABP interface. Irrespective of the ABP (alpha-actinin: left column or fimbrin: right column) and deformation mode (extension: top row; bending middle row; or torsion: bottom row), the SED per unit work is less at the actin-actin interface (blue heatmap) than at the actin-ABP interface (yellow heatmap). In each panel, four neighboring actins are shown relative to a central G-actin in surface representation The ABP (alpha-actinin or fimbrin) is colored gray and rendered as a cartoon.

A simple model explains how ABPs increase F-actin's stiffness and stability (Fig.6-4). The space between adjacent actins within the same strand function like cracks in a beam. When an ABP binds to F-actin it staples together the adjacent protomers, thereby bridging the crack. Once bridged the tip of the crack supports less load when deformed, which makes the crack less likely to propagate and cause fracture. Moreover, because a crack makes the filament locally thinner and consequently more compliant, bridging the crack with an ABP necessarily increases the stiffness of the filaments in all deformation modes. This model is consistent with observations reported here and with experimental evidence showing both alpha-actinin [29, 104] and fimbrin [104, 32] stabilize F-actin.

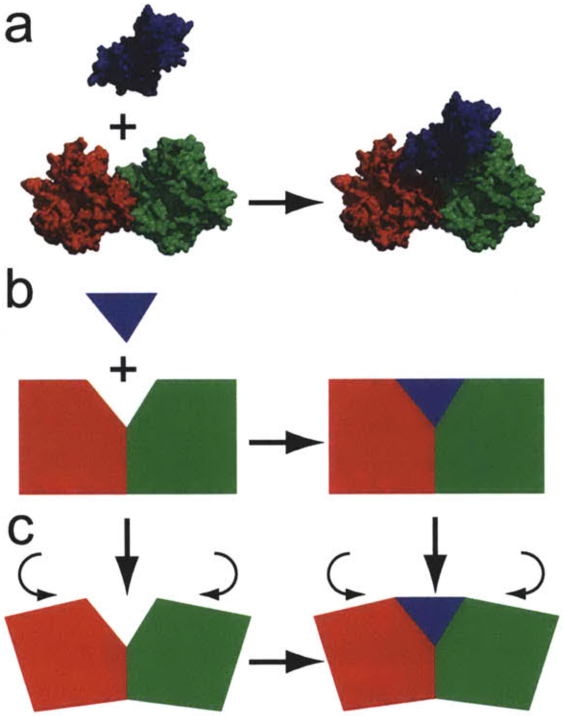

Figure 6-4: Crack-bridging by actin binding proteins. The ABP staples adjacent actin within the same strand. The staple relaxes stress at the crack, thereby increasing the stiffness of the filament and increasing its stability.

6.4 Concluding remarks

We characterized the affect of ABP decoration on the flexibility and stability of F-actin under simple deformations. We find that the ABPs alpha-actinin and fimbrin increase the stiffness of F-actin by redistributing the load-bearing responsibilities from the actin-actin interface to the actin-ABP interface. Further insight into F-actin mechanics may be gained by similar computational experiments as those presented here, but using other decorated [56, 107, 114] or bare [58, 174, 13] filament models. We anticipate that a systems-level understanding of how the over 160 ABPs mediate the flexibility and stability of F-actin may enable precise descriptive models of cytoskeletal mechanics.

Chapter 7

Perspective

In this work we investigated topics loosely grouped under the umbrella of molecular biomechanics of proteins and protein assemblies: the source of cadherin binding specificity (Chap. 2); unsupervised methods for detecting allostery from protein conformational dynamics (Chap. 3); the consequence of allostery on the structure of transverse F-actin bundles (Chap. 4); and the role of geometry in controlling the mechanical behavior of F-actin (Chap. 5 and 6). By way of summary we recapitulate our key findings.

Key findings

- The cadherin-cadherin dimer interface is enriched with putative specificity determining residues.

- There still exists an unmet need for unsupervised methods and benchmarks for detecting allostery in proteins from conformation dynamics.

- Cooperative binding of actin binding proteins to bundled F-actin promotes cross-linking over other modes of F-actin decoration.

- The structure of actin subdomain 2 mediates F-actin flexibility.

- Fimbrin and alpha-actinin relax strain energy at protomer-protomer interfaces in F-actin.

Outlook

We next mention a few points discussed casually with colleagues over the past few years. Through these discussion it is clear that many of us share a goal of solving problems in biology but there are many different views on the strategies for making progress. Below I outline a few thoughts on harnessing the vast amounts of structural data for solving problems in biology.

Biomechanics at the nanometer length scale

Biomechanics spans many length scales, from the molecular mechanics of protein to the movements joints and limbs. A key area going forward is an understanding of the biomechanics and function of large molecular complexes [93]. Before we can really understand large complexes, we must decide the level of abstraction that is permissible. In other words, we must decide how detailed a model needs to be for it to be useful. For example, physics permits the construction of very high resolution cryo-electron microscopy maps, although technology currently lags behind [63]. Is atomic resolution necessary for modeling large molecular complexes, or does such detail make the model more complicated that necessary? Hopefully time will tell.

Choosing the appropriate computational tool

A second need in computational biology is a means of disseminating computational strategies that helps people choose the most appropriate method of those that are available. It is now seems very easy to run many different computer analyses on the same data set (see Chap. 3). As the number of tools increases, an individual's capacity to take in new information does not, mostly due time constraints. Biologist are then left with a dilemma of choosing which approach to take. In my work the one chosen is often the easiest to use, but in some cases its simply the most popular choice (i.e. best brand). Going forward, computational and wetlab biologists need a principled way to choose between different strategies for any given problem worth solving. To my knowledge such guidelines do not exist.

Integration

A still unmet challenge in computational biology is how to quantitatively integrate diverse forms of data to construct models. An example of a system where one needs to integrate diverse forms of data is at the interface of molecular evolution and molecular dynamics. A new view of protein evolution claims that highly dynamic proteins are promiscuous, and the inherent promiscuity favors evolvability [171]. Right now there seems to be no principled framework for weaving together quantitative data sets like those presented in Chapters 2 and 3. Quantitative integration of diverse datasets may provide ample opportunity for exercising creativity for solving problems in biology.

Appendix A

Supporting materials for Chapter 2

Theoretical background

Our analysis of homologous cadherin protein sequences relies heavily on tools from information theory, originally developed by Shannon while employed at Bell Labs [149], and reviewed extensively in [85]. We use a small but useful fraction of Shannon's treatise, namely, entropy and mutual information.

The entropy of a discrete random variable X with n outcomes $\{x_1, ..., x_n\}$ is

$$H(X) \quad = \quad -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{A.1}$$

where $p(x_i)$ is the probability mass function of the outcome x_i. A random variable with only one outcome has an entropy equal to zero. In contrast, a random variable with n outcomes that all occur with equal probability $p_i = 1/n$ has a maximal entropy, $H = \log(n)$. It is sometimes convenient to scale the entropy of random variable to unity, which requires only dividing Eqn. A.1 by $\log(n)$, or alternatively using n as the base of the logarithm. Note that a outcome with zero probability is permissible since

$$\lim_{p \to 0+} p \log p \quad = \quad 0$$

Also note that although the limiting value is zero, small values for p still contribute to entropy because the function $f(p) = p \log p$ grows quickly.

The co-variation of two random variables X and Y can be quantified by a metric called the mutual information

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{A.2}$$

where $p(x_i, y_j)$ is the joint distribution of X and Y. We typically require that

$$p(Y) = \sum_{i=1}^{n} p(x_i, Y)$$

although that assumption is unnecessary [48]. Mutual information is minimized when the random variables X and Y are independent. When independent, $p(x_i, y_j) = p(x_i)p(y_j)$ and the logarithm in Eqn. A.2 is always zero. Mutual information may also be written in terms of joint, marginal, and conditional entropies

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X|Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

It then follows that $MI(X,Y) \leq \min\{H(X), H(Y)\}$ [111].

Often one does not know either the marginal or joint distribution that is of interest. Instead, one has a table of co-occurrences N_{ij} from which to infer the distribution. The frequentist approach to estimating probability distributions is to use the frequency of a realization $N_i / \sum_i N_i$ as an approximation for $p(x_i)$. Approximating probability with frequency converges as the amount of data increases, but the amount of data necessary for convergence is not typically known *a priori*. In general, a Bayesian approach to estimating probability distribution is more principled. In this case one can apply Bayes' rule to estimate the joint probability from the data

$$P\left[p(x_i, y_j)|n_{xy}\right] = \frac{P\left[n_{xy}|p(x_i, y_j)\right] P\left[p(x_i, y_j)\right]}{P(n_{xy})}$$

Estimating the posterior distribution can be computationally intensive, so for this work we compiled a data set that is much large than the empirically determined bound of 125 sequences [111], and use the simple frequentist approach.

Paralog specificity analysis

Hsap cdh1	Mmus cdh11	p-value
D1†	G1‡	3×10^{-7}
I4	W4‡	2×10^{-8}
V22	G22‡	8×10^{-8}
I24†	L24‡	1×10^{-16}
G40	G41	1×10^{-11}
G42	G43	4×10^{-14}
A43	A44	1×10^{-11}
V50	I47	3×10^{-8}
T57	S54	4×10^{-6}
E89†	E87‡	1×10^{-7}

Table A.1: Predicted specificity-determining residues corresponding to clade I. Residues with established functions include those that participate in Type I cadherin strand-swapping, †, Type II cadherin strand-swapping, ‡, or residues that that coordinate Ca^{2+}, *.

Hsap cdh1	Mmus cdh11	p-value
I4	W4‡	4×10^{-8}
I7	F7‡	2×10^{-11}
N12	E12*	4×10^{-18}
V22	G22‡	1×10^{-7}
I24†	L24‡	7×10^{-13}
M92†	S90‡	2×10^{-6}
I94	F92‡	6×10^{-13}

Table A.2: Predicted specificity-determining residues corresponding to clade II. Residues with established functions include those that participate in Type I cadherin strand-swapping, †, Type II cadherin strand-swapping, ‡, or residues that that coordinate Ca^{2+}, *.

