Contributions to the analysis of proteins

by

Reza Sharifi Sedeh

Master of Science in Mechanical Engineering (2005)
Sharif University of Technology, Tehran, Iran

Bachelor of Science in Mechanical Engineering (2003)
University of Tehran, Tehran, Iran

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author .....
Department of Mechanical Engineering
May 3, 2011

Certified by .....
Klaus-Jürgen Bathe
Professor of Mechanical Engineering
Thesis Supervisor

Certified by .....
Mark Bathe
Assistant Professor of Biological Engineering
Thesis Supervisor

Accepted by .....
David E. Hardt
Chairman, Department Committee on Graduate Students

Contributions to the analysis of proteins

by

Reza Sharifi Sedeh

Submitted to the Department of Mechanical Engineering
on May 3, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Proteins are essential to organisms and play a central role in almost every biological process. The analysis of the conformational dynamics and mechanics of proteins using numerical methods, such as normal mode analysis (NMA), provides insight into their functional mechanisms. However, despite the fact that much effort has been focused on improving NMA over the last few decades, the analysis of large-scale protein motions is still infeasible due to computational limitations.

In this work, first, we identify the usefulness and effectiveness of the subspace iteration (SSI) procedure, otherwise widely used in structural engineering, for the analysis of proteins. We also develop a novel technique for the selection of iteration vectors in protein NMA, which significantly increases the effectiveness of the method. The SSI procedure also lends itself naturally to efficient NMA of multiple neighboring macromolecular conformations, as demonstrated in a conformational change pathway analysis of adenylate kinase.

Next, we present a new algorithm to account for the effects of solvent-damping on slow protein conformational dynamics. The algorithm proves to be an effective approach to calculating the diffusion coefficients of proteins with various molecular weights, as well as their Langevin modes and corresponding relaxation times, as demonstrated for the small molecule crambin.

Finally, the structure of *Homo sapiens* fascin-1, an actin-binding protein that is present predominantly in filopodia, is examined and described in detail. Application of a sequence conservation analysis to the protein indicates highly conserved surface patches near the putative actin-binding domains of fascin. A novel conformational dynamics analysis suggests that these domains are coupled via an allosteric mechanism that may have important functional implications for F-actin bundling by fascin.

Thesis Supervisor: Klaus-Jürgen Bathe
Title: Professor of Mechanical Engineering

Thesis Supervisor: Mark Bathe
Title: Assistant Professor of Biological Engineering

Acknowledgments

This thesis could not have been completed without the support, encouragement, and inspiration of a number of wonderful people who made my life at MIT memorable. Although it would be impossible to name all of them, I would like to gratefully acknowledge those who contributed most to this work during my four-year and a half PhD study.

First, I would like to thank the chair of my thesis committee and my thesis supervisor Professor Klaus-Jürgen Bathe for his enthusiastic guidance and constant support throughout the course of my PhD studies. I have always considered myself extraordinarily lucky to have the opportunity to benefit from his tremendous experience and insight into the world of finite element methods. Although world-famous and extremely busy, he has always taken the time not only to discuss my research progress and to answer my technical questions but also to chat and to give me invaluable advice about non-academic challenges. I am really indebted to him for all the patience, flexibility, and encouragement he provided over the last four years and a half, especially during stressful periods.

I would also like to thank Professor Mark Bathe, my other thesis supervisor, who provided me the opportunity to work on a number of extremely interesting bioengineering projects. My special thanks go to him for being a constant source of guidance, knowledge, and encouragement throughout this work. No matter how busy his schedule, he has always made time to listen to my research results, give me great feedback, and share a number of wonderful ideas. I am deeply indebted to him for his invaluable support and assistance during my research and preparation of this dissertation.

Additionally, I am grateful to Professor Nicolas Hadjiconstantinou, the other member of my thesis committee, for his thoughtful comments and brilliant suggestions that significantly improved the quality of this work. I was extremely privileged to have

the opportunity to discuss my ideas with him during committee meetings. He has my deepest appreciation.

I also thank the members of the "Finite Element Method" group in the Department of Mechanical Engineering and the "Laboratory for Computational Biology & Biophysics" group in the Department of Biological Engineering for providing me a wonderful working environment. In addition, I am grateful to the staff of ADINA R&D Inc. for their constant support with the use of ADINA.

I would also like to thank all my friends at MIT and all over the world whose invaluable support and encouragement made my PhD studies at one of the most prestigious universities not only possible but also enjoyable. Thank you all!

Finally, I wish to express my infinite gratitude to my family: Esmaeil Sharifi Sedeh (father), Shahnaz Samiei Esfahani (mother), Arezoo and Sara (sisters), and Omid and Arash (brothers) for their unconditional support, sacrifice, understanding, trust, and encouragement during, and prior to, my PhD. This thesis is proudly dedicated to my parents, to whom I am forever indebted for their endless love and prayers.

Contents

List of Figures

List of Tables

Introduction

Proteins are essential to organisms and play a central role in almost every biological process. Based on their functions, proteins can be divided into different classes. Structural proteins such as F-actin and microtubules are a class of proteins that are used in the cytoskeleton of cells and are responsible for the cell geometry. Another class of proteins are enzymes, which are catalysts and accelerate the chemical reactions occurring within organisms. There are also many other proteins that play roles in cell adhesion, cell cycle, cell signalling, etc.

The conformational dynamics and mechanics of proteins are of great importance to many biological functions, ranging from transcription and translation to cell division and migration. Numerical methods, such as molecular dynamics (MD) and normal mode analysis (NMA), may give insight into the mechanical properties and dynamic behavior of proteins. Unlike MD, which needs to perform time-consuming time-integrations of the full set of governing equations of motion, NMA examines only harmonic oscillations of the protein around its ground-state conformation. As a result, NMA can be employed to analyze many protein motions that are currently inaccessible to MD. For example, NMA has proven successful in analyzing the functional motions associated with large macromolecules, such as myosin [5, 6], kinesin [5, 7], microtubules [8], and F-actin [9].

Over the last few decades, significant effort has been directed towards further improving the computational efficiency and accuracy of NMA for analyzing the conformational dynamics and mechanics of proteins. For example, one of the main time-consuming parts of NMA, which has attracted much attention, is solving the eigenvalue problem associated with the protein model. However, in spite of all the

effort [10, 11], the all-atom NMA of many protein motions, such as conformational change pathways of large macromolecules, is still almost infeasible due to the lack of a computationally efficient and robust eigenvalue solver. Additionally, since the effects of solvent friction on proteins are generally ignored in NMA, the time scales of protein functional motions cannot be predicted correctly using eigensolutions. Also, it is expected that the normal modes of proteins are altered substantially when the effects of solvent-damping are incorporated into NMA [12].

The present work focuses on both developing a computationally efficient and robust eigenvalue solver and incorporating the solvent-damping effects into NMA. Also, here NMA along with other computational procedures, such as sequence conservation analysis, are employed to gain insight into the functional mechanism of *Homo sapiens* fascin-1, an F-actin crosslinking protein.

In Chapter 1, we first review briefly the standard subspace iteration (SSI) method, a widely used eigenvalue solver in engineering problems [13]. Then, we present a new algorithm to optimize the number of iteration vectors employed in the method [14]. We subsequently apply the improved method to two proteins to illustrate its use in protein NMA. A particularly important observation is that with the new variant of the SSI method CPU time scales linearly with the number of eigenpairs sought [14], as in the Lanczos method [15]. Additionally, it is demonstrated that the SSI method is well-suited to the analysis of protein conformational change pathways, where hundreds of normal mode analyses may be performed in nearby conformations [16].

In Chapter 2, we first review the Langevin mode analysis developed by Lamm and Szabo [17] to incorporate the effects of solvent-damping into the standard NMA. Then, we present a new algorithm that calculates a solvent friction matrix using the finite element method (FEM) to account for the solvent-damping effects. The algorithm proves successful in calculating the diffusion coefficients of a sphere and 10 proteins with various molecular weights, ranging from 7 kDa to 233 kDa. We subsequently couple the solvent friction matrix and the stiffness and mass matrices calculated using the FEM [18] to obtain the Langevin modes and corresponding relaxation times of

crambin, a small protein with 46 amino acids. The obtained results are then compared with those calculated using bead models [19].

In Chapter 3, we first examine the structure of *Homo sapiens* fascin-1 [20], an actin-binding protein that is present predominantly in filopodia. The structure reveals a novel arrangement of four tandem β-trefoil domains that form a bi-lobed structure with approximate pseudo 2-fold symmetry. We subsequently apply sequence conservation analysis to the protein to investigate its structurally and functionally important regions. The results confirm the importance of the hydrophobic core residues that stabilize the β-trefoil fold, as well as the interfacial residues that are likely to stabilize the overall fascin molecule. Additionally, sequence conservation analysis indicates highly conserved surface patches near the putative actin-binding domains of fascin. Conformational dynamics analysis also suggests these domains to be coupled via an allosteric mechanism that might have important functional implications for F-actin crosslinking by fascin.

Finally, we present our conclusions.

Chapter 1

The subspace iteration method in protein normal mode analysis

Normal mode analysis (NMA) plays an important role in relating the conformational dynamics of proteins to their biological function [11]. In classical NMA [21, 22], protein atomic degrees of freedom are treated explicitly in solving the generalized eigenvalue problem in a biologically relevant conformation, typically for the lowest twenty to one hundred normal modes that represent the largest conformational fluctuations of the molecule. In the analysis of conformational transitions, numerous normal mode analyses may be performed for the same protein in nearby conformations [23].

NMA provides a considerable computational advantage over molecular dynamics because of the elimination of time-integration and explicit solvent degrees of freedom. Nevertheless, significant effort has been directed towards further improving the computational efficiency of NMA to enable its application to ever-larger supramolecular complexes including viral capsids, molecular motors, and the ribosome (Ref. [16] and references therein). Particular attention has been directed to the development and application of coarse-grained protein models such as elastic network and related models [18, 24], whereas somewhat less attention has been paid to the development of algorithms that improve the computational efficiency of all-atom protein NMA itself. Such developments are of interest because they preserve the explicit representation of atomic degrees of freedom and their solvent-mediated interactions as modeled by

implicit solvent force-fields. The explicit representation of atomic interactions is important to model accurately a number of biological processes, including interactions between proteins and nucleic acids [25], as well as small molecules in rational drug design [26]. Additionally, the role of allosteric regulation of binding affinity and catalysis by at-a-distance mutations remains an interesting and open area of research that may require all-atom modeling to understand fully [27].

The subspace iteration method was originally developed by K. J. Bathe for the solution of frequencies and mode shapes of macroscopic structures such as buildings and bridges using finite element analysis (FEA) [28, 29]. In those applications, relatively few frequencies and corresponding mode shapes were sought, such as the lowest 10–20 eigenpairs in models containing a total of 1000–10,000 degrees of freedom. Since its development, however, the subspace iteration method has been used extensively in the FEA of considerably larger systems reaching millions of degrees of freedom, and naturally has attracted significant attention for improvements as a result (see for example Refs. [30–37]).

The subspace iteration method is a particularly attractive approach to protein NMA because the procedure (1) is designed specifically for the calculation of the lowest eigenpairs of large systems; (2) uses previously calculated eigenvectors from nearby conformations to speed up significantly the solution of eigenpairs in nearby conformations of interest; (3) is computationally robust; and (4) is amenable to parallel-processing.

The original development of the method was based on the earlier use of the Ritz method, and relates to the works of Bauer [38] and Rutishauser [39]. Key developments for its practical use in structural engineering were the specific steps in the iteration method, the construction of the starting iteration vectors, the use of an effective number of iteration vectors, the use of error measures, and the Sturm sequence check [28]. A convergence analysis of the subspace iteration method is given in Ref. [40]. The method is also abundantly used in the solution of linearized buckling problems [13], which is applicable to calculations of the stability of the cytoskeletal polymers filamentous actin and microtubules, as well as viral capsids and other

supramolecular assemblies with mechanically related biological function [18].