Hsap cdh1	Mmus cdh11	p-value
D1[†]	G1[‡]	2×10^{-6}
I4	W4[‡]	1×10^{-9}
P5[†]	N5	2×10^{-6}
I7	F7[‡]	6×10^{-10}
N12	E12[*]	3×10^{-16}
L21	V21	2×10^{-7}
V22	G22[‡]	4×10^{-11}
I24[†]	L24[‡]	7×10^{-20}
M92[†]	S90[‡]	1×10^{-8}
I94	F92[‡]	7×10^{-14}

Table A.3: Predicted specificity-determining residues corresponding to clade III. Residues with established functions include those that participate in Type I cadherin strand-swapping, [†], Type II cadherin strand-swapping, [‡], or residues that that coordinate Ca^{2+}, [*].

Hsap cdh1	Mmus cdh11	p-value
D1[†]	G1[‡]	5×10^{-7}
I4	W4[‡]	2×10^{-8}
I7	F7[‡]	3×10^{-8}
N12	E12[*]	4×10^{-15}
L21	V21	1×10^{-6}
V22	G22[‡]	2×10^{-9}
I24[†]	L24[‡]	6×10^{-17}
L60	I57	5×10^{-6}
M92[†]	S90[‡]	3×10^{-10}
I94	F92[‡]	6×10^{-13}

Table A.4: Predicted specificity-determining residues corresponding to clade IV. Residues with established functions include those that participate in Type I cadherin strand-swapping, [†], Type II cadherin strand-swapping, [‡], or residues that that coordinate Ca^{2+}, [*].

Appendix B

Supporting materials for Chapter 3

Relation between the Pearson and generalized correlation coefficient

The generalized correlation is a generalization of the Pearson correlation. In the case of a block covariance matrix

$$\Sigma = \begin{pmatrix} \mathbf{I}_d & r_P\mathbf{I}_d \\ r_P\mathbf{I}_d & \mathbf{I}_d \end{pmatrix}$$

where $\mathbf{I}_d$ is a d-dimension identity matrix, the linearized mutual information, LMI, relates to the Pearson correlation coefficient, r_P, via

$$LMI = -\frac{d}{2}\ln\left(1 - r_P^2\right)$$

Thus, r_P is equivalent to r_{LMI} when the fluctuations are colinear.

Benchmarks for algorithms intended to detect allostery

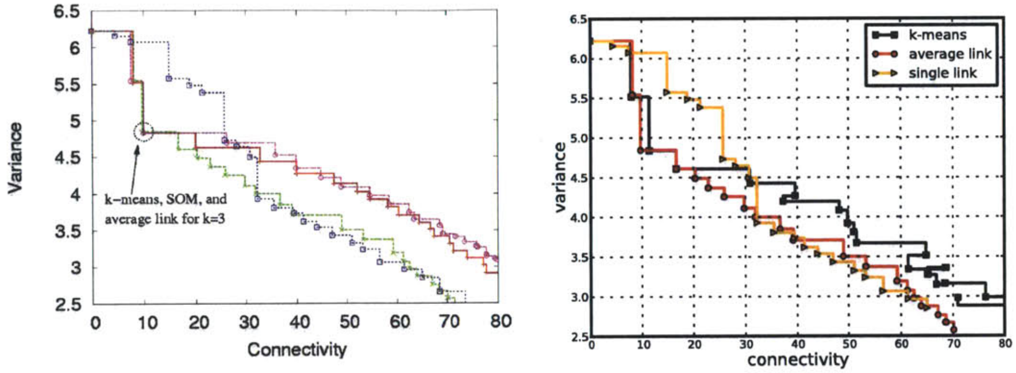

Figure B-1: Test of internal clustering validation metrics on Leukemia data. The consensus number of clusters is three, as demonstrated by a "knee" at the third data point. The data and analysis protocol is from [66].

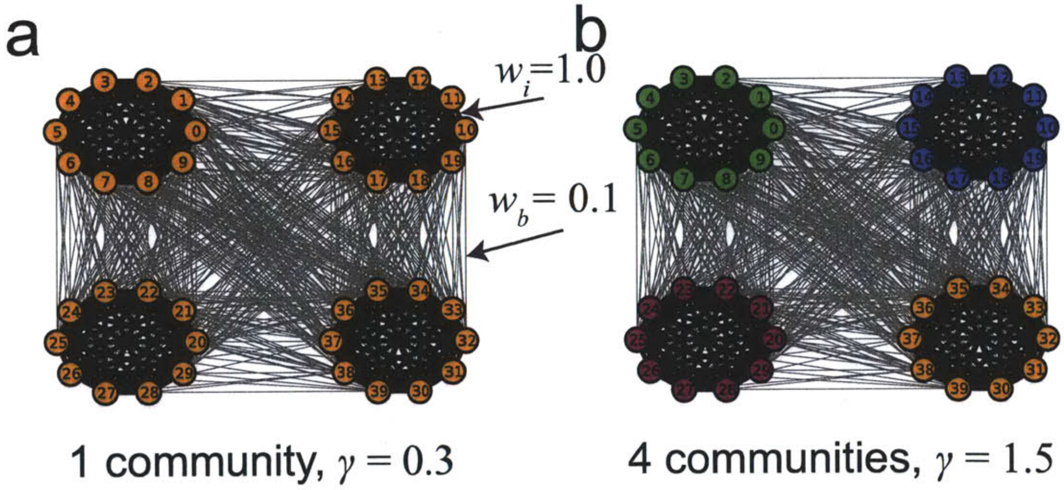

Figure B-2: .

Validation of community detection computer program on a test case from [70]. The network is a dense graph with four communities. The edge weight between nodes within the same community is $w_i = 1$ while the edge weight between nodes from different communities is $w_b = 0.1$. When the resolution parameter is $\gamma = 0.3$ the algorithm detects one large community. The algorithm detects four communities when $\gamma = 1.5$.

Predicted allosteric networks in Hemoglobin

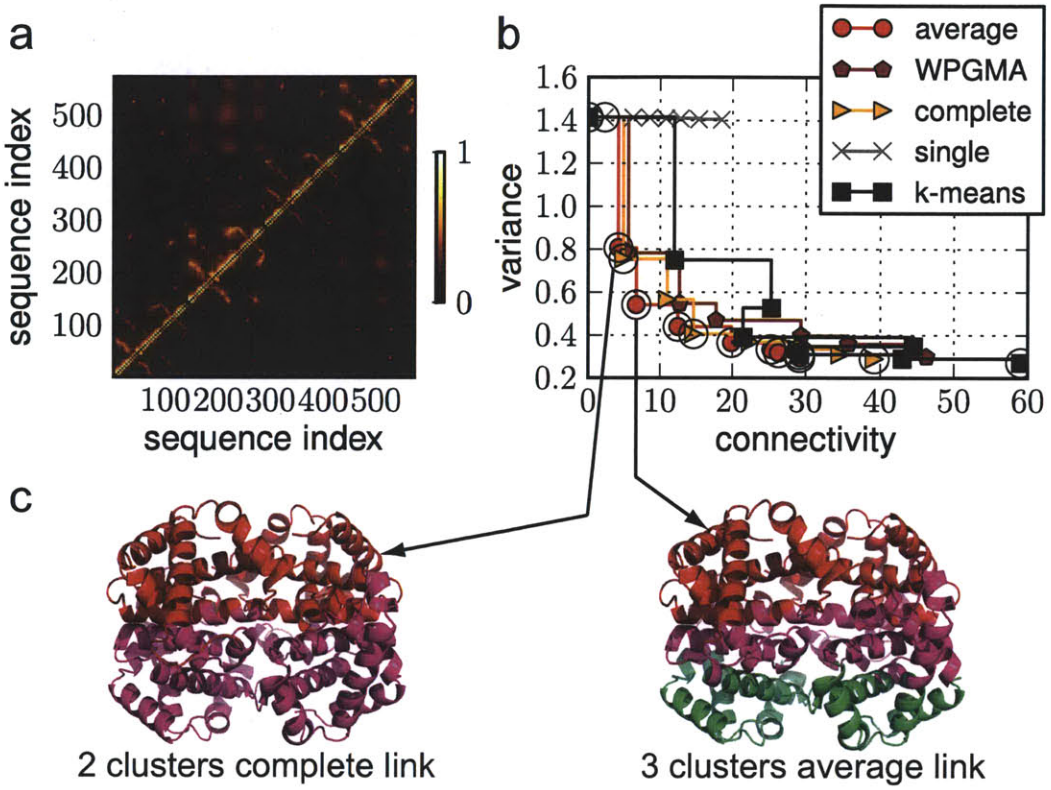

Figure B-3: Pareto-optimal dynamics-based allosteric network in hemoglobin. The first pareto optimal solution splits the α and β chains (c). The next optimal solution in the hierarchy, three clusters, splits the interface between the α and β chains into a separate cluster

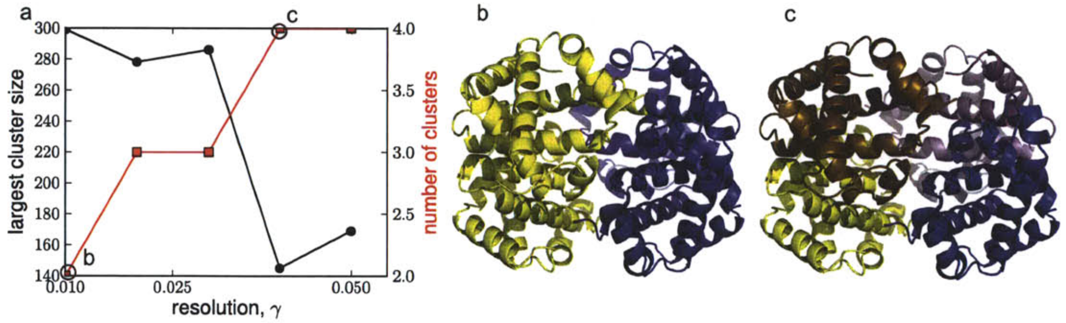

Figure B-4: Correlated dynamical networks in Hemoglobin inferred by community detection. The lowest resolution mapping divides the $\alpha_1\beta_2$ half. The lowest resolution mapping divides the $\alpha_1\beta_2$ half from the $\alpha_2\beta_1$ half (b). At a higher resolution each chain is its own community.