An additional leading approach to NMA in the structural mechanics community is the Lanczos method [15], advanced particularly by Paige [41] and others [42]. Initially, the Lanczos method exhibited instabilities due to loss of orthogonality of the iteration vectors employed. This shortcoming, however, has been largely overcome, and when implemented properly the method is highly efficient. A particular asset of the method is that computational effort scales about linearly (neglecting the effort for the initial factorization) with the number of eigenpairs sought, a property that is not generally satisfied by the traditional subspace iteration method. An important property of both the subspace iteration and Lanczos procedures is that they solve directly for the eigenpairs sought instead of calculating intermediate matrices first, as if all eigenvalues were desired. This property contrasts with the approach of the Householder-QR method [13], for example, which becomes prohibitively expensive computationally and in memory as the size of coefficient matrices increases. At present, the Lanczos and subspace iteration methods are the two most widely used techniques for the solution of large eigenvalue problems in FEA, when coefficient matrices are of order 10,000−10,000,000. For these reasons, any significant improvements to these methods are of great interest.

Recently, considerable effort has been directed towards using parallel processing in FEA, in shared-memory and distributed-memory processing modes. Whereas the Lanczos method can intrinsically (largely) be parallelized only in the factorization of the stiffness matrix and the forward reduction and back-substitution of the *individual* vectors, the subspace iteration method allows in addition the parallel solution of *multiple* iteration vectors which can result in a large computational benefit. However, there is also interest in improving the method in other ways, and in particular, for the solution of eigenproblems in which relatively many eigenpairs need to be calculated.

As mentioned earlier, a key step in the subspace iteration method is the establishment of effective starting iteration vectors, which implies using an optimal number of iteration vectors. The objective of the present work is to apply the subspace iteration method to the normal mode analysis of proteins, and to introduce a significant

improvement upon the original implementation regarding the choice of the number of iteration vectors. In the following sections, we first review briefly the standard subspace iteration method and discuss its inherent value for the solution of frequencies and mode shapes of proteins. We, subsequently, present a new algorithm to establish an effective number of iteration vectors, illustrating the use of this algorithm in some applications. A particularly important observation is that computational effort increases linearly with the number of eigenpairs sought in the solutions obtained with the improved subspace iteration method, as in the Lanczos method. To focus on our new development only, and to compare results obtained with the traditional and improved methods, we employ a basic implementation without parallelization of the code, running in-core on a single processor workstation. Moreover, we provide only relative solution times, which are largely independent of the machine used. Although these times thereby represent practically "machine-independent" algorithmic improvements, actual solution times will naturally depend on the specific machine employed and will decrease as computational hardware becomes more efficient.

1.1 Methods

1.1.1 The basic subspace iteration method

We consider the generalized eigenvalue problem,

$$\mathbf{K}\boldsymbol{\varphi} = \lambda \mathbf{M}\boldsymbol{\varphi} \tag{1.1}$$

where $\mathbf{K}$ and $\mathbf{M}$ are symmetric matrices of order n, $\mathbf{K}$ is positive definite, and $\mathbf{M}$ is positive semidefinite. We seek the smallest p eigenvalues λ_1, λ_2, ..., λ_p and corresponding eigenvectors $\boldsymbol{\varphi}_1$, $\boldsymbol{\varphi}_2$, ..., $\boldsymbol{\varphi}_p$ with the ordering,

$$\lambda_1 \leq \lambda_2 \ldots \leq \lambda_p \tag{1.2}$$

The eigenpairs $(\lambda_i, \boldsymbol{\varphi}_i)$ satisfy,

$$\mathbf{K}\boldsymbol{\varphi}_i = \lambda_i \mathbf{M}\boldsymbol{\varphi}_i; \qquad i = 1, \ ..., \ p \tag{1.3}$$

and

$$\begin{aligned} {\boldsymbol{\varphi}_i}^{\mathrm{T}}\mathbf{M}\boldsymbol{\varphi}_j &= \delta_{ij} \\ {\boldsymbol{\varphi}_i}^{\mathrm{T}}\mathbf{K}\boldsymbol{\varphi}_j &= \lambda_i \delta_{ij} \end{aligned} \tag{1.4}$$

where δ_{ij} is the Kronecker delta. The basic equations used in the subspace iteration method are as follows [13]:

Step 1: Establish q starting iteration vectors in $\mathbf{X}_1$

Step 2: Iterate with k= 1, 2, 3, ..., until convergence

$$\mathbf{K}\overline{\mathbf{X}}_{k+1} = \mathbf{M}\mathbf{X}_k \tag{1.5}$$

$$\begin{aligned} \mathbf{K}_{k+1} &= {\overline{\mathbf{X}}_{k+1}}^{\mathrm{T}}\mathbf{K}\overline{\mathbf{X}}_{k+1} \\ \mathbf{M}_{k+1} &= {\overline{\mathbf{X}}_{k+1}}^{\mathrm{T}}\mathbf{M}\overline{\mathbf{X}}_{k+1} \end{aligned} \tag{1.6}$$

$$\mathbf{K}_{k+1}\mathbf{Q}_{k+1} = \mathbf{M}_{k+1}\mathbf{Q}_{k+1}\boldsymbol{\Lambda}_{k+1} \tag{1.7}$$

$$\mathbf{X}_{k+1} = \overline{\mathbf{X}}_{k+1}\mathbf{Q}_{k+1} \tag{1.8}$$

Step 3: Perform the Sturm sequence check.

Hence, the procedure consists of three distinct solution steps. First, the q starting iteration vectors in $\mathbf{X}_1$ are established, $q > p$, where $\mathbf{X}_1$ is a matrix of dimension $n \times q$. Second, iteration is performed using Eqs. 1.5–1.8, for $k = 1, 2, \ldots$ until the convergence tolerance below is satisfied, where $\mathbf{Q}_{k+1}$ and $\boldsymbol{\Lambda}_{k+1}$ store the eigenvectors and eigenvalues corresponding to the subspace matrices $\mathbf{K}_{k+1}$ and $\mathbf{M}_{k+1}$. Finally, the Sturm sequence check is performed.

Let ${\lambda_i}^{(k)}$ be the approximation for λ_i calculated in the $(k-1)^{\mathrm{th}}$ iteration, we have convergence to an accuracy of $2 \times s$ digits in the eigenvalues when for $i = 1, \ ..., \ p$ (see Ref. [13]),

$$\left[1 - \frac{\left(\lambda_i^{(k)}\right)^2}{\left(\mathbf{q}_i^{(k)}\right)^{\mathrm{T}} \mathbf{q}_i^{(k)}}\right]^{1/2} \leq 10^{-2s} \tag{1.9}$$

where $\mathbf{q}_i^{(k)}$ is the vector in the matrix $\mathbf{Q}_k$ corresponding to $\lambda_i^{(k)}$. The eigenvectors will only be accurate to s digits and the theoretical convergence rate of the vectors is λ_i/λ_{q+1}. Thus, there is a higher convergence rate for a smaller eigenvalue and its corresponding eigenvector. Although these convergence rates correspond to the theoretical values [13, 40], they are usually also observed in actual computations. The Sturm sequence check is carried out to ensure that the lowest p eigenpairs, that is, $(\lambda_i, \boldsymbol{\varphi}_i)$, $i = 1, ..., p$, have indeed been calculated [13, 28]. If the Sturm sequence check is not passed, the iteration is continued with a larger number of iteration vectors.

Considering Eqs. 1.5–1.8, it is seen that the method can be programmed efficiently for parallel computations. The factorization of the coefficient matrix and the forward reductions and back-substitutions of each individual vector can be parallelized. In addition, the solution of the q vectors can be distributed to different processors and also the computation of the subspace matrices $\mathbf{K}_{k+1}$ and $\mathbf{M}_{k+1}$ can be parallelized.

An important difference between the coefficient matrices of structural FE assemblages and of proteins is that the latter have much larger bandwidths because of long-range nonbonded electrostatic, and to a lesser extent van der Waals, interactions that introduce broad coupling between protein atoms. Thus, for a given number of degrees of freedom, the factorization of the matrix and solution of the vectors in Eq. 1.5 constitute a much larger computational effort than in standard FE solutions. Although parallel processing can be very important for this reason, we do not address this computational issue further in the present work.

Using the earlier equations, it is critical to establish effective starting iteration vectors for two reasons. First, if the subspace of these vectors contains the exact eigenvectors, theory states that a single iteration will result in the exact eigenvalues and vectors sought. Here, we simply use the algorithm of Ref. [28] (also given in Ref. [13]), to construct the starting iteration vectors. In cases where better starting

vectors are known from an existing solution, such as in conformational change pathway analyses of proteins where eigensolutions may be performed numerous times for small changes in protein conformation [23], the algorithm of Ref. [13] is used only for the first eigensolution. Thereafter, the previous solution from the nearest-neighbor conformation provides the starting iteration vectors for the next eigensolution. Second, an effective value of q needs to be used because the convergence rate to an eigenvector is given by λ_i/λ_{q+1}. If q $(> p)$ is small, a relatively large number of iterations are required to converge. In contrast, if q is large, fewer iterations are required for convergence, but each iteration is computationally more costly. Thus, use of an optimal value of q is highly desirable. Calculation of an effective value of q for the frequency and mode shape solutions of proteins is addressed in the next section.

1.1.2 The algorithm to calculate the number of starting iteration vectors

An important observation regarding proteins is that the magnitudes of their eigenvalues increase nearly linearly with increasing wave-number [43, 44], as shown for T4-lysozyme in Fig. 1-1. This characteristic of proteins may be used to find an effective value of q for the subspace iteration method.

Assume that we order the iteration vectors in $\mathbf{X}_k$ naturally so that they correspond to increasing eigenvalues, with the first vector corresponding to λ_1. Then the last iteration vector to converge is the p^{th} vector in $\mathbf{X}_k$ and its rate of convergence is $\frac{\lambda_p}{\lambda_{q+1}}$. Additionally, after the i^{th} iteration, the norm of the vector difference between the p^{th} $\mathbf{M}$-orthonormalized eigenvector and its current approximation (the error vector $\boldsymbol{\varepsilon}$) is given by,

$$\|\boldsymbol{\varepsilon}\,(\text{current})\| = \left(\frac{\lambda_p}{\lambda_{q+1}}\right)^i \|\boldsymbol{\varepsilon}\,(\text{initial})\| \tag{1.10}$$

where $\|\boldsymbol{\varepsilon}\,(\text{initial})\|$ is the initial error vector. To reach s-digits of accuracy in the eigenvector we need,

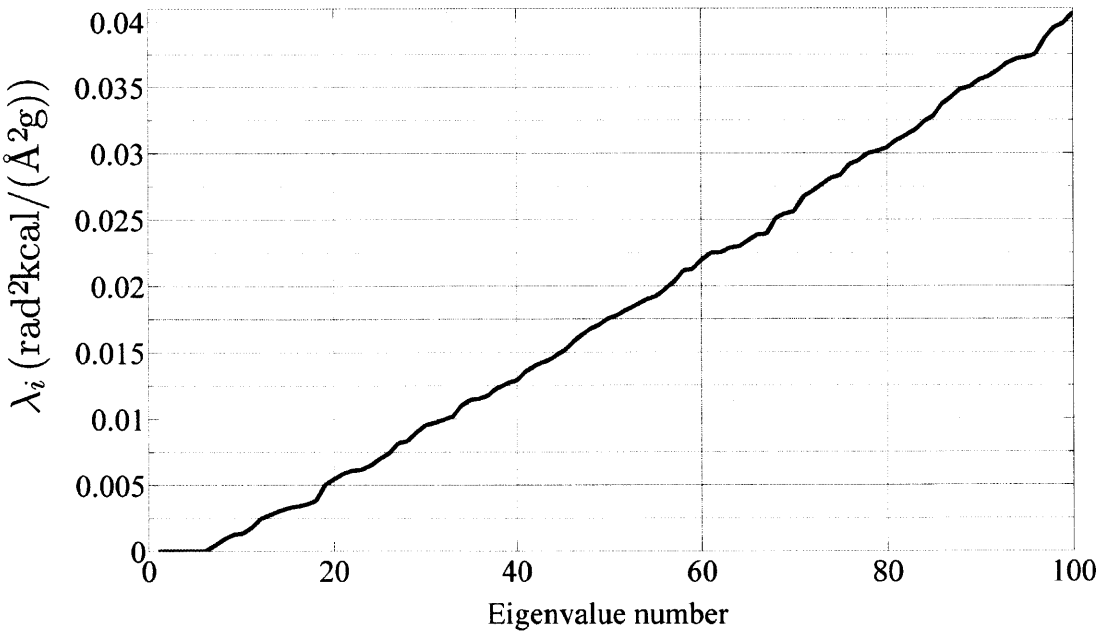

Figure 1-1 – The lowest one hundred eigenvalues (λ_i) of T4-lysozyme (Protein Data Bank ID 3LZM) [1]. (The first six zero eigenvalues correspond to rigid body modes.)

$$\left(\frac{\lambda_p}{\lambda_{q+1}}\right)^i \|\boldsymbol{\varepsilon}\,(\text{initial})\| \leq 10^{-s} \tag{1.11}$$

and, therefore, require l iterations for the vector to converge, where l is given by,

$$l = \frac{\ln\left(10^{-s}/\,\|\boldsymbol{\varepsilon}\,(\text{initial})\|\right)}{\ln\left(\lambda_p/\lambda_{q+1}\right)} \tag{1.12}$$

Next, we use the fact that the eigenvalue magnitudes increase linearly and assume that for different values of q, the norm of the initial error vector for the p^{th} iteration vector is the same. Additionally, the first six eigenvalues are zero. This implies that the $\mathbf{K}$ matrix is singular. To use the subspace iteration method, we perform a shift ρ on the $\mathbf{K}$ matrix to have a positive definite matrix, see Ref. [13]. We use ρ to be a very small value, $\rho = -1 \times 10^{-6}$. Therefore, $\frac{\lambda_p}{\lambda_{q+1}}$ is approximately equal to $\frac{(p-6-\rho)}{(q-5-\rho)}$. Since ρ is very small, it can be neglected and $\frac{\lambda_p}{\lambda_{q+1}}$ is approximated as $\frac{(p-6)}{(q-5)}$. Then Eq. 1.12 gives us directly,

$$l = \frac{\ln\left(10^{-s} / \left\|\boldsymbol{\varepsilon}\left(\text{initial}\right)\right\|\right)}{\ln\left(\left(p-6\right)/\left(q-5\right)\right)} \tag{1.13}$$

However, an operation count tells that the following number of numerical operations are needed for l iterations with q vectors [13],

$$TCC = \frac{\ln\left(10^{-s} / \left\|\boldsymbol{\varepsilon}\left(\text{initial}\right)\right\|\right)}{\ln\left(\left(p-6\right)/\left(q-5\right)\right)}\left(2nmq + 2nq^2 + 3nq\right) \tag{1.14}$$

where TCC is the Total Cost of Computation for l iterations, n is the order of the $\mathbf{K}$ and $\mathbf{M}$ matrices, and m is the half-bandwidth (assumed to be full) of the $\mathbf{K}$ matrix. As the column heights of $\mathbf{K}$ vary, an average or effective value for m must be used [13]. Although we refer to TCC in Eq. 1.14, in reality we only have the total number of *arithmetical* operations. As our only purpose is to find an effective value of q for each p, and we also know that,

$$c = \ln\left(10^{-s} / \left\|\boldsymbol{\varepsilon}\left(\text{initial}\right)\right\|\right)$$

where c is an unknown constant, we may use,

$$TCC = \frac{c}{\ln\left(\left(p-6\right)/\left(q-5\right)\right)}\left(2nmq + 2nq^2 + 3nq\right) \tag{1.15}$$

Minimizing this expression with respect to q we find an approximation for the best q to obtain the p eigenvalues and vectors in the least amount of computational time. Because a closed-form solution does not exist, we solve for q by iteration. Note that this analysis does not provide the actual computational effort required (since the constant c is unknown) but only that the minimum is obtained when using the value of q given by minimizing TCC in Eq. 1.15.

Fig. 1-2 shows the normalized actual solution time and TCC to calculate the lowest 100 eigenvalues with six digits of accuracy for T4-lysozyme using different numbers of iteration vectors. The iteration times are normalized by the maximum actual iteration time and, since the constant c in Eq. 1.15 is unknown, TCC is scaled such that the iteration times are equal at the minimum of TCC.

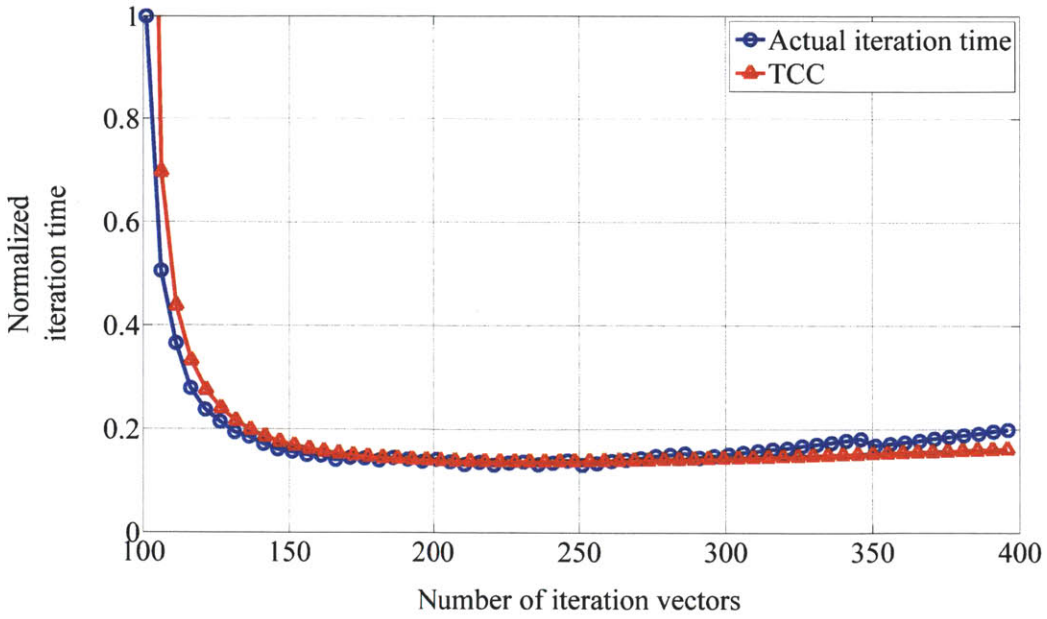

Figure 1-2 – Normalized actual iteration time and normalized TCC to calculate the first one hundred eigenvalues for T4-lysozyme (Protein Data Bank ID 3LZM) [1].

As seen in Fig. 1-2, prediction of the relative computational cost of calculating the lowest eigenvalues with different numbers of iteration vectors by Eq. 1.15 is acceptable. Next we illustrate the use of the value of q in the normal mode analyses of two proteins.

1.2 Results

1.2.1 Illustrative solutions

In this section we use the subspace iteration method for the calculation of the frequencies and normal modes of two proteins. In each case we use the standard subspace iteration method as published in Refs. [13, 28] including the algorithm to construct *all* starting iteration vectors. We use the standard value q= min $\{2p, p+8\}$, referred to as the traditional subspace iteration method, and this method with the value of q that minimizes TCC in Eq. 1.15, referred to as the improved subspace iteration method. We intentionally do not use any other acceleration techniques, such as given for example in Ref. [30], to identify clearly the improvements achieved solely by use

of the value of q derived earlier.

In each solution we employ the skyline solver of Ref. [13] for Eq. 1.5. Although we recognize that a sparse solver could lead to significantly improved solution times [45], we do not expect our fundamental observations regarding the performance of the method to be affected. We note that the solution times given always include all operations of the subspace iterations. Additionally, in an effort to present machine-independent conclusions regarding performance of the algorithms, we present normalized solution times instead of actual solution times, where normalized time is equal to actual time divided by the maximum solution time measured in each case.

G-actin

The initial structure of ADP-bound G-actin is taken from the work of Otterbein et al. [2] (Protein Data Bank ID 1J6Z; residue numbers 4–372). The stiffness matrix of order 10,608 for this protein was computed in CHARMM version 34b1 [46] using the implicit solvation model EEF1 [47]. Steepest descent minimization followed by adopted-basis Newton-Raphson minimization is performed in the presence of successively reduced harmonic constraints on backbone atoms to achieve a final root-mean-square (RMS) energy gradient of $2\times10^{-4}\ \frac{\text{kcal}}{(\text{mol}\times\text{Å})}$ with corresponding RMS deviation between the X-ray and energy-minimized structures of 1.4 Å (Fig. 1-3). Computations are performed on an Intel Xeon 5120 with 1.86 GHz and 4 GB RAM in single processor mode.

Considering the eigenvalue problem, different numbers of the lowest eigenvalues with six digits of accuracy of this protein have been obtained using the traditional and improved subspace iteration methods. Fig. 1-4 provides normalized solution times versus the required number of lowest eigenvalues for G-actin, and also provides in parentheses the number of iteration vectors q used in the improved subspace iteration method in each case. It is evident that a significant improvement in the subspace iteration method is achieved by use of the calculated values of q.

As already noted, normalized solution times in Fig. 1-4 are defined as the actual solution times divided by the maximum solution time encountered in the analysis. The maximum solution time (13,939 seconds clock-time) in this case is the time

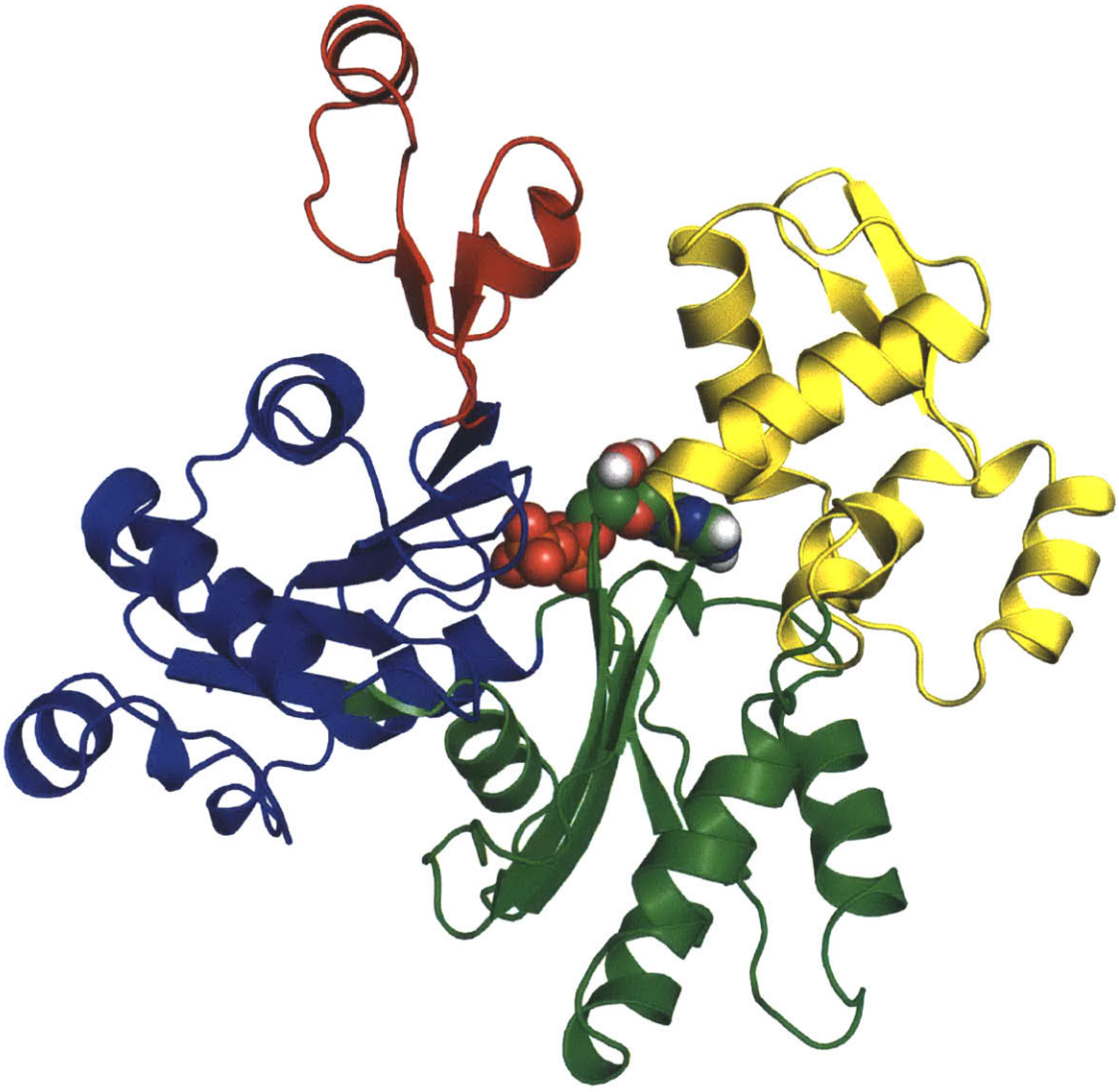

Figure 1-3 – G-actin-ADP. Schematic representation of the energy-minimized molecular structure analyzed with subdomains colored according to the definition of Kabsch et al. [48], Subdomain 1 is colored blue, subdomain 2 is colored red, subdomain 3 is colored green, and subdomain 4 is colored yellow. ADP is shown in van der Waals representation. Figure rendered using PyMOL [49].