Predicted allosteric networks in β-trypsin

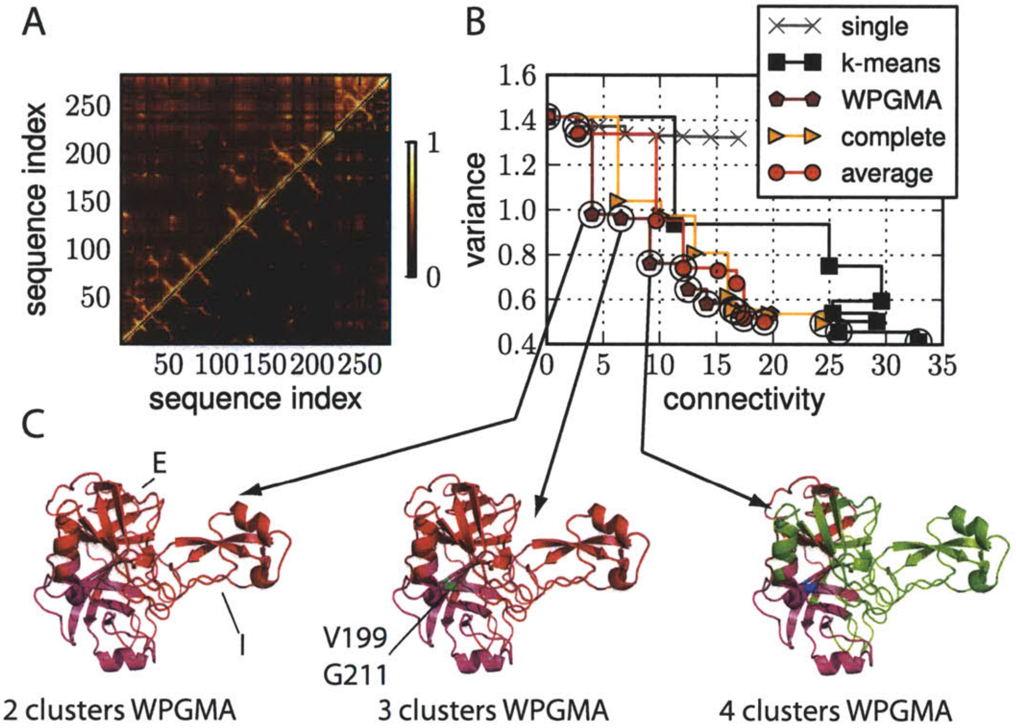

Figure B-5: Pareto-optimal dynamics-based allosteric network in β-trypsin. The different between the 2 and 3 clusters solutions is a small connectivity penalty due V199 and G211 forming a distinct cluster.

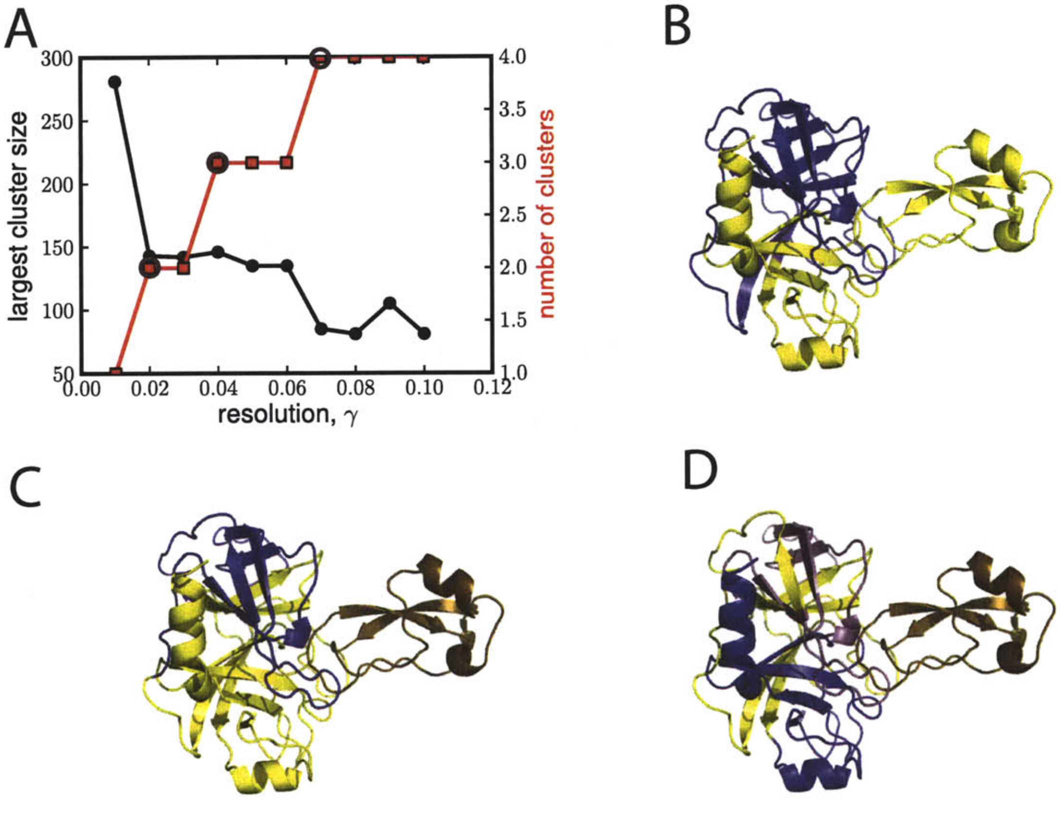

Figure B-6: Correlated dynamical networks in β-trypsin+inhibitor inferred by community detection. At the lowest resolution β-trypsin divides into two communities, one which includes the inhibitor (B). At higher resolutions the structure splits with the inhibitor forming its own community (C), followed by further partitioning of β-trypsin at the highest resolution tested (D).

Appendix C

Supporting materials for Chapter 5

Molecular mechanics benchmarks

We compared our continuum model against a molecular mechanics (MM) approach via Normal Mode Analysis (NMA). NMA is a natural choice for comparing models because the approach is designed to capture the fundamental global motion of a structured from Newton's Second law. We computed equilibrium thermal fluctuations using FEM and MM and compared the results. Note that our approach is not equivalent to validating our model against experiments, and we do not claim the MM approach is representative of experimental thermal fluctuations observed via spectroscopy or X-ray crystallography. Our validation only tests whether or not the FEM framework captures the same physics as the broadly accepted MM approach.

An open question asks how well does the FEM match other methods, and when does the FEM approach fail to capture atomistic detail. Bathe has shown FEM performs as well or better ENM, RTB, and all atom models for predicting the fluctuation of α-carbon via normal mode analysis (NMA) of globular proteins [16]. Here we extend Bathe's analysis to multi-domain proteins. We do limit the scope of our analysis to systems where our molecules can be simulated by both FEM and a atomistic methods that are more computationally expensive. To that end we compare the FEM approach to RTB for the case of a F-actin dimer. We choose this system because it is computationally feasible to solve it with RTB and its close relation to the models we build in Chap. 5. Moreover, the dimer includes an interface, which we hypothesized contains important atomistic details that a continuum model could fail to capture.

We follow a common approach for evaluating new molecular models by comparing the residue thermal fluctuation predicted from NMA [164, 12, 16]. We start with the thermal fluctuation tensor of an atom number i, $\mathbf{C}_i = \langle \mathbf{x}_i \mathbf{x}_i^T \rangle$, where the averages are computed in the standard way from normal mode-based fluctuation vectors, $\mathbf{x}_{ik}$, detailed in [26].

$$\langle \mathbf{x}_i \mathbf{x}_i^T \rangle = k_B T \sum_{k=7}^{k=3N} \frac{\mathbf{x}_{ik} \mathbf{x}_{ik}^T}{\omega_k^2}$$

From the fluctuation tensor we compute the root means square fluctuation (RMSF) of the atom from the trace, $\sqrt{tr\left(\mathbf{C}\right)}$. Physically, the RMSF is a scalar representation of the range of motion on an atom in the molecule.

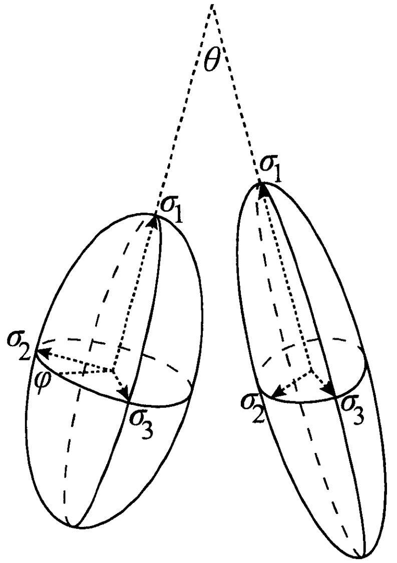

Figure C-1: The anisotropy of fluctuating atoms described by an analogy to an ellipsoid (inspired by [181]). The three radii of the ellipsoid represent the magnitude of the principal fluctuations, $\sigma_1 \geq \sigma_2 \geq \sigma_3$. The principal fluctuations determine the anisotropy scalar, A_1. The three axes of the ellipsoid correspond to the principal directions of the fluctuations. The angle between principal directions of two atoms defines the fluctuation misdirection.

In addition to magnitude of the fluctuations, we are also interested in any potential bias in the directionality, which we refer to as the anisotropy. To quantify anisotropy we first remove the dependence of our arbitrarily chosen coordinates system on **C** by diagonalizing it to obtain three principal fluctuations $\{\sigma_1^2, \sigma_2^2, \sigma_3^2\}$. From the principal fluctuations we compute a measure of anisotropy of the fluctuations from [80].

$$A_1 = \sqrt{\frac{\sigma_1^2}{1/2\,(\sigma_2^2 + \sigma_3^2)}} - 1$$

The quantity A_1 equals 0 for an atom that fluctuates in the first principal direction as much as the other two on average. Larger values of A_1 correspond to atoms that fluctuate predominantly in the first principal direction Finally, to characterize the ability of FEM to capture the direction of the anisotropy fluctuations predicted by the RTB approach we defined the anisotropy misdirection as the angle between the RTB and FEM principal fluctuation directions, θ_1.