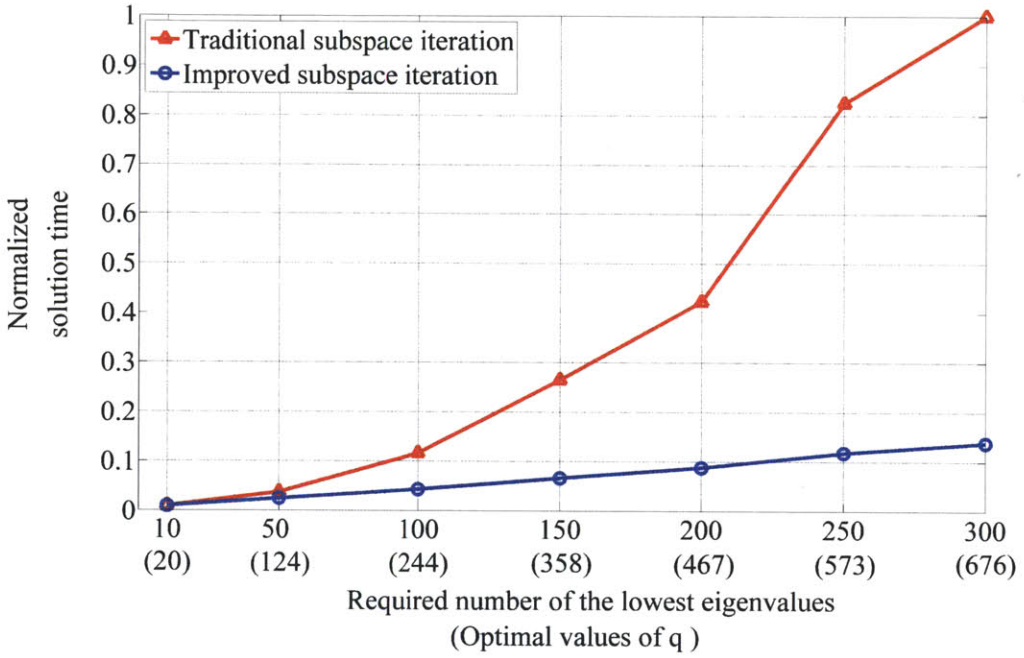

Figure 1-4 – Normalized solution times versus required number of the lowest eigenvalues with six digits of accuracy for G-actin (Protein Data Bank ID 1J6Z) [2] using the traditional and improved subspace iteration methods; the value of q used in each case with the improved subspace iteration method is given in parentheses.

required to compute the lowest 300 eigenpairs with the traditional subspace iteration method. This solution time is quite large for the reasons mentioned earlier.

Pertussis toxin

The next protein examined is pertussis toxin (chains A–F). Initial coordinates are taken from the work of Stein et al. [3] (Protein Data Bank ID 1PRT). Like for G-actin, CHARMM version 34b1 [46] with the implicit solvation model EEF1 [47] is used to obtain the energy-minimized structure (Fig. 1-5) and calculate the Hessian, which has dimension of order 26,664. Steepest descent minimization followed by adopted-basis Newton-Raphson minimization is performed in the presence of successively reduced harmonic constraints on backbone atoms to achieve a final root-mean-square (RMS) energy gradient of 3×10^{-4} $\frac{\text{kcal}}{(\text{mol}\times\text{Å})}$ with corresponding RMS deviation between the X-ray and energy-minimized structures of 1.6 Å. Computations are also performed on an Intel Xeon 5120 with 1.86 GHz and 4 GB RAM in single processor mode.

Fig. 1-6 shows the measured normalized solution times versus the required number of the lowest eigenvalues for this molecule, and also gives in parentheses the number of iteration vectors q used in the improved subspace iteration method in each case. Again, significant computational savings are achieved when the improved iteration method is used.

1.2.2 Conformational change pathway analysis of adenylate kinase

To illustrate the benefit of employing the subspace iteration procedure to analyze conformational change pathways of proteins, we apply the procedure to the open-to-closed transition of adenylate kinase (PDBIDs 4AKE [50] and 1AKE [51] for the open and closed conformers, respectively)(Figs. 1-7-A and 1-7-B). In the absence of molecular dynamics or other all-atom trajectory, we employ the elastic-based FE model applied previously to protein NMA to generate the conformational change pathway [18]. The initial model is defined by the open conformation of the protein.

Figure 1-5 – Pertussis toxin. Schematic representation of the energy-minimized molecular structure analyzed with subdomains colored according to the definition of Stein et al. [3], S1 is colored green, S2 is cyan, S3 is purple, S4 is red, and S5 is yellow. Figure rendered using PyMOL [49].

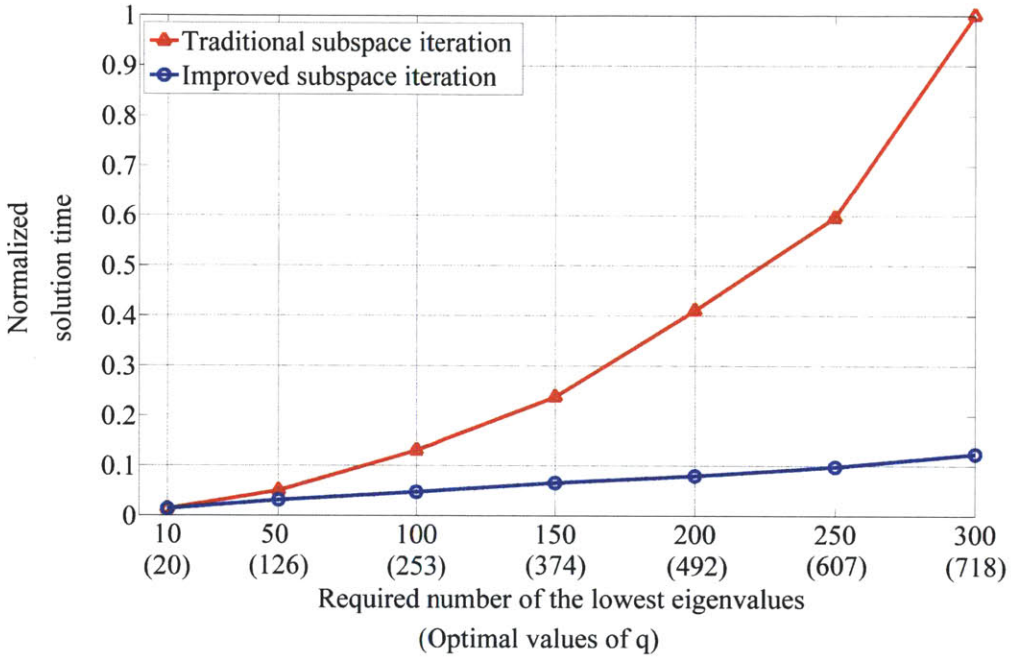

Figure 1-6 – Normalized solution times versus required number of the lowest eigenvalues with six digits of accuracy for one of two molecules from pertussis toxin (Protein Data Bank ID 1PRT; Chains A–F) [3] using the traditional and improved subspace iteration methods; the value of q used in each case with the improved subspace iteration method is given in parentheses.

Following Ref. [18] the molecular volume is defined by the solvent excluded surface (SES) using MSMS ver. 2.6.1 [52]. This SES is then decimated to a coarsened surface using the surface simplification algorithm QSLIM [53–55], as implemented in MeshLab [56]. Finally, the decimated SES is imported into the finite element analysis program ADINA ver. 8.5 (Watertown, MA), where the molecular volume is meshed automatically using 3D four-node tetrahedral elements [18]. The protein is assumed to behave as a linear, isotropic material with homogeneous mass density of 1420 $\frac{\text{kg}}{\text{m}^3}$, elastic Young's modulus of 4.9 GPa, and Poisson's ratio of 0.3. The mass density is obtained from the molecular weight and molecular volume of the open conformation. The Young's modulus is obtained by fitting thermal fluctuations of α-carbon atoms in the finite element model to those obtained using the Rotation Translation Block procedure [57, 58] at room temperature in CHARMM, where one block per residue and the implicit solvation model EEF1 [47] are employed (see Appendix B).

The conformational change pathway of adenylate kinase is generated according to the procedure of Tama, Miyashita, and Brooks [59]. Starting from the initial, open conformation, $\mathbf{K}$ and $\mathbf{M}$ matrices are generated for the FE model using ADINA. The traditional subspace iteration procedure is then used to calculate the first 100 eigenpairs of the model with four digits of accuracy for the eigenvalues. The FE model interpolation functions are used to interpolate the eigenvectors, $\boldsymbol{\varphi}_i{}^k$, corresponding to the FE nodal positions to their values, $\mathbf{C}_i{}^k$, at the positions of the α-carbons, where i and k denote the number of the eigenvector and conformation, respectively. To generate the next conformation, the difference vector between the positions of the α-carbons in the k^{th} conformation and those of the closed conformation, $\boldsymbol{\Delta}\mathbf{r}^k$, is projected onto the eigenvectors corresponding to the α-carbons, $c_i{}^k = \beta^k \boldsymbol{\Delta}\mathbf{r}^k \cdot \mathbf{C}_i{}^k$, where β^k is a parameter between zero and one [23, 59] (see Appendix A). $c_i{}^k$ is the contribution of the i^{th} eigenvector to the displacement of the α-carbons in the k^{th} step. Positions of all non-α-carbon atoms are updated using the FE displacement interpolation functions in the current conformation. This procedure is repeated until the root-means-quare-difference (RMSD) between the current positions of α-carbons and those of the closed conformer is less than or equal to 1 Å. In this approach to

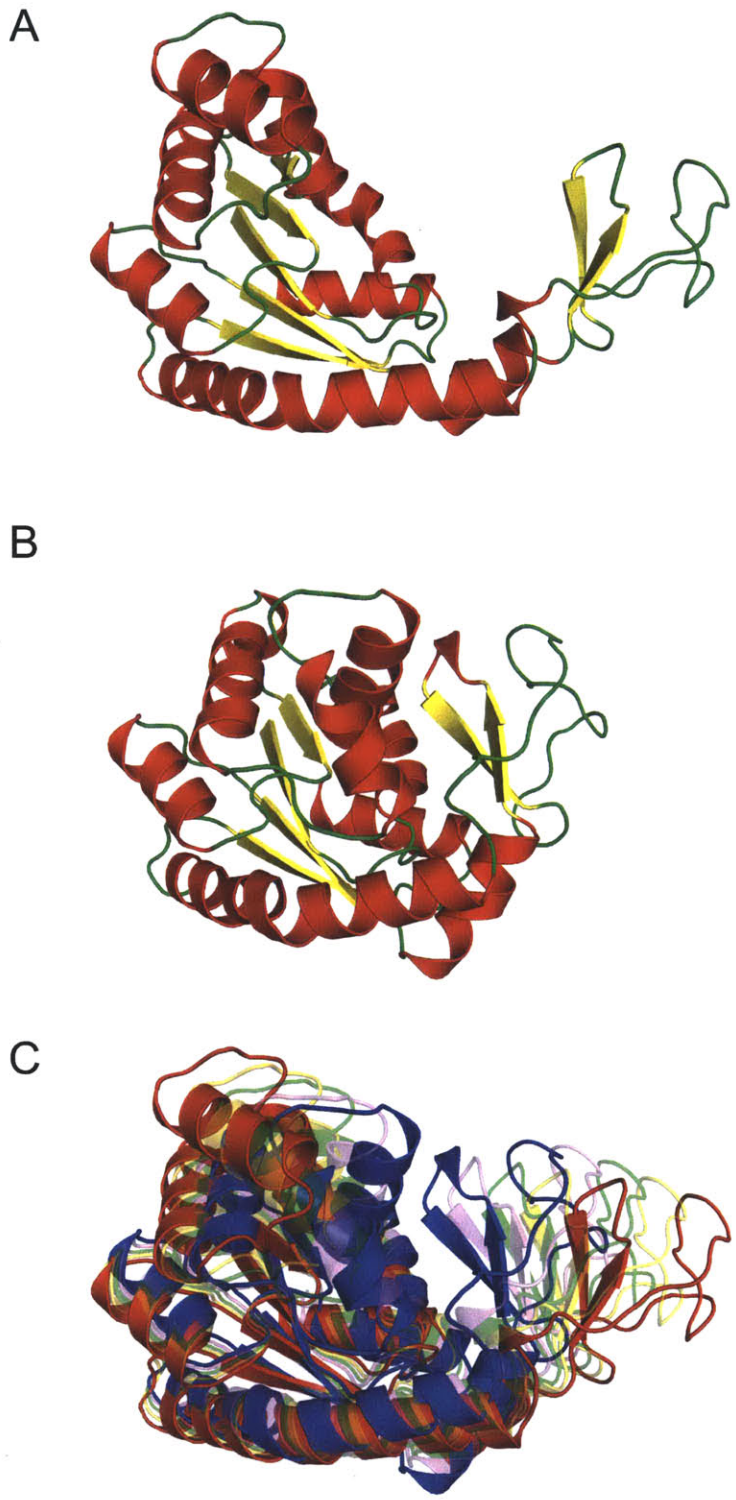

Figure 1-7 – Conformational change pathway of adenylate kinase. (A) Schematic representation of the open conformation of adenylate kinase (Protein Data Bank ID 4AKE [50]). (B) Schematic representation of the closed conformer of adenylate kinase (Protein Data Bank ID 1AKE [51]). (C) Schematic representation of the open-to-closed transition. The root-mean-square-difference between the positions of α-carbons in the closed conformer and that of the red, yellow, green, violet, and blue conformations is 7.14, 5.25, 3.5, 1.75, and 0 Å, respectively. Figures rendered using PyMOL [49].

generating the conformational change pathway, the eigenvectors of the current conformation are used as the starting vectors for the eigenvalue problem of the next conformation, excluding the first step, which is also excluded from the solution time per conformation presented below because it constitutes a small and invariant component of the total solution time in each case. An initial conformational change pathway of 1843 conformations is generated, from which subsets of 1001, 101, 11, and 1 conformation are chosen with nearly constant differences in RMSD between α-carbon positions of each successive conformation and the closed conformation (see Appendix A) (Fig. 1-7-C). Computations are performed on an Intel Xeon E5405 with 2.00 GHz and 16 GB RAM in single processor mode.

The solution time per conformation for the subspace iteration procedure decreases monotonically with increasing number of conformations employed in the conformational change pathway (Fig. 1-8). Normalized time is equal to the actual solution time divided by the maximum solution time measured in the 100 normal mode case. As an increasing number of conformations is employed, normal mode solutions from neighboring conformations become increasingly better choices for the starting normal modes of neighboring conformations, resulting in the observed decrease in solution time per conformation. This result is true whether 20 or 100 eigenvectors are solved for (Fig. 1-8), and is additionally expected to be independent of the number of degrees of freedom in the model. Although it is of interest to understand the detailed solution-time properties of the subspace iteration procedure in conformational change pathway analysis (e.g., dependence of solution time per conformation scaling with model size, number of normal modes computed, etc.), such analysis is reserved for future work.

1.3 Important properties of the subspace iteration method

In evaluating the effectiveness of any numerical procedure, it is clearly valuable to make a thorough comparison with existing methods [10, 15, 21]. In the present

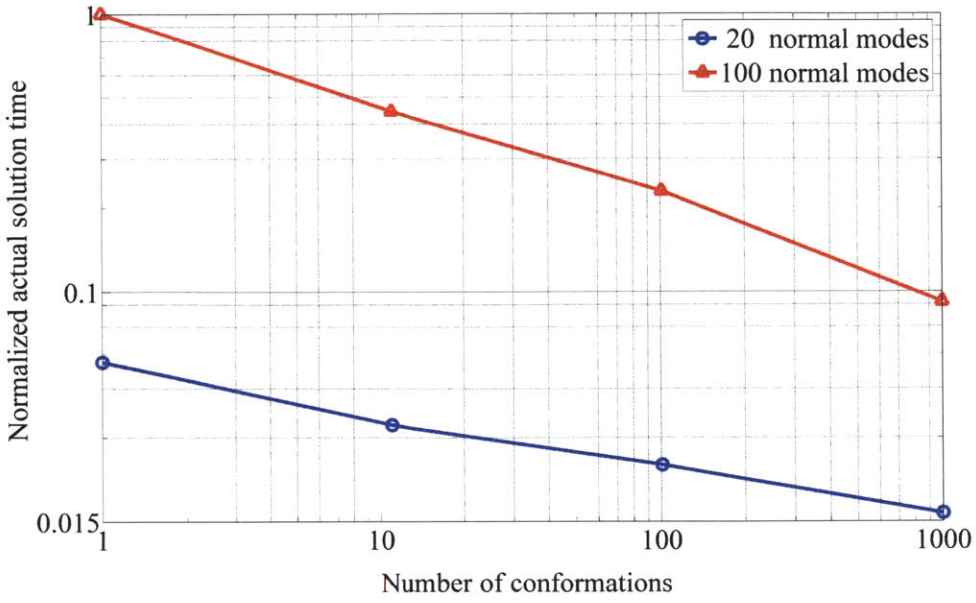

Figure 1-8 – Normalized actual solution time per conformation for the subspace iteration method versus the number of conformations analyzed in the conformational change pathway of adenylate kinase using 100 and 20 normal modes.

case, such comparison is unfortunately complicated by a number of factors, including the requirement that each method employs the same convergence tolerance and is implemented in the optimal manner. Even then, results would depend on whether the computation is performed in- or out-of-core, the type of parallel processing used, the degree of energy-minimization performed in the use of some methods, and so on. While such a comparison would clearly be of value, it is outside the scope of the present work. Nevertheless, we would like to point out several important properties of the subspace iteration procedure, and in particular contrast these properties with corresponding properties of the Lanczos method.

The subspace iteration procedure converges monotonically and robustly to the number of frequencies and mode shapes sought. In each subspace iteration, inverse iteration is performed on a q-dimensional subspace and a Rayleigh-Ritz analysis extracts the best approximations to the p normal modes sought. Best here refers to minimization of the Rayleigh quotient on the subspace [13, 40]. As the q-dimensional subspace is rotated towards the least dominant p-dimensional subspace within each

iteration, the NM approximations become more accurate. If only low accuracy in the normal modes is needed, only a few subspace iterations may be required.

Solution time in the Lanczos method scales approximately linearly with the number of eigenpairs computed. The traditional subspace iteration does not typically display this scaling when many frequencies and mode shapes are calculated (e.g., > 20) and a single processor is employed. In the present work, however, we observed that the subspace iteration method with the improved selection of the number of iteration vectors also resulted in linear scaling of solution time with the number of normal modes sought. As expected, we additionally observed a significant decrease in computational time when the NMA was performed on multiple neighboring conformations, because the method uses normal mode solutions from neighboring conformations to accelerate subsequent solutions. This is an important property of the subspace iteration procedure that is not a property of methods that start with individual vectors, such as the Lanczos algorithm. Additional acceleration might be achieved for NMA of single conformations by exciting principally the dihedral angles to choose starting vectors that span a subspace that is closer to the required least dominant subspace than the algorithm employed here [13, 28]. In addition, acceleration techniques published previously could be implemented [30, 35].

A final important computational property of any NMA procedure is the possibility to use parallel processing (with shared and distributed memory), such as implemented for the Lanczos procedure in the publically available program ARPACK [60]. Although the calculations in the subspace iterations (Eqs. 1.5–1.8) lend themselves naturally to parallel processing, the actual benefits achievable in comparison to the Lanczos procedure, which operates sequentially on individual vectors, remain to be established. Use of a combination of the basic steps in the subspace iteration and Lanczos methods, using the best ingredients of each technique and taking into account parallel processing, would be of interest to reach a more effective method. Further investigation is required to identify the appropriate next steps to take in this research direction.

1.4 Concluding remarks

The objective of this chapter was to present the application of the subspace iteration method to the normal mode analysis of proteins and to provide an algorithm for the calculation of an effective number of iteration vectors. We demonstrated use of an algorithm to calculate the number of iteration vectors q to find p eigenpairs that improves the effectiveness of the subspace iteration method significantly for proteins. The algorithm results in computation time scaling linearly with the number of eigenpairs sought, as demonstrated for G-actin and pertussis toxin. The subspace iteration method is well suited to protein NMA because relatively small subsets of the total available normal modes are typically sought and numerous analyses may be performed for relatively similar conformations in conformational change pathway analyses [23]. In such cases, the previously calculated eigensolution provides an excellent set of initial iteration vectors for the subsequent solution, as demonstrated here for the open-to-closed conformational change of adenylate kinase. The subspace iteration method is additionally attractive because it is robust, in that it converges monotonically to the desired eigenvalue solution for any positive semidefinite stiffness matrix. This is of significant utility in all-atom protein NMA for two reasons. First, energy minimization to tight tolerance in the energy gradient is time-consuming and often challenging due to the rugged energy landscape of proteins, and second, energy minimization often distorts the protein structure such that it deviates significantly from the experimental crystal structure. For these reasons, and due to its relative computational efficiency, the robust Rotational Translational Blocks procedure [57, 58] has gained significant popularity. However, this procedure assumes single or larger blocks of residues to be rigid, in contrast with the present implementation that retains all atomic degrees of freedom. Although the significant reduction in number of degrees of freedom in the former approach renders its computational efficiency high, an interesting area of future research concerns the integration of computationally robust NMA methods with efficient reduced degree-of-freedom approaches that retain internal residue flexibility, as initially proposed in Ref. [57]. Incorporation of such

procedures into the finite element method would enable simultaneously calculations of protein mechanical response, as well as NMs.

Chapter 2

Finite element framework for Langevin modes of proteins

Protein motions such as conformational changes, folding/unfolding, and ligand association/dissociation generally occur in a physiological solvent, a viscous environment within cells. Hence, to analyze the true dynamic behavior of a protein, both the protein and the solvent have to be modeled simultaneously, as in all-atom, explicit-solvent molecular dynamics [61]. However, in practice, especially for the above-mentioned long-time and large length-scale motions, the time-integration of the full set of governing equations of motion performed in the molecular dynamics is infeasible. Hence, coarse-grained models have been developed to speed up the analysis of the dynamic behavior of proteins. These models can describe many protein motions which are currently inaccessible to the standard molecular dynamics. For example, protein folding and unfolding have been investigated, respectively, using lattice models [62–64] and steered molecular dynamics [65]. Also, the elastic network model (ENM), a coarse-grained normal mode analysis (NMA), has been used to analyze the conformational change pathways of proteins [7, 66–68]. Generally, the effects of solvent friction on proteins are ignored in these normal mode analyses. Consequently, the frequencies of proteins calculated from the set of the governing equations of motion in a vacuum cannot be used to predict the actual time-scales of functional protein motions in a solvent. Also, the normal modes of proteins are altered significantly when incorporating

the effects of solvent-damping into the normal mode analyses [12, 17].