We generated the FEM model of the F-actin dimer from the Oda structural model [124] (PDB ID 2ZWH). The molecular surface was computed using PyMOL's `surf` routine [40] and the tetrahedral mesh was derived from the computer program TetGen [154]. We solved for 200 normal modes using the commercial FEM software ADINA. The same atomic coordinates were then used to compute the RTB solution. First we minimized the crystal structure using successive rounds of minimization with decreasing harmonic restraints on the heavy atoms until the unrestrained energy gradient was less that 1×10^{-4} kcal/mol/Å. Next we computed 200 non-degenerate normal modes using the molecular mechanics program CHARMM [25, 27] for comparison with the FEM solution.

C.0.1 Validation of finite element framework

Figure C-2 shows that the FEM approach can match RTB in terms of thermal fluctuations. The RMSF of the α-carbons from the two approaches overlap in general (Fig. C-2A). Moreover, the RMSF correlates both within the protomers and at the interface between protomers (Fig. C-2B). The deviation between FEM and RTB among α-carbons with greatest RMSF shows up as spread in the scatter above 3 Å.

The anisotropy of the thermal fluctuations A_1 is also captured by the FEM procedure (Fig.C-2C-D). The FEM approach underestimates the most anisotropic fluctuations (Fig.C-2C), but the FEM and RTB approaches still correlate positively (Fig.C-2D), although not as much as the RMSF. Similarly the FEM approach capture the RTB-predicted anisotropy direction (Fig.C-2E-F).

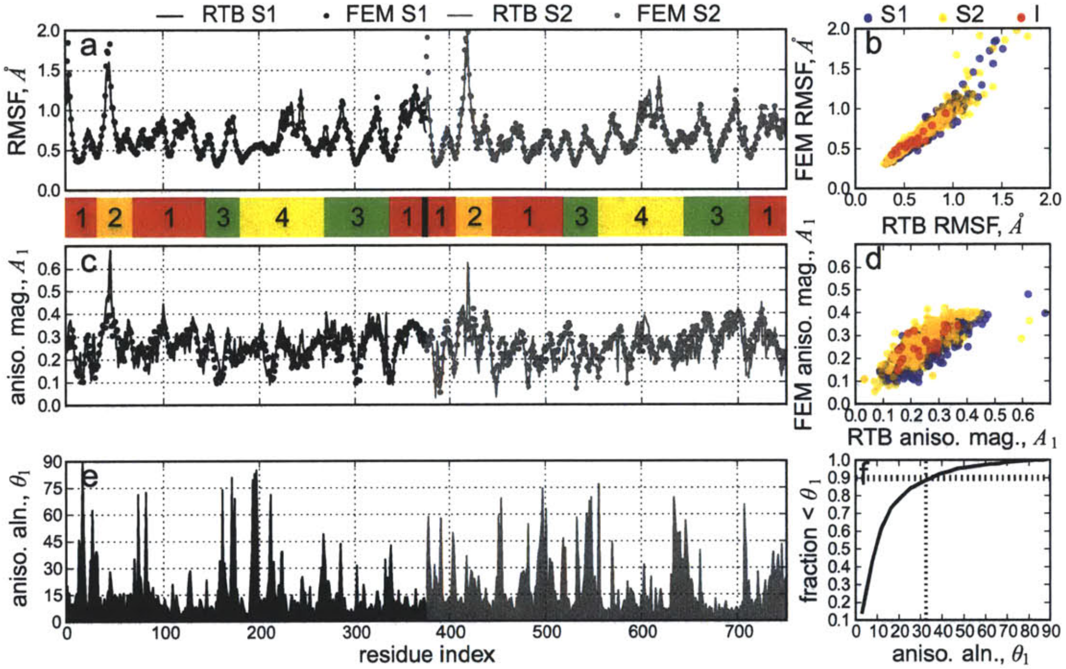

Figure C-2: Correspondence between computationally predicted thermal fluctuations from the continuum (FEM) or atomistic (RTB) models of the Oda *et al.* F-actin dimer model. The root mean squared fluctuations as a function of sequence from the FEM model (dots) predict the RTB results (solid line) (a). The subdomain numbering is shown schematically below the abscissa. The Pearson correlation coefficient is equal to 0.96, 0.95, and 0.91 for the subset F-actin dimer residues belonging to subunit 1 (blue), subunit 2 (yellow), or at the interface (red), respectively (b). The A_1 anisotropy of the fluctuations also positively correlate (c-d) with a Pearson correlation coefficients equal to 0.78, 0.76, and 0.64 for subunit 1, subunit 2, and the interface, respectively. The fluctuations predicted by FEM or RTB are typically parallel (e), with 90% of the C_α principal directions at least 34° from parallel (f)

Although we found good agreement between the FEM and RTB models when tested on an F-actin dimer, we had no data to show the agreement would be as good or better with larger molecules studied in this project. We therefore tested the hypothesis that the correspondence between FEM and RTB improves with the size of the molecule. To this end we created models of the Oda filament comprising 1, 3, or 12 actin subunits. We chose the 200 modes for computing the monomer fluctuations and computed the contribution of the 200th mode to the fluctuations to be about 0.01 Å. Larger molecules required fewer modes for the fluctuation series to converge to the same tolerance. The agreement of the RTB and FEM anisotropy metrics improve with the size of the molecule (Fig.C-3).

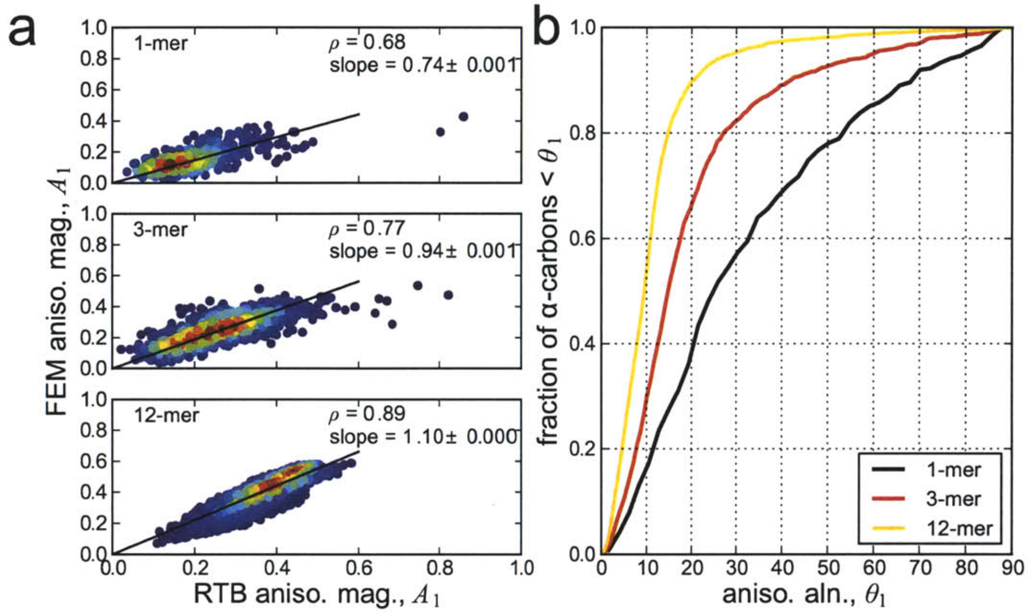

Figure C-3: The continuum approximation to the molecular mechanics model converges with increasing size of the simulated macromolecule. The correlation between FEM and RTB anisotropy magnitudes A_1 increases with molecular size (a). Shown are scatter plots of α-carbon fluctuation anisotropy A_1 computed by RTB (abscissa) or FEM (ordinated). The color of the dots reflect the density of data points at the coordinate. The correlation from monomer to trimer to dodecamer increase from 0.68 to 0.77 to 0.89. The scatter is fit with a linear function $A_1^{FEM}(A_1^{RTB}) = \text{slope} \times A_1^{RTB} + N(0, \sigma)$ where we assume the slope is random variable from a normal distribution and N is unbiased Gaussian noise with $\sigma \sim \Gamma(\alpha = \beta = 1)$. The direction of the FEM principal fluctuations tend to align with the RTB principal fluctuation direction with increasing molecular size too (b). The fraction of α-carbons that are severely misaligned decreases from monomer to dimer to dodecamer.

We hypothesized that FEM works as well as RTB at predicting anisotropy because anisotropy is predominantly a simple function of geometry. To test this we asked whether the location of the atoms in the molecule is predictive of the magnitude of the fluctuation anisotropy. By plotting fluctuation anisotropy A_1 versus distance from the center of geometry we identified a piecewise linear function with a positive slope (Fig. C-4), indicating the location of residues in the protein is predictive of anisotropy.

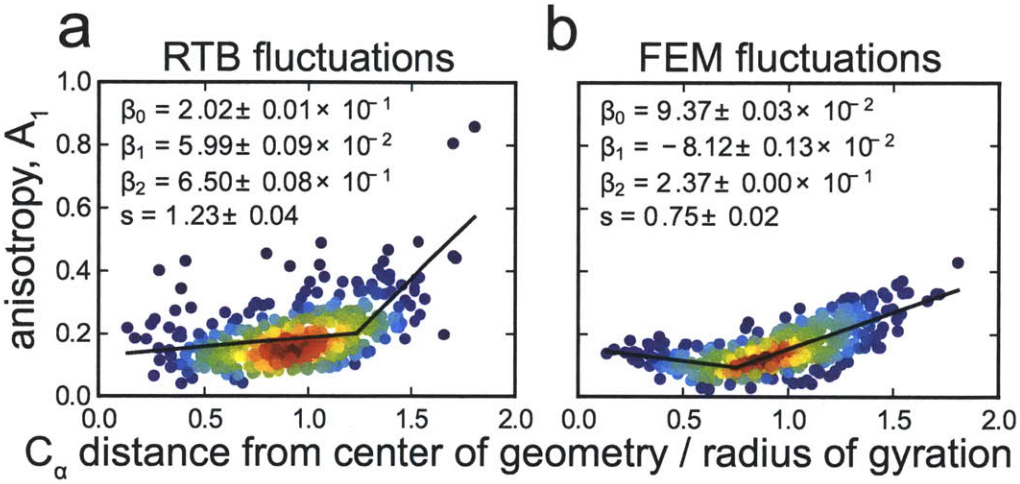

Figure C-4: The anisotropy magnitude of α-carbon fluctuations is weakly dependent on the distance from the center of geometry of the molecule. The scatter plots show that RTB (a) or FEM (b) fluctuation anisotropy increases with normalized (arbitrarily) distance from center of geometry. The color of the dots represent the density of data points at the coordinate. The data are from NMA of an Oda model monomer.