The Langevin mode analysis (LMA) developed based on the Langevin dynamics formalism by Lamm and Szabo [17] can account for the effects of solvent friction on the normal modes and corresponding time-scales of proteins. In the Langevin dynamics formalism [17, 69–71], the effects of solvent friction are implicitly applied to proteins. In practice, the Langevin dynamics simulations themselves are computationally too expensive to be used for the dynamic behavior analysis of many large proteins, while the LMA can be more readily applied to the analysis of the proteins. Recently, the LMA has been used by Miller et al. to examine the dynamic behavior of myosin in a solvent [12]. They employed bead models [72–80] to incorporate solvent-damping into the ENM. In their bead model, one bead was located at the position of every C_α [12]. The radii of beads were calibrated to the experimental translational and rotational diffusion coefficients of proteins to model solvent drag. The coupling of the friction matrix calculated using the bead model and the stiffness and mass matrices obtained from the ENM resulted in the Langevin modes of myosin. Due to solvent viscosity, the Langevin modes of the protein were significantly different from its normal modes computed in a vacuum. Additionally, the first Langevin modes were shown to be over-damped [12].

The bead models generally used in the LMA [12, 81] have some problematic aspects such as bead overlapping [82] and volume corrections for rotation [83] and viscosity [84]. Also, in these models for the calculation of the hydrodynamic interactions between pairs of atoms, it is assumed that the intervening space between the pairs is filled only with a solvent and the presence of other atoms in the space is totally ignored [81]. Additionally, although solvent friction takes place at the surface of proteins [19], the bead models used in the LMA assume that the frictional forces act at the centers of C_α. In contrast to the bead models [12, 81], the finite element method (FEM) can model solvent drag on the protein surface [13]. Additionally, the FEM encounters none of the above-mentioned problems of the bead models and the frictional forces acting on the surface converge to the exact solution when the finite element size is reduced to zero [13].

The solvent friction matrix used in the LMA may be computed by embedding the protein in a Stokes-fluid that is modeled using the FEM, as is commonly performed in FE fluid-solid interaction analyses [85, 86]. A unit velocity in each of the three x_1-, x_2-, and x_3-directions can be imposed on one node located on the protein surface, and the resultant forces acting on all the protein surface nodes can be calculated and substituted for the corresponding column of the friction matrix. To establish the whole friction matrix, $3M$ separate FE fluid simulations need to be performed using the available finite element software programs, which render this approach prohibitively costly. M is the number of protein surface nodes. However, since the flow around the protein surface is governed by the Stokes equations, the fluid can be modeled as a solid. Providing that the solid properties are chosen correctly, the whole friction matrix can be obtained accurately with one finite element solid simulation [13]. Hence, the computational cost is significantly reduced. The stiffness and mass matrices of the protein model can be calculated from the elastic-body approximation developed by M. Bathe [18]. Finally, the Langevin modes of the protein can be obtained using the friction, stiffness, and mass matrices from the FEM.

In this chapter, we first review the LMA developed by Lamm and Szabo [17] to incorporate the effects of solvent-damping into the standard NMA. Then, we present a new algorithm that calculates a solvent friction matrix using the FEM to account for the solvent-damping effects. The algorithm proves successful in calculating the diffusion coefficients of a sphere and 10 proteins with various molecular weights, ranging from 7 kDa to 233 kDa. We subsequently couple the solvent friction matrix and the stiffness and mass matrices calculated using the FEM [18] to obtain the Langevin modes and corresponding relaxation times of crambin, a small protein with 46 amino acids. The obtained results are then compared with those calculated using bead models [19].

2.1 Methods

2.1.1 Langevin mode analysis

Langevin mode analysis has been developed by Lamm and Szabo [17] to incorporate the effects of solvent-damping into the standard NMA [21]. The basic theory of Langevin modes is based on the Langevin dynamics formalism given below [81]:

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{Z}\dot{\mathbf{q}} + V'(\mathbf{q}) = \mathbf{f}(t) \tag{2.1}$$

where $\mathbf{M}$ is the $3N \times 3N$ diagonal mass matrix, $\mathbf{Z}$ is the $3N \times 3N$ friction matrix, V is the potential energy function, $\mathbf{q}$ is the position vector, $\dot{\mathbf{q}}$ is the velocity vector, $\ddot{\mathbf{q}}$ is the acceleration vector, and $\mathbf{f}(t)$ is the vector of external stochastic forces as a function of time that satisfies the following conditions:

$$\begin{aligned} \langle f_i(t) \rangle &= 0 \\ \langle f_i(t) \cdot f_j(t') \rangle &= \frac{2Z_{ij}\delta(t-t')}{k_B T} \end{aligned} \tag{2.2}$$

Here k_B is Boltzmann's constant, T is temperature, $\delta(t-t')$ is the Kronecker delta, $f_i(t)$ is component i of $\mathbf{f}(t)$ and Z_{ij} is the ij^{th} component of the viscous damping matrix. N is the number of particles in the Langevin dynamics model.

Expanding the potential energy function in a Taylor series around a minimum $\mathbf{q}^0$ and neglecting the terms higher than the quadratic order, we can obtain the Langevin equations governing the linearized protein response as follows [81]:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{Z}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f}(t) \tag{2.3}$$

where the ij^{th} component of the stiffness matrix $\mathbf{K}$ is,

$$K_{ij} = \frac{\partial^2 V}{\partial q_i \partial q_j} = \frac{\partial^2 V}{\partial x_i \partial x_j} \tag{2.4}$$

and the displacement vector $\mathbf{x}$ is,

$$\mathbf{x} = (\mathbf{q} - \mathbf{q}^0) \tag{2.5}$$

Eq. 2.3 can be recast into the following matrix form:

$$\begin{aligned}
\begin{pmatrix} \dot{\mathbf{x}} \\ \ddot{\mathbf{x}} \end{pmatrix} &= \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{M}^{-1}\mathbf{f}(t) \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{F} & -\boldsymbol{\gamma} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{R}(t) \end{pmatrix} \\
&= \mathbf{A} \begin{pmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{R}(t) \end{pmatrix}
\end{aligned} \tag{2.6}$$

where $\mathbf{I}$ is a $3N \times 3N$ identity matrix, and $\mathbf{F}$, $\boldsymbol{\gamma}$, and $\mathbf{R}(t)$ are, respectively, obtained from pre-multiplying the inverse of the mass matrix by the stiffness matrix, the friction matrix, and the vector of external stochastic forces.

Langevin modes and their corresponding eigenvalues can be obtained by diagonalizing the $6N \times 6N$ matrix $\mathbf{A}$ [12, 17, 81],

$$\mathbf{AW} = \mathbf{W}\boldsymbol{\Lambda} \tag{2.7}$$

Here $\mathbf{W}$ is the $6N \times 6N$ matrix containing the Langevin modes as columns and $\boldsymbol{\Lambda}$ is the $6N \times 6N$ diagonal matrix of eigenvalues.

LMA can be performed in the FEM, where N is the number of nodes in a protein FEM model and the stiffness, mass, and friction matrices are obtained from the FEM model.

2.1.2 Properties of Langevin modes

A Langevin mode consists of $6N$ elements, of which the upper half corresponds to displacements; the lower one, to velocities. As a result, the bottom $3N$ elements can be obtained by multiplying the corresponding eigenvalue by the top ones. The

$6N \times 6N$ matrix $\mathbf{W}$ can be denoted as [81],

$$\mathbf{W} = \begin{pmatrix} \mathbf{L} \\ \mathbf{L\Lambda} \end{pmatrix} \tag{2.8}$$

where $\mathbf{L}$ is a $3N \times 6N$ matrix containing the upper halves of Langevin modes.

Since $\mathbf{A}$ is non-symmetric, $\mathbf{L}$ and $\mathbf{\Lambda}$ are generally complex. Complex eigenvalues and their corresponding eigenvectors exist in conjugate pairs. Since the matrices $\mathbf{F}$ and $\boldsymbol{\gamma}$ are non-negative definite, the real components of the eigenvalues are non-positive. Also, the negative of the inverse of a real component is the relaxation time corresponding to the eigenvalue [81].

Additionally, the special structure of the non-symmetric matrix $\mathbf{A}$ allows us to factor the matrix into a product of two symmetric matrices [17]:

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{F} & -\boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & -\boldsymbol{\gamma} \end{pmatrix} \begin{pmatrix} -\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{2.9}$$

We can also analytically invert the first matrix in the right hand side of the above equation as follows:

$$\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & -\boldsymbol{\gamma} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\gamma} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \tag{2.10}$$

Using Eqs. 2.9 and 2.10, Eq. 2.7 can be recast into the following form [17]:

$$\begin{pmatrix} -\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{W} = \begin{pmatrix} \boldsymbol{\gamma} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \mathbf{W\Lambda} \tag{2.11}$$

The above equation is a generalized eigenvalue problem, and the eigenvectors can be normalized as follows:

$$\mathbf{W}^{\mathrm{T}} \begin{pmatrix} \boldsymbol{\gamma} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \mathbf{W} = \mathbf{I} \tag{2.12}$$

Then we can write the inverse of the matrix $\mathbf{W}$ as,

$$\mathbf{W}^{-1} = \mathbf{W}^{\mathrm{T}} \begin{pmatrix} \boldsymbol{\gamma} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \tag{2.13}$$

2.1.3 Calculation of the friction matrix from the FEM

The FEM is a well-established numerical procedure that is widely used in engineering [13, 87] (see for example Refs. [88, 89]). The method can be used to model the protein embedded in a Stokes-fluid and consequently calculate the solvent friction matrix, where the boundary of the protein is assumed to be the solvent-excluded surface (SES). The SES of a protein (Fig. 2-1-A) is defined as the closest contact point of a 1.4 Å radius solvent-probe rolled over the protein van der Waals surface [18]. We compute the SES by MSMS ver. 2.6.1 [52]. Subsequently, the surface is coarsened (Fig. 2-1-B) using the surface simplification algorithm QSLIM [53–55], as implemented in MeshLab [56]. Then the coarsened SES is imported into the finite element program ADINA ver. 8.7.1. The space from the SES to the surface of a sphere with a diameter of approximately 400 times the largest dimension of the protein is meshed with 8-node hexahedral elements (Figs. 2-1-C and 2-1-D). The element size changes from the finest (near the SES) to the coarsest (near the sphere surface) level in eleven layers, while the adjacent layers are glued to each other and the interfacing surfaces have the same displacements.