Geometry of F-actin models

Structure	PDB ID	Shift Å	Twist (degrees)
Mode 1	N/A	27.53	166.64
Mode 2	N/A	27.60	166.6
Mode 3	N/A	27.60	166.6
Mode 4	N/A	27.60	166.6
Mode 5	N/A	27.60	166.67
α-actinin	3LUE	26.6	167.2
Fimbrin	3BYH	27.30	166.5
Oda	2ZWH	27.59	166.4
Fujii	3MFP	27.6	166.7

Table C.1: Geometric properties of F-actin models.

Bibliography

[1] EMDB statistics. http://www.ebi.ac.uk/pdbe/emdb/statistics.html.

[2] Genome.gov | human genome project completion: Frequently asked questions. http://www.genome.gov/11006943.

[3] RCSB PDB. http://www.pdb.org.

[4] Summary of NDA approvals & receipts, 1938 to the present. http://www.fda.gov/.

[5] UniProtKB/TrEMBL release statistics | UniProt | the universal protein resource | EBI. http://www.ebi.ac.uk/uniprot/TrEMBLstats/.

[6] Monika Abedin and Nicole King. The premetazoan ancestry of cadherins. *Science*, 319(5865):946–948, February 2008.

[7] N.A. Ahmed and D.V. Gokhale. Entropy expressions and their estimators for multivariate distributions. *Information Theory, IEEE Transactions on*, 35(3):688–692, 1989.

[8] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997. PMID: 9254694.

[9] T. Angelini, H. Liang, W. Wriggers, and G. Wong. Direct observation of counterion organization in f-actin polyelectrolyte bundles. *The European Physical Journal E: Soft Matter and Biological Physics*, 16(4):389–400, April 2005.

[10] Thomas E Angelini, Hongjun Liang, Willy Wriggers, and Gerard C L Wong. Like-charge attraction between polyelectrolytes induced by counterion charge density waves. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15):8634–8637, July 2003. PMID: 12853566.

[11] William R. Atchley, Kurt R. Wollenberg, Walter M. Fitch, Werner Terhalle, and Andreas W. Dress. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Molecular Biology and Evolution*, 17(1):164 –178, January 2000.

[12] A Atilgan. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, January 2001.

[13] Christopher H S Aylett, Qing Wang, Katharine A Michie, Linda A Amos, and Jan Löwe. Filament structure of bacterial tubulin homologue TubZ. *Proceedings of the National Academy of Sciences of the United States of America*, October 2010. PMID: 20974911.

[14] Barabasi and Albert. Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–512, October 1999. PMID: 10521342.

[15] James R Bartles. Parallel actin bundles and their multiple actin-bundling proteins. *Current Opinion in Cell Biology*, 12(1):72–78, February 2000.

[16] Mark Bathe. A finite element framework for computation of protein normal modes and mechanical response. *Proteins*, 70(4):1595–1609, March 2008. PMID: 17975833.

[17] Mark Bathe, Claus Heussinger, Mireille M A E Claessens, Andreas R Bausch, and Erwin Frey. Cytoskeletal bundle mechanics. *Biophysical Journal*, 94(8):2955–2964, April 2008. PMID: 18055529.

[18] Werner Baumgartner, Markus W Wendeler, Agnes Weth, Rainer Koob, Detlev Drenckhahn, and Reinhard Gessner. Heterotypic trans-interaction of LI- and e-cadherin and their localization in plasmalemmal microdomains. *Journal of Molecular Biology*, 378(1):44–54, April 2008. PMID: 18342884.

[19] D ben-Avraham and M M Tirion. Dynamic and elastic properties of f-actin: a normal-modes analysis. *Biophysical Journal*, 68(4):1231–1245, April 1995. PMID: 7787015 PMCID: 1282021.

[20] D Berndorff, R Gessner, B Kreft, N Schnoy, A M Lajous-Petter, N Loch, W Reutter, M Hortsch, and R Tauber. Liver-intestine cadherin: molecular cloning and characterization of a novel ca(2+)-dependent cell adhesion molecule expressed in liver and intestine. *The Journal of Cell Biology*, 125(6):1353–1369, June 1994. PMID: 8207063.

[21] Blatt, Wiseman, and Domany. Superparamagnetic clustering of data. *Physical Review Letters*, 76(18):3251–3254, April 1996. PMID: 10060920.

[22] Titus J Boggon, John Murray, Sophie Chappuis-Flament, Ellen Wong, Barry M Gumbiner, and Lawrence Shapiro. C-cadherin ectodomain structure and implications for cell adhesion mechanisms. *Science (New York, N.Y.)*, 296(5571):1308–1313, May 2002. PMID: 11964443.

[23] Itamar Borukhov, Robijn F Bruinsma, William M Gelbart, and Andrea J Liu. Structural polymorphism of the cytoskeleton: a model of linker-assisted filament aggregation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3673–3678, March 2005. PMID: 15731355.

[24] E. A. Boucher and R. M. Nisbet. Application of the transfer-matrix method to equilibria of polymer reactions and adsorption from solution. *Chemical Physics Letters*, 40(1):61–65, May 1976.

[25] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[26] Bernard R. Brooks, Dušanka Janežič, and Martin Karplus. Harmonic analysis of large systems. i. methodology. *Journal of Computational Chemistry*, 16(12):1522–1542, 1995.

[27] BR Brooks, CL Brooks, AD Mackerell, L Nilsson, RJ Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, A Caflisch, L Caves, Q Cui, AR Dinner, M Feig, S Fischer, J Gao, M Hodoscek, W Im, K Kuczera, T Lazaridis, J Ma, V Ovchinnikov, E Paci, RW Pastor, CB Post, JZ Pu, M Schaefer, B Tidor, RM Venable, HL Woodcock, X Wu, W Yang, DM York, and M Karplus. CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, July 2009.

[28] Cristina Marino Buslje, Javier Santos, Jose Maria Delfino, and Morten Nielsen. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, 25(9):1125 –1131, May 2009.

[29] M L Cano, L Cassimeris, M Fechheimer, and S H Zigmond. Mechanisms responsible for f-actin stabilization after lysis of polymorphonuclear leukocytes. *The Journal of Cell Biology*, 116(5):1123–1134, March 1992. PMID: 1740469.

[30] Bin Chen, Anli Li, Dennis Wang, Min Wang, Lili Zheng, and James R. Bartles. Espin contains an additional actin-binding site in its n terminus and is a major actin-bundling protein of the sertoli Cell–Spermatid ectoplasmic specialization junctional plaque. *Molecular Biology of the Cell*, 10(12):4327–4339, December 1999. PMID: 10588661 PMCID: 25761.

[31] Chien Peter Chen, Shoshana Posy, Avinoam Ben-Shaul, Lawrence Shapiro, and Barry H Honig. Specificity of cell-cell adhesion by classical cadherins: Critical role for low-affinity dimerization through beta-strand swapping. *Proceedings of the National Academy of Sciences of the United States of America*, 102(24):8531–8536, June 2005. PMID: 15937105.

[32] Dongmei Cheng, Joyce Marner, and Peter A. Rubenstein. Interaction in vivo and in vitro between the yeast fimbrin, SAC6P, and a polymerization-defective yeast actin (V266G and L267G). *Journal of Biological Chemistry*, 274(50):35873 –35880, December 1999.

[33] Jhih-Wei Chu and Gregory A. Voth. Allostery of actin filaments: Molecular dynamics simulations and coarse-grained analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13111 –13116, 2005.

[34] Paulo Cignoni. MeshLab, version 1.3.0b. May 2010.

[35] M M A E Claessens, C Semmrich, L Ramos, and A R Bausch. Helical twist controls the thickness of f-actin bundles. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26):8819–8822, July 2008. PMID: 18579789.

[36] Mireille M. A. E. Claessens, Mark Bathe, Erwin Frey, and Andreas R. Bausch. Actin-binding proteins sensitively mediate f-actin bundle stiffness. *Nat Mater*, 5(9):748–753, 2006.

[37] Enrique M De La Cruz, Jeremy Roland, Brannon R McCullough, Laurent Blanchoin, and Jean-Louis Martiel. Origin of twist-bend coupling in actin filaments. *Biophysical Journal*, 99(6):1852–1860, September 2010. PMID: 20858430.

[38] Fabien M Decaillot, Katia Befort, Dominique Filliol, ShiYi Yue, Philippe Walker, and Brigitte L Kieffer. Opioid receptor random mutagenesis reveals a mechanism for g protein-coupled receptor activation. *Nat Struct Mol Biol*, 10(8):629–636, 2003.

[39] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol*, 2, May 2006.

[40] Warren DeLano. PyMOL molecular graphics system, 2009.

[41] Yi der Chen. Binding of n-mers to one-dimensional lattices with longer than close-contact interactions. *Biophysical Chemistry*, 27(1):59–65, July 1987.

[42] D J DeRosier and L G Tilney. How actin filaments pack into bundles. *Cold Spring Harbor Symposia on Quantitative Biology*, 46 Pt 2:525–540, 1982. PMID: 6955098.

[43] Thomas Glen Dietterich, Suzanna Becker, and Zoubin Ghahramani. *Advances in neural information processing systems 14: proceedings of the 2002 conference.* MIT Press, 2002.

[44] C. G. Dos Remedios, D. Chhabra, M. Kekic, I. V. Dedova, M. Tsubakihara, D. A. Berry, and N. J. Nosworthy. Actin binding proteins: Regulation of cytoskeletal microfilaments. *Physiological Reviews*, 83(2):433 –473, April 2003.

[45] Duke Duguay, Ramsey A Foty, and Malcolm S Steinberg. Cadherin-mediated cell adhesion and tissue segregation: qualitative and quantitative determinants. *Developmental Biology*, 253(2):309–323, January 2003. PMID: 12645933.

[46] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333 –340, February 2008.

[47] Joe Felsenstein. PHYLIP (Phylogeny inference package) version 3.6, 2005.

[48] Andrew D. Fernandes and Gregory B. Gloor. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, 26(9):1135 –1139, May 2010.

[49] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman. The pfam protein families database. *Nucleic Acids Research*, 38(Database):D211–D222, November 2009.

[50] Daniel A. Fletcher and R. Dyche Mullins. Cell mechanics and the cytoskeleton. *Nature*, 463(7280):485–492, January 2010.

[51] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Bert Overduin, Bethan Pritchard, Harpreet Singh Riat, Daniel Rios, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Giulietta Spudich, Y Amy Tang, Stephen Trevanion, Jana Vandrovcova, Albert J Vilella, Simon White, Steven P Wilder, Amonida Zadissa, Jorge Zamora, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, Jan Vogel, and Stephen M J Searle. Ensembl 2011. *Nucleic Acids Research*, 39(Database issue):D800–806, January 2011. PMID: 21045057.

[52] Jessica H Fong, Amy E Keating, and Mona Singh. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biology*, 5(2):R11, 2004. PMID: 14759261.

[53] Ramsey A Foty and Malcolm S Steinberg. The differential adhesion hypothesis: a direct evaluation. *Developmental Biology*, 278(1):255–263, February 2005. PMID: 15649477.

[54] Takashi Fujii, Atsuko H Iwane, Toshio Yanagida, and Keiichi Namba. Direct visualization of secondary structures of f-actin by electron cryomicroscopy. *Nature*, 467(7316):724–728, October 2010. PMID: 20844487.

[55] Yuan-cheng Fung and Pin Tong. *Classical and computational solid mechanics.* World Scientific, 2001.

[56] Vitold E Galkin, Albina Orlova, William Brieher, Hao Yuan Kueh, Timothy J Mitchison, and Edward H Egelman. Coronin-1A stabilizes f-actin by bridging adjacent actin protomers and stapling opposite strands of the actin filament. *Journal of Molecular Biology*, 376(3):607–613, February 2008. PMID: 18177666.

[57] Vitold E. Galkin, Albina Orlova, Olga Cherepanova, Marie-Christine Lebart, and Edward H. Egelman. High-resolution cryo-EM structure of the f-actin–fimbrin/plastin ABD2 complex. *Proceedings of the National Academy of Sciences*, 105(5):1494–1498, February 2008.

[58] Vitold E Galkin, Albina Orlova, Chris Rivera, R Dyche Mullins, and Edward H Egelman. Structural polymorphism of the ParM filament and dynamic instability. *Structure (London, England: 1993)*, 17(9):1253–1264, September 2009. PMID: 19748346.

[59] Vitold E Galkin, Albina Orlova, Anita Salmazo, Kristina Djinovic-Carugo, and Edward H Egelman. Opening of tandem calponin homology domains regulates their affinity for f-actin. *Nature Structural & Molecular Biology*, 17(5):614–616, May 2010. PMID: 20383143.

[60] Vitold E Galkin, Albina Orlova, Gunnar F Schroder, and Edward H Egelman. Structural polymorphism in f-actin. *Nat Struct Mol Biol*, 17(11):1318–1323, November 2010.

[61] M. L. Gardel, J. H. Shin, F. C. MacKintosh, L. Mahadevan, P. Matsudaira, and D. A. Weitz. Elastic behavior of Cross-Linked and bundled actin networks. *Science*, 304(5675):1301 –1305, May 2004.

[62] F Gittes, B Mickey, J Nettleton, and J Howard. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *The Journal of Cell Biology*, 120(4):923–934, February 1993. PMID: 8432732.

[63] Robert M. Glaeser and Richard J. Hall. Reaching the information limit in Cryo-EM of biological macromolecules: Experimental aspects. *Biophysical Journal*, 100(10):2331–2337, May 2011.

[64] J R Glenney, P Kaulfus, P Matsudaira, and K Weber. F-actin binding and bundling properties of fimbrin, a major cytoskeletal protein of microvillus core filaments. *Journal of Biological Chemistry*, 256(17):9283 –9288, 1981.

[65] M J Greenberg, C-L A Wang, W Lehman, and J R Moore. Modulation of actin mechanics by caldesmon and tropomyosin. *Cell Motility and the Cytoskeleton*, 65(2):156–164, February 2008. PMID: 18000881.

[66] Julia Handl, Joshua Knowles, and Douglas B Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics (Oxford, England)*, 21(15):3201–3212, August 2005. PMID: 15914541.

[67] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27:83–85, 2005. 10.1007/BF02985802.

[68] Daniel Häussinger, Thomas Ahrens, Thomas Aberle, Jürgen Engel, Jörg Stetefeld, and Stephan Grzesiek. Proteolytic e-cadherin activation followed by solution NMR and x-ray crystallography. *The EMBO Journal*, 23(8):1699–1708, April 2004. PMID: 15071499.

[69] Daniel Häussinger, Thomas Ahrens, Hans-Jürgen Sass, Olivier Pertz, Jürgen Engel, and Stephan Grzesiek. Calcium-dependent homoassociation of e-cadherin by NMR spectroscopy: changes in mobility, conformation and mapping of contact regions. *Journal of Molecular Biology*, 324(4):823–839, December 2002. PMID: 12460580.

[70] Tapio Heimo, Jussi M Kumpula, Kimmo Kaski, and Jari Saramäki. Detecting modules in dense weighted networks with the potts method. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08007, 2008.

[71] Judith Herzfeld. Crowding-induced organization in cells: spontaneous alignment and sorting of filaments with physiological control points. *Journal of Molecular Recognition: JMR*, 17(5):376–381, October 2004. PMID: 15362095.

[72] Claus Heussinger, Felix Schller, and Erwin Frey. Statics and dynamics of the wormlike bundle model. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 81(2 Pt 1):021904, February 2010. PMID: 20365592.

[73] Gábor Hild, Beáta Bugyi, and Miklós Nyitrai. Conformational dynamics of actin: effectors and implications for biological function. *Cytoskeleton (Hoboken, N.J.)*, 67(10):609–629, October 2010. PMID: 20672362.

[74] T L Hill, E Eisenberg, and L Greene. Theoretical model for the cooperative equilibrium binding of myosin subfragment 1 to the actin-troponin-tropomyosin complex. *Proceedings of the National Academy of Sciences of the United States of America*, 77(6):3186–3190, June 1980. PMID: 10627230.

[75] T L Hill, E Eisenberg, and L E Greene. Alternate model for the cooperative equilibrium binding of myosin subfragment-1-nucleotide complex to actin-troponin-tropomyosin. *Proceedings of the National Academy of Sciences of the United States of America*, 80(1):60–64, January 1983. PMID: 6572009.

[76] Kenneth C. Holmes, Isabel Angert, F. Jon Kull, Werner Jahn, and Rasmus R. Schroder. Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature*, 425(6956):423–427, 2003.

[77] Kenneth C. Holmes, David Popp, Werner Gebhard, and Wolfgang Kabsch. Atomic model of the actin filament. *Nature*, 347(6288):44–49, 1990.

[78] M. Hosek. Polymer-induced bundling of f actin and the depletion force. *Physical Review E*, 69(5), 2004.

[79] Paco Hulpiau and Frans van Roy. Molecular evolution of the cadherin superfamily. *The International Journal of Biochemistry & Cell Biology*, 41(2):349–369, February 2009.

[80] T Ichiye and M Karplus. Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation. *Proteins*, 2(3):236–259, 1987. PMID: 3447180.

[81] N W Ikebuchi and D M Waisman. Calcium-dependent regulation of actin filament bundling by lipocortin-85. *The Journal of Biological Chemistry*, 265(6):3392–3400, February 1990. PMID: 2137457.

[82] H Inuzuka, S Miyatani, and M Takeichi. R-cadherin: a novel ca(2+)-dependent cell-cell adhesion molecule expressed in the retina. *Neuron*, 7(1):69–79, July 1991. PMID: 1712604.

[83] H Isambert, P Venier, A C Maggs, A Fattoum, R Kassab, D Pantaloni, and M F Carlier. Flexibility of actin filaments derived from thermal fluctuations. effect of bound nucleotide, phalloidin, and muscle regulatory proteins. *The Journal of Biological Chemistry*, 270(19):11437–11444, May 1995. PMID: 7744781.

[84] Jr Glenney J R, N Geisler, P Kaulfus, and K Weber. Demonstration of at least two different actin-binding sites in villin, a calcium-regulated modulator of f-actin organization. *The Journal of Biological Chemistry*, 256(15):8156–8161, August 1981. PMID: 6790532.

[85] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, May 1957.

[86] D T Jones, W R Taylor, and J M Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences: CABIOS*, 8(3):275–282, June 1992. PMID: 1633570.

[87] W Kabsch, H G Mannherz, D Suck, E F Pai, and K C Holmes. Atomic structure of the actin:DNase i complex. *Nature*, 347(6288):37–44, September 1990. PMID: 2395459.

[88] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogenbonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[89] Bryan J Katafiasz, Marvin T Nieman, Margaret J Wheelock, and Keith R Johnson. Characterization of cadherin-24, a novel alternatively spliced type II cadherin. *The Journal of Biological Chemistry*, 278(30):27513–27519, July 2003. PMID: 12734196.

[90] Kazutaka Katoh, Kei ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518, 2005. PMID: 15661851.