The fluid flow around the SES is commonly modeled as an incompressible, steady-state Stokes flow [19]. Considering a stationary Cartesian reference frame (x_i, i=1, 2, 3) and using index notation, the governing equations of the flow can be written as follows [13]:

momentum:
$$\frac{\partial \tau_{ij}}{\partial x_j} + f_i{}^B = 0 \tag{2.14}$$

constitutive:
$$\tau_{ij} = -p\delta_{ij} + 2\mu e_{ij} \tag{2.15}$$

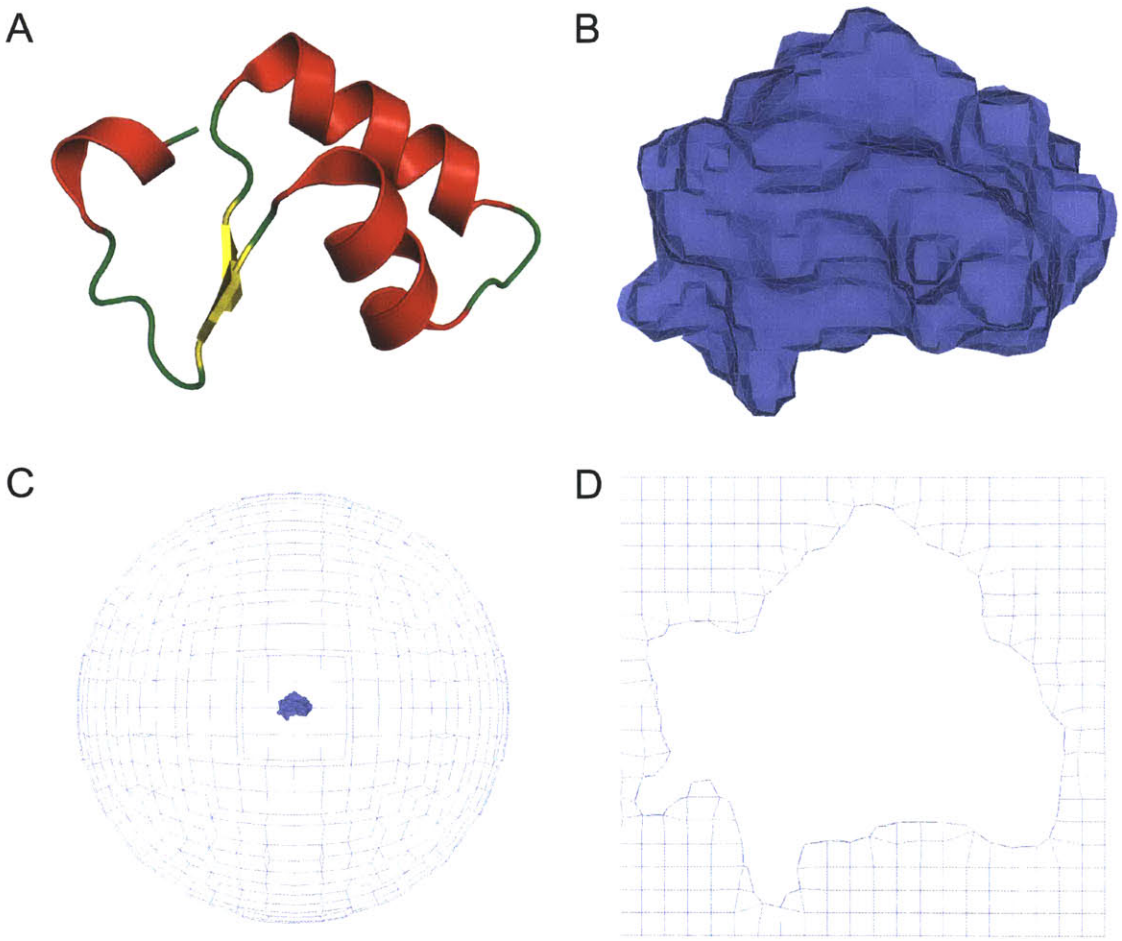

Figure 2-1 – Finite element solvent model of crambin (Protein Data Bank ID 2FD7). **A** shows the schematic representation of the energy-minimized molecular structure, which is colored according to its secondary structures; **B** shows the coarsened SES imported into ADINA; **C** shows the spherical volume mesh employed to model the solvent around the SES (for visual purposes, the size of the SES has been increased); **D** shows the close-up of the mesh surrounding the protein (in cross-section).

continuity: $$v_{i,i} = 0 \tag{2.16}$$

where,

v_i = velocity of fluid flow in direction x_i

τ_{ij} = components of stress tensor

f_i^B = components of body force vector

p = pressure

δ_{ij} = Kronecker delta

μ = fluid (laminar) viscosity

e_{ij} = components of velocity strain tensor $= \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)$

The above equations (Eqs. 2.14, 2.15, and 2.16) may be used in the FE fluid analysis of the solvent model to compute the friction matrix. In that case, velocities at the nodes on the sphere surface (Fig. 2-1-C) are set to zero. Additionally for one of the nodes, zero pressure is chosen. Then, a unit velocity in each of the three x_1-, x_2-, and x_3-directions may be applied to one of the nodes located on the protein surface, while the other velocity degrees of freedom of the protein surface nodes are set to zero. Subsequently, the resultant forces at the protein surface nodes are computed and inserted into the corresponding column of the friction matrix. We need to perform $3M$ separate FE fluid simulations using the commercial finite element software programs such as ADINA to calculate the whole friction matrix. This number of simulations render the calculation of the matrix infeasible. However, since the governing equations of motion for an incompressible, steady-state Stokes fluid flow (Eqs. 2.14, 2.15, and 2.16), under some circumstances, can be equivalent to those of an incompressible, isotropic, linear elastic solid (Eqs. 2.17 and 2.18), the flow around the SES can be modeled as the static displacement of the incompressible solid [13].

equilibrium:
$$\frac{\partial \tau_{ij}}{\partial x_j} + f_i{}^B = 0 \tag{2.17}$$

constitutive:
$$\tau_{ij} = -p\delta_{ij} + 2G\varepsilon'_{ij} \tag{2.18}$$

where,

G = shear modulus

ε'_{ij} = components of deviatoric strain tensor = $\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) - \frac{1}{3}\cancel{\frac{\partial u_i}{\partial x_i}}\delta_{ij}$

u_i = displacement of solid in direction x_i

The prerequisites for this equivalency to hold are that Poisson's ratio of the solid has to be chosen close to 0.5 (for example, 0.4999) and its shear modulus needs to be equal to the fluid viscosity. Additionally, the velocity in the incompressible, steady-state Stokes fluid flow is equivalent to the displacement in the incompressible solid.

To compute the friction matrix using the FE solid analysis of the solvent model, except for a fraction (20%), the nodes on the sphere surface (Fig. 2-1-C) are restrained in all three directions. This boundary condition approximately simulates zero velocity and pressure at infinity. Additionally, a unit displacement in each of the three x_1-, x_2-, and x_3-directions is applied to one of the nodes located on the protein surface, while the other displacement degrees of freedom of the protein surface nodes are set to zero. Subsequently, the resultant forces that are exerted on the nodes located on the SES are calculated and substituted for the corresponding column of the friction matrix. Similarly, the other columns can be calculated. Since the solid is incompressible, the displacement/pressure formulation must be used in the FEM [13]. Here 8-node displacement/pressure solid elements (8/1 elements) [13] are used in the FE solid analysis of the solvent model (Figs. 2-1-C and 2-1-D).

Unlike the calculation of the friction matrix using the FE fluid analysis of the

solvent model, the calculation of the matrix does not require $3M$ separate FE solid simulations. Instead, the stiffness matrix of the solvent model can be decomposed once [13] and used for different boundary conditions and loadings applied to calculate the friction matrix. Hence, calculating the friction matrix using the FE solid analysis is substantially faster than calculating that of the FE fluid analysis. The friction matrix calculated here is a $3M \times 3M$ matrix, $\tilde{\mathbf{Z}}$, that corresponds to the nodes on the protein surface. Subsequently, the $3N \times 3N$ friction matrix $\mathbf{Z}$ corresponding to all the nodes in the protein model can be obtained from $\tilde{\mathbf{Z}}$, where the components of $\mathbf{Z}$ which do not correspond to the nodes on the SES are zero.

2.1.4 Calculation of the friction matrix from bead models

Bead models [19] can be applied for the calculation of protein diffusion coefficients [76, 77, 90, 91]. Of the available models, the best results are obtained by use of the shell model [19]. In this model, the surface of proteins is covered by small beads and diffusion coefficients are calculated using Oseen or modified-Oseen tensors [75, 92, 93]. Bead models can also be used to analyze protein dynamics in a solvent. A combination of elastic network and bead models has proven successful in calculating protein Langevin modes [12].

Here the $3M \times 3M$ friction matrix, $\tilde{\mathbf{Z}}$, is calculated by use of a shell-type model. In this model, equal-size beads are positioned at the nodes located on the SES (Fig. 2-1-B) and a matrix $\mathbf{T}$ is obtained as follows [12]:

$$\begin{aligned}
\mathbf{T}_{ij} &= (6\pi\mu\sigma)^{-1}\mathbf{I} && \text{when } i = j \\
&= (8\pi\mu \mathrm{r}_{ij})^{-1}\left(\mathbf{I} + \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^{\mathrm{T}}}{\mathrm{r}_{ij}^{2}} + 2\frac{\sigma^2}{\mathrm{r}_{ij}^{2}}\left(\frac{1}{3}\mathbf{I} - \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^{\mathrm{T}}}{\mathrm{r}_{ij}^{2}}\right)\right) && \text{when } i \neq j \text{ and } \mathrm{r}_{ij} \geq 2\sigma \\
&= (6\pi\mu\sigma)^{-1}\left(\left(1 - \frac{9}{32}\frac{\mathrm{r}_{ij}}{\sigma}\right)\mathbf{I} + \frac{3}{32}\frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^{\mathrm{T}}}{\sigma \mathrm{r}_{ij}}\right) && \text{when } i \neq j \text{ and } \mathrm{r}_{ij} < 2\sigma
\end{aligned} \tag{2.19}$$

where σ is the hydrodynamic radius of the beads, $\mathbf{r}_{ij}$ is the vector from node i to node

j, and r_{ij} is its magnitude. $\tilde{\mathbf{Z}}$ is obtained by inverting $\mathbf{T}$. Subsequently, the $3N \times 3N$ friction matrix $\mathbf{Z}$ can be computed from $\tilde{\mathbf{Z}}$.

2.1.5 Calculation of diffusion coefficients from the friction matrix

A 6×6 resistance tensor, Ξ, is usually defined to express the hydrodynamic resistance of an object [19]. To calculate Ξ, first, we need to partition the $3M \times 3M$ friction matrix $\tilde{\mathbf{Z}}$ into 3×3 blocks, $\tilde{\mathbf{Z}}_{ij}$. Then Ξ can be obtained using the following equations:

$$\Xi_{\mathrm{tt}} = \sum_i \sum_j \tilde{\mathbf{Z}}_{ij} \tag{2.20}$$

$$\Xi_{\mathrm{tr}} = \sum_i \sum_j \mathbf{U}_i \tilde{\mathbf{Z}}_{ij} \tag{2.21}$$

$$\Xi_{\mathrm{rr}} = \sum_i \sum_j \mathbf{U}_i \tilde{\mathbf{Z}}_{ij} {\mathbf{U}_j}^{\mathrm{T}} \tag{2.22}$$

$$\Xi = \begin{pmatrix} \Xi_{\mathrm{tt}} & {\Xi_{\mathrm{tr}}}^{\mathrm{T}} \\ \Xi_{\mathrm{tr}} & \Xi_{\mathrm{rr}} \end{pmatrix} \tag{2.23}$$

$$\mathbf{U}_i = \begin{pmatrix} 0 & -z_i & y_i \\ z_i & 0 & -x_i \\ -y_i & x_i & 0 \end{pmatrix} \tag{2.24}$$

where Ξ_{tt}, Ξ_{rr}, and Ξ_{tr} are the 3×3 blocks of Ξ, which correspond to translation, rotation, and translation-rotation coupling, respectively, and x_i, y_i, and z_i are the coordinates of node i in the stationary Cartesian reference frame (x_j, j=1, 2, 3).

A 6×6 diffusion matrix, $\mathbf{D}$, can be obtained from Ξ using the generalized Einstein relationship,

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{\mathrm{tt}} & {\mathbf{D}_{\mathrm{tr}}}^{\mathrm{T}} \\ \mathbf{D}_{\mathrm{tr}} & \mathbf{D}_{\mathrm{rr}} \end{pmatrix} = k_B T \Xi^{-1} \tag{2.25}$$

where $\mathbf{D}$, like Ξ, has been partitioned into 3×3 blocks [19].