[91] P Katsamba, K Carroll, G Ahlsen, F Bahna, J Vendome, S Posy, M Rajebhosale, S Price, T M Jessell, A Ben-Shaul, L Shapiro, and Barry H Honig. Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28):11594–11599, July 2009. PMID: 19553217.

[92] D.-N. Kim, J. Altschuler, C. Strong, G. McGill, and M. Bathe. Conformational dynamics data bank: a database for conformational dynamics of proteins and supramolecular protein assemblies. *Nucleic Acids Research*, 39(Database):D451–D455, November 2010.

[93] Do-Nyun Kim, Cong-Tri Nguyen, and Mark Bathe. Conformational dynamics of supramolecular protein assemblies. *Journal of Structural Biology*, 173(2):261–270, February 2011. PMID: 20854912.

[94] Taeyoon Kim, Wonmuk Hwang, Hyungsuk Lee, and Roger D. Kamm. Computational analysis of viscoelastic properties of crosslinked actin networks. *PLoS Comput Biol*, 5(7):e1000439, July 2009.

[95] Michael G Klein, Wuxian Shi, Udupi Ramagopal, Yiider Tseng, Denis Wirtz, David R Kovar, Christopher J Staiger, and Steven C Almo. Structure of the actin crosslinking core of fimbrin. *Structure (London, England: 1993)*, 12(6):999–1013, June 2004. PMID: 15274920.

[96] Jonathan Knight. Physics meets biology: Bridging the culture gap. *Nature*, 419(6904):244–246, 2002.

[97] H Kojima, A Ishijima, and T Yanagida. Direct measurement of stiffness of single actin filaments with and without tropomyosin by in vitro nanomanipulation. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12962–12966, December 1994. PMID: 7809155.

[98] Y Kong. Ligand binding on ladder lattices. *Biophysical Chemistry*, 81(1):7–21, September 1999. PMID: 17030328.

[99] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, October 2008.

[100] L D Landau and E.M. Lifshitz. *Mechanics, Third Edition: Volume 1*. Butterworth-Heinemann, 3 edition, January 1976.

[101] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris,

A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, and J Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. PMID: 11237011.

[102] Oliver F Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins*, 62(4):1053–1061, March 2006. PMID: 16355416.

[103] Pekka Lappalainen and David G. Drubin. Cofilin promotes rapid actin filament

turnover in vivo. *Nature*, 388(6637):78–82, July 1997.

[104] M.-C. Lebart, F. Hubert, C. Boiteau, S. Ventéo, C. Roustan, and Y. Benyamin. Biochemical characterization of the L-Plastin-Actin interaction shows a resemblance with that of ?-Actinin and allows a distinction to be made between the two Actin-Binding domains of the molecule†. *Biochemistry*, 43(9):2428–2437, March 2004.

[105] L Levi, J Douek, M Osman, T C Bosch, and B Rinkevich. Cloning and characterization of BS-cadherin, a novel cadherin from the colonial urochordate botryllus schlosseri. *Gene*, 200(1-2):117–123, October 1997. PMID: 9373145.

[106] Guohui Li and Qiang Cui. A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Biophysical Journal*, 83(5):2457–2474, November 2002. PMID: 12414680.

[107] Xiaochuan (Edward) Li, Larry S. Tobacman, Ji Young Mun, Roger Craig, Stefan Fischer, and William Lehman. Tropomyosin position on F-Actin revealed by EM reconstruction and computational chemistry. *Biophysical Journal*, 100(4):1005–1013, February 2011.

[108] Xiumei Liu and Gerald H Pollack. Mechanics of f-actin characterized with microfabricated cantilevers. *Biophysical Journal*, 83(5):2705–2715, November 2002. PMID: 12414703 PMCID: 1302355.

[109] S W Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (New York, N.Y.)*, 286(5438):295–299, October 1999. PMID: 10514373.

[110] Pter Lnrt, Christian P Bacher, Nathalie Daigle, Arthur R Hand, Roland Eils, Mark Terasaki, and Jan Ellenberg. A contractile nuclear actin network drives chromosome congression in oocytes. *Nature*, 436(7052):812–818, August 2005. PMID: 16015286.

[111] L C Martin, G B Gloor, S D Dunn, and L M Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford, England)*, 21(22):4116–4124, November 2005. PMID: 16159918.

[112] Atsushi Matsumoto and Wilma K Olson. Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophysical Journal*, 83(1):22–41, July 2002. PMID: 12080098 PMCID: 1302125.

[113] Brannon R McCullough, Laurent Blanchoin, Jean-Louis Martiel, and Enrique M De la Cruz. Cofilin increases the bending flexibility of actin filaments: implications for severing and cell mechanics. *Journal of Molecular Biology*, 381(3):550–558, September 2008. PMID: 18617188.

[114] Amy McGough, Brian Pope, Wah Chiu, and Alan Weeds. Cofilin changes the twist of F-Actin: implications for actin filament dynamics and cellular function. *The Journal of Cell Biology*, 138(4):771–781, August 1997. PMID: 9265645 PMCID: 2138052.

[115] R K Meyer and U Aebi. Bundling of actin filaments by alpha-actinin depends on its molecular length. *The Journal of Cell Biology*, 110(6):2013–2024, June 1990. PMID: 2351691.

[116] Tim Meyer, Marco D'Abramo, Adam Hospital, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluis Gelpí, and Modesto Orozco. MoDEL (Molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. *Structure*, 18(11):1399–1409, November 2010.

[117] Vesselin Z Miloushev, Fabiana Bahna, Carlo Ciatto, Goran Ahlsen, Barry Honig, Lawrence Shapiro, and Arthur G Palmer. Dynamic properties of a type II cadherin adhesive domain: implications for the mechanism of strand-swapping of classical cadherins. *Structure (London, England: 1993)*, 16(8):1195–1205, August 2008. PMID: 18682221.

[118] Leonid A Mirny and Mikhail S Gelfand. Using orthologous and paralogous proteins to identify specificity determining residues. *Genome Biology*, 3(3):PREPRINT0002, 2002. PMID: 11897020.

[119] O. Miyashita, J. N. Onuchic, and P. G. Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences*, 100(22):12570 –12575, October 2003.

[120] MEJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[121] Carien M. Niessen and Barry M. Gumbiner. Cadherin-mediated cell sorting not determined by binding or adhesion specificity. *J. Cell Biol.*, 156(2):389–400, January 2002.

[122] A Nose, K Tsuji, and M Takeichi. Localization of specificity determining sites in cadherin cell adhesion molecules. *Cell*, 61(1):147–155, April 1990. PMID: 2317870.

[123] Akinao Nose, Akira Nagafuchi, and Masatoshi Takeichi. Expressed recombinant cadherins mediate cell sorting in model systems. *Cell*, 54(7):993–1001, September 1988.

[124] Toshiro Oda, Mitsusada Iwasa, Tomoki Aihara, Yuichiro Maéda, and Akihiro Narita. The nature of the globular- to fibrous-actin transition. *Nature*, 457(7228):441–445, January 2009. PMID: 19158791.

[125] Toshiro Oda and Yuichiro Maéda. Multiple conformations of f-actin. *Structure (London, England: 1993)*, 18(7):761–767, July 2010. PMID: 20637412.

[126] S Ono, Y Yamakita, S Yamashiro, P T Matsudaira, J R Gnarra, T Obinata, and F Matsumura. Identification of an actin binding region and a protein kinase c phosphorylation site on human fascin. *The Journal of Biological Chemistry*, 272(4):2527–2533, January 1997. PMID: 8999969.

[127] A. Orlova and E. H. Egelman. A conformational change in the actin subunit can change the flexibility of the actin filament. *Journal of Molecular Biology*, 232(2):334–341, July 1993.

[128] Albina Orlova, Vitold E Galkin, Margaret S VanLoock, Eldar Kim, Alexander Shvetsov, Emil Reisler, and Edward H Egelman. Probing the structure of f-actin: cross-links constrain atomic models and modify actin dynamics. *Journal of Molecular Biology*, 312(1):95–106, September 2001.

[129] M Ozawa and R Kemler. Correct proteolytic cleavage is required for the cell adhesive function of uvomorulin. *The Journal of Cell Biology*, 111(4):1645 – 1650, October 1990.

[130] Porntula Panorchan, Melissa S Thompson, Kelly J Davis, Yiider Tseng, Konstantinos Konstantopoulos, and Denis Wirtz. Single-molecule analysis of cadherin-mediated cell-cell adhesion. *Journal of Cell Science*, 119(Pt 1):66–74, January 2006. PMID: 16371651.

[131] Saurabh D Patel, Carlo Ciatto, Chien Peter Chen, Fabiana Bahna, Manisha Rajebhosale, Natalie Arkus, Ira Schieren, Thomas M Jessell, Barry Honig, Stephen R Price, and Lawrence Shapiro. Type II cadherin ectodomain structures: implications for classical cadherin specificity. *Cell*, 124(6):1255–1268, March 2006. PMID: 16564015.

[132] Robert J Pelham and Fred Chang. Actin dynamics in the contractile ring during cytokinesis in fission yeast. *Nature*, 419(6902):82–86, September 2002. PMID: 12214236.

[133] Jim Pfaendtner, Edward Lyman, Thomas D. Pollard, and Gregory A. Voth. Structure and dynamics of the actin filament. *Journal of Molecular Biology*, 396(2):252–263, February 2010.

[134] Anastasia V Pivovarova, Sofia Yu Khaitlina, and Dmitrii I Levitsky. Specific cleavage of the DNase-I binding loop dramatically decreases the thermal stability of actin. *The FEBS Journal*, 277(18):3812–3822, September 2010. PMID: 20718862.

[135] Douglas Poland. DNA melting profiles from a matrix method. *Biopolymers*, 73(2):216–228, February 2004.

[136] Thomas D Pollard and Gary G Borisy. Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112(4):453–465, February 2003. PMID: 12600310.

[137] Shoshana Posy, Lawrence Shapiro, and Barry Honig. Sequence and structural determinants of strand swapping in cadherin domains: do all cadherins bind through the same adhesive interface? *Journal of Molecular Biology*, 378(4):954–968, May 2008. PMID: 18395225.

[138] A K Prakasam, V Maruthamuthu, and D E Leckband. Similarities between heterophilic and homophilic cadherin adhesion. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42):15434–15439, October 2006. PMID: 17023539.

[139] Alka Prasad, Huaying Zhao, John M Rutherford, Nicole Housley, Corey Nichols, and Susan Pedigo. Effect of linker segments on the stability of epithelial cadherin domain 2. *Proteins*, 62(1):111–121, January 2006. PMID: 16287100.

[140] Stephen R Price, Natalia V De Marco Garcia, Barbara Ranscht, and Thomas M Jessell. Regulation of motor neuron pool sorting by differential expression of type II cadherins. *Cell*, 109(2):205–216, April 2002.

[141] Ewa Prochniewicz, Neal Janson, David D Thomas, and Enrique M De la Cruz. Cofilin increases the torsional flexibility and dynamics of actin filaments. *Journal of Molecular Biology*, 353(5):990–1000, November 2005. PMID: 16213521.

[142] I Rayment, H M Holden, M Whittaker, C B Yohn, M Lorenz, K C Holmes, and R A Milligan. Structure of the actin-myosin complex and its implications for muscle contraction. *Science (New York, N.Y.)*, 261(5117):58–65, July 1993. PMID: 8316858.

[143] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, July 2006.

[144] Antonis Rokas. The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annual Review of Genetics*, 42:235–251, 2008. PMID: 18983257.

[145] Michel F Sanner, Arthur J Olson, and JeanClaude Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, March 1996.

[146] Damon Scoville, John D. Stamm, Christian Altenbach, Alexander Shvetsov, Kaveh Kokabi, Peter A. Rubenstein, Wayne L. Hubbell, and Emil Reisler. Effects of binding factors on structural elements in F-Actin†. *Biochemistry*, 48(2):370–378, January 2009.

[147] Reza Sharifi Sedeh, Mark Bathe, and KlausJrgen Bathe. The subspace iteration method in protein normal mode analysis. *Journal of Computational Chemistry*, 31(1):66–74, January 2010.

[148] Reza Sharifi Sedeh, Alexander A Fedorov, Elena V Fedorov, Shoichiro Ono, Fumio Matsumura, Steven C Almo, and Mark Bathe. Structure, evolutionary conservation, and conformational dynamics of homo sapiens fascin-1, an f-actin crosslinking protein. *Journal of Molecular Biology*, 400(3):589–604, July 2010. PMID: 20434460.

[149] CE Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[150] Y Shimoyama, T Shibata, M Kitajima, and S Hirohashi. Molecular cloning and characterization of a novel human classic cadherin homologous with mouse muscle cadherin. *The Journal of Biological Chemistry*, 273(16):10011–10018, April 1998. PMID: 9545347.

[151] Y Shimoyama, G Tsujimoto, M Kitajima, and M Natori. Identification of three human type-II classic cadherins and frequent heterophilic interactions between different subclasses of type-II classic cadherins. *The Biochemical Journal*, 349(Pt 1):159–167, July 2000. PMID: 10861224.

[152] Homin Shin, Kirstin R Purdy Drew, James R Bartles, Gerard C L Wong, and Gregory M Grason. Cooperativity and frustration in protein-mediated parallel actin bundles. *Physical Review Letters*, 103(23):238102, December 2009. PMID: 20366178.

[153] Homin Shin and Gregory M. Grason. Structural reorganization of parallel actin bundles by crosslinking proteins: Incommensurate states of twist. *Physical Review E*, 82(5):051919, November 2010.

[154] Hang Si. TetGen: a quality tetrahedral mesh generator and Three-Dimensional delaunay triangulator, December 2009.

[155] D A Smith and M A Geeves. Cooperative regulation of myosin-actin interactions by a continuous flexible chain II: actin-tropomyosin-troponin and regulation by calcium. *Biophysical Journal*, 84(5):3168–3180, May 2003. PMID: 12719246.

[156] Antonio Del Sol, Marcos J Araúzo-Bravo, Dolors Amoros, and Ruth Nussinov. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome Biology*, 8(5):R92, 2007. PMID: 17531094.

[157] Marcos Sotomayor and Klaus Schulten. The allosteric role of the ca2+ switch in adhesion and elasticity of c-cadherin. *Biophysical Journal*, 94(12):4621–4633, June 2008. PMID: 18326636.

[158] M S Steinberg. Mechanism of tissue reconstruction by dissociated cells. II. time-course of events. *Science (New York, N.Y.)*, 137:762–763, September 1962. PMID: 13916688.

[159] M S Steinberg. On the mechanism of tissue reconstruction by dissociated cells. i. population kinetics, differential adhesiveness. and the absence of directed migration. *Proceedings of the National Academy of Sciences of the United States of America*, 48:1577–1582, September 1962. PMID: 13916689.

[160] M S Steinberg. On the mechanism of tissue reconstruction by dissociated cells, III. free energy relations and the reorganization of fused, heternomic tissue fragments. *Proceedings of the National Academy of Sciences of the United States of America*, 48(10):1769–1776, October 1962. PMID: 16591009.

[161] Gurol M. Suel, Steve W. Lockless, Mark A. Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Mol Biol*, 10(1):59–69, January 2003.

[162] K Sugimoto, S Honda, T Yamamoto, T Ueki, M Monden, A Kaji, K Matsumoto, and T Nakamura. Molecular cloning and characterization of a newly identified member of the cadherin family, PB-cadherin. *The Journal of Biological Chemistry*, 271(19):11548–11556, May 1996. PMID: 8626716.

[163] M Takeichi. The cadherins: cell-cell adhesion molecules controlling animal morphogenesis. *Development (Cambridge, England)*, 102(4):639–655, April 1988. PMID: 3048970.

[164] F Tama, F X Gadea, O Marques, and Y H Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41(1):1–7, October 2000. PMID: 10944387.

[165] Jay X. Tang and Paul A. Janmey. The polyelectrolyte nature of f-actin and the mechanism of actin bundle formation. *Journal of Biological Chemistry*, 271(15):8556–8563, April 1996.

[166] Vladimir B. Teif. General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to OR operator of phage lambda. *Nucl. Acids Res.*, 35(11):e80, June 2007.

[167] Vladimir B. Teif. Predicting Gene-Regulation functions: Lessons from temperate bacteriophages. *Biophysical Journal*, 98(7):1247–1256, April 2010.

[168] Vladimir B Teif, Daniel Harries, Dmitri Y Lando, and Avinoam Ben-Shaul. Matrix formalism for site-specific binding of unstructured proteins to multicomponent lipid membranes. *Journal of Peptide Science: An Official Publication of the European Peptide Society*, 14(4):368–373, April 2008. PMID: 18186025.

[169] Vladimir B. Teif and Karsten Rippe. Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. *Nucl. Acids Res.*, 37(17):5641–5655, September 2009.

[170] D C Teller, T Okada, C A Behnke, K Palczewski, and R E Stenkamp. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of g-protein-coupled receptors (GPCRs). *Biochemistry*, 40(26):7761–7772, July 2001. PMID: 11425302.

[171] Nobuhiko Tokuriki and Dan S. Tawfik. Protein dynamism and evolvability. *Science*, 324(5924):203 –207, April 2009.

[172] Yuri Tsuda, Hironori Yasutake, Akihiko Ishijima, and Toshio Yanagida. Torsional rigidity of single actin filaments and actin–actin bond breaking force under torsion measured directly by in vitro micromanipulation. *Proceedings of the National Academy of Sciences of the United States of America*, 93(23):12937–12942, November 1996. PMID: 8917522 PMCID: 24024.

[173] William S J Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, August 2002. PMID: 12112692.

[174] F van den Ent, L A Amos, and J Löwe. Prokaryotic origin of the actin cytoskeleton. *Nature*, 413(6851):39–44, September 2001. PMID: 11544518.

[175] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy,

K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351, February 2001. PMID: 11181995.

[176] D J Vestal and B Ranscht. Glycosyl phosphatidylinositol–anchored t-cadherin mediates calcium-dependent, homophilic cell adhesion. *The Journal of Cell Biology*, 119(2):451–461, October 1992. PMID: 1400585.

[177] N Volkmann, D DeRosier, P Matsudaira, and D Hanein. An atomic model of actin filaments cross-linked by fimbrin and its implications for bundle assembly and function. *The Journal of Cell Biology*, 153(5):947–956, May 2001. PMID: 11381081.

[178] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67 –72, January 2009.

[179] Steven J Winder and Kathryn R Ayscough. Actin-binding proteins. *Journal of Cell Science*, 118(Pt 4):651–654, February 2005. PMID: 15701920.

[180] S Yamashiro-Matsumura and F Matsumura. Purification and characterization of an f-actin-bundling 55-kilodalton protein from HeLa cells. *The Journal of Biological Chemistry*, 260(8):5087–5097, April 1985. PMID: 3886649.

[181] Lei Yang, Guang Song, and Robert L. Jernigan. Comparisons of experimental and computed protein anisotropic temperature factors. *Proteins: Structure, Function, and Bioinformatics*, 76(1):164–175, 2009.

[182] R Yasuda, H Miyata, and K Kinosita. Direct measurement of the torsional rigidity of single actin filaments. *Journal of Molecular Biology*, 263(2):227–236, October 1996. PMID: 8913303.

[183] Zeyun Yu, Michael J Holst, Yuhui Cheng, and J Andrew McCammon. Feature-preserving adaptive mesh generation for molecular shape modeling and simulation. *Journal of Molecular Graphics & Modelling*, 26(8):1370–1380, June 2008. PMID: 18337134.

Colophon

This thesis was typeset in LaTeX on the MIT Athena using the template MIT provides.

The bibliography was organized using the Firefox add-on Zotero. Zotero permits easy incorporation of citations through a web-based browsing interface. Importantly, the data is stored on external servers and is therefore accessible from any computer with an internet connection.

The Figure were generated as PDFs whenever possible. As needed bitmaps were included in PNG format. All figures that required annotation were processed using Adobe Illustrator CS4, then resaved as PDFs.

The figures add a considerable amount of information to the PDF files. The resolution of the file was reduced using the Ghostscript on the *nix environment `gs`.