Translational (D_t) and rotational (D_r) diffusion coefficients can be computed as follows:

$$D_t = \frac{1}{3}\mathrm{Tr}\left(\mathbf{D}_{\mathrm{tt}}\right) \tag{2.26}$$

$$D_r = \frac{1}{3}\mathrm{Tr}\left(\mathbf{D}_{\mathrm{rr}}\right) \tag{2.27}$$

where the symbol Tr indicates the trace of a tensor. Consequently, translational (f_t) and rotational (f_r) friction coefficients can be obtained using the following equations [19]:

$$f_t = \frac{k_B T}{D_t} \tag{2.28}$$

$$f_r = \frac{k_B T}{D_r} \tag{2.29}$$

The blocks $\mathbf{D}_{\mathrm{tt}}$ and $\mathbf{D}_{\mathrm{tr}}$ depend on the origin of the Cartesian system, while $\mathbf{D}_{\mathrm{rr}}$ does not vary with the change of origin [94]. However, the diffusion matrix $\mathbf{D}$ has to be calculated at the center of diffusion, D, to best match the experimental diffusion coefficients [95], where D is a location at which $\mathbf{D}_{\mathrm{tr}}$ is symmetric [94]. To calculate the correct diffusion matrix, we can first compute the matrix at some arbitrary origin, O, and then transfer the origin to D using $\mathbf{r}_{\mathrm{OD}}$, the position vector of D with respect to O, and recalculate the matrix [19].

$$\mathbf{r}_{\mathrm{OD}} = \begin{pmatrix} x_{\mathrm{OD}} \\ y_{\mathrm{OD}} \\ z_{\mathrm{OD}} \end{pmatrix} = \begin{pmatrix} {D_{\mathrm{rr}}}^{yy} + {D_{\mathrm{rr}}}^{zz} & -{D_{\mathrm{rr}}}^{xy} & -{D_{\mathrm{rr}}}^{xz} \\ -{D_{\mathrm{rr}}}^{xy} & {D_{\mathrm{rr}}}^{xx} + {D_{\mathrm{rr}}}^{zz} & -{D_{\mathrm{rr}}}^{yz} \\ -{D_{\mathrm{rr}}}^{xz} & -{D_{\mathrm{rr}}}^{yz} & {D_{\mathrm{rr}}}^{xx} + {D_{\mathrm{rr}}}^{yy} \end{pmatrix}^{-1} \begin{pmatrix} {D_{\mathrm{tr}}}^{yz} - {D_{\mathrm{tr}}}^{zy} \\ {D_{\mathrm{tr}}}^{zx} - {D_{\mathrm{tr}}}^{xz} \\ {D_{\mathrm{tr}}}^{xy} - {D_{\mathrm{tr}}}^{yx} \end{pmatrix} \tag{2.30}$$

2.1.6 Calculation of the stiffness and mass matrices

Recently, the FEM has proven successful in calculating the lowest normal modes of proteins [18]. Since the lowest normal modes are determined dominantly by the shape of proteins [96], the material type used in the current FE models is not of great importance for obtaining the normal modes. Hence, proteins can be modeled simply as homogeneous, isotropic, linear elastic continua [18].

To calculate the stiffness and mass matrices of a protein, first, we use the SES explained above (Fig. 2-1-B) to define the volume of the protein model. Then the volume is discretized with 3D 4-node tetrahedral elements using the finite element software program ADINA ver. 8.7.1. Since the stiffness and mass matrices need to be coupled with the friction matrix, the nodes generated on the protein surface in this model have to be the same as those of the solvent model used for calculating the friction matrix (see Section 2.1.3).

Here the mass density of the protein model is defined as the division of the molecular mass by the volume of the model, and Poisson's ratio is set to 0.3 [18]. Knowing the mass density and Poisson's ratio, we can adjust Young's modulus of the protein model to fit the FEM α-carbon atom fluctuation results to the RTB results [18]. Given the protein volume and the material properties, the FEM can use numerical integrations to calculate the stiffness and mass matrices.

2.2 Results

2.2.1 Diffusion coefficients of a sphere with a radius of 25 Å surrounded by 20 °C water

There are analytical solutions for the diffusion coefficients of spheres [19]. Hence, to examine the accuracy of the FEM in calculating solvent friction matrices, one can compute the diffusion coefficients of a sphere from the FEM and compare the results with the analytical solutions. Also, this study can help us better model globular proteins, which have spherical shapes, within water.

Here we employ the FEM to calculate the diffusion coefficients of a 25-Å-radius sphere located in 20 °C water. To this end, first, we need to compute the friction matrix using the FEM. We embed the sphere with the radius, r_{in}, of 25 Å in an incompressible, isotropic, linear elastic solid outer sphere with the radius r_{out}. As explained in Section 2.1.3, Poisson's ratio of the solid material is set to 0.4999 and its shear modulus, G, is chosen equal to 60.34 $\frac{\text{Da}}{(\text{psec}\times\text{Å})}$, which is the viscosity, μ, of water at 20 °C [90]. Considering the dynamic behavior and molecular properties of proteins, dalton (Da), angstrom (Å), and picosecond (psec) are, respectively, chosen as the mass, length, and time units used in the FEM. The following equations define the chosen set of non-SI units based on the SI units:

$$\begin{aligned} \text{Da} &= 1.6605 \times 10^{-27}\ \text{kg} \\ \text{Å} &= 10^{-10}\ \text{m} \\ \text{psec} &= 10^{-12}\ \text{sec} \end{aligned} \tag{2.31}$$

As described before, the intervening space between the inner and outer sphere surfaces is meshed with 8/1 solid elements [13]. The element size varies from the finest (near the inner sphere surface) to the coarsest (near the outer sphere surface) level in several layers, while the adjacent layers are glued to each other (Fig. 2-2).

To simulate the zero velocity and pressure at infinity, except for a fraction, r_{free}, the nodes located on the outer sphere surface are restrained in all three directions. Subsequently, as explained before, the friction matrix and the diffusion coefficients can be computed by applying different displacements and boundary conditions to the nodes located on the inner sphere surface. Results show that changing r_{free}, in the range of 0.1 to 0.9, has almost no effect on the errors in the calculated diffusion coefficients (Fig. 2-3). Hence, from here on in the FE models, r_{free} is set to 0.2.

In this chapter, the error in a calculated diffusion coefficient is defined as follows:

$$\text{Error} = 100 \left| \frac{{D_{\text{t/r}}}^{\text{exact}} - {D_{\text{t/r}}}^{\text{calc}}}{{D_{\text{t/r}}}^{\text{exact}}} \right| \tag{2.32}$$

where ${D_{\text{t/r}}}^{\text{exact}}$ is the exact value of the translational (t) or rotational (r) diffusion coefficient obtained from experiments or analytical solutions, and ${D_{\text{t/r}}}^{\text{calc}}$ is the cal-

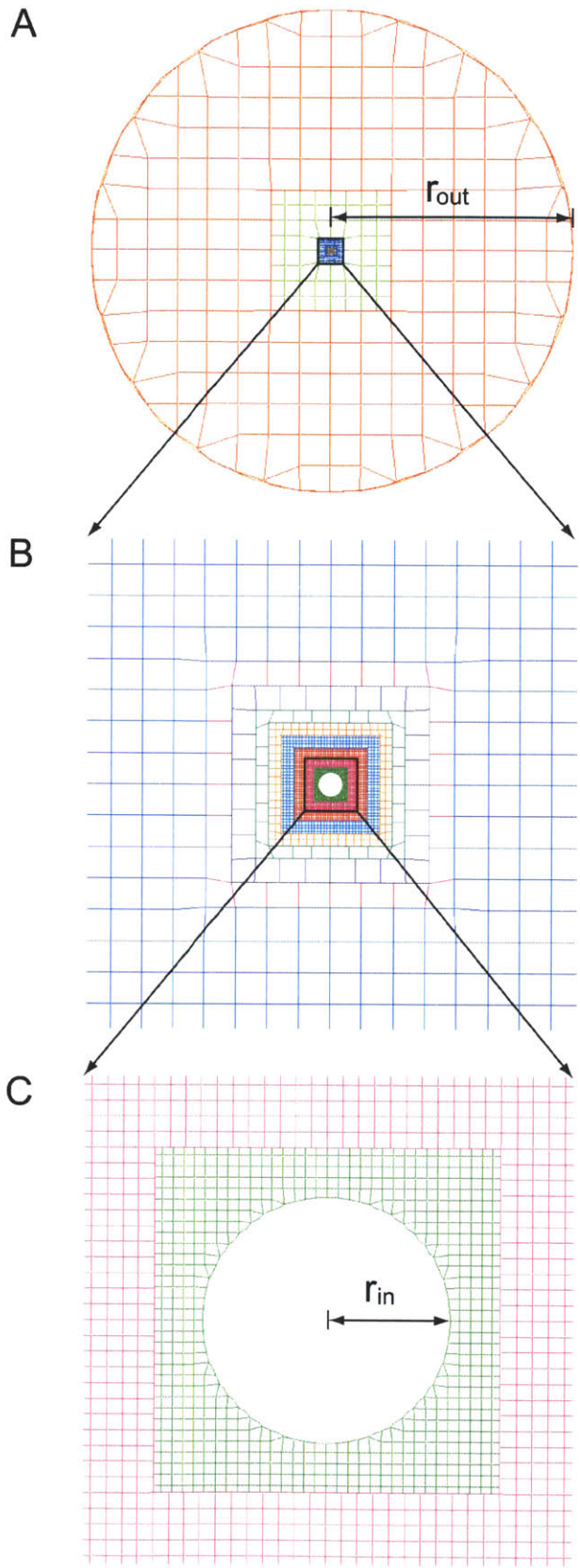

Figure 2-2 – The mesh between the inner and outer sphere surfaces (in cross-section). **A** shows all the layers of mesh used in the FE model, while **B** and **C** show only 9 and 2 layers, respectively, surrounding the inner sphere. Different colors indicate different layers of mesh.

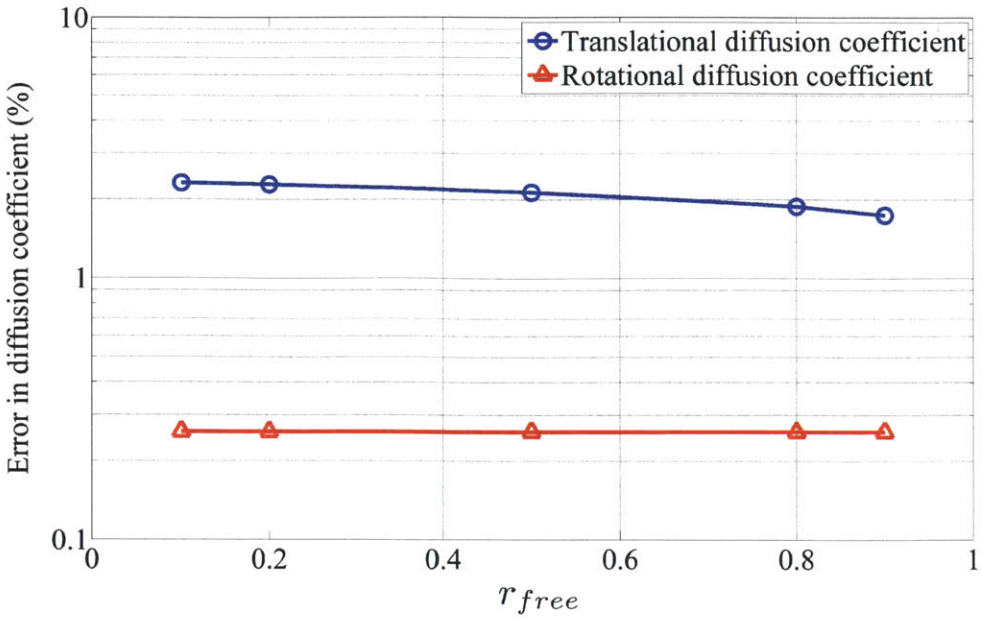

Figure 2-3 – Error in the calculated translational and rotational diffusion coefficients of the inner sphere versus the fraction of the nodes on the outer sphere surface that are unrestrained, r_{free}. Here r_{out} is 400 times r_{in} and the ratio of the radius of the inner sphere, r_{in}, to the element size near the inner sphere surface, h, is chosen equal to 11.43.

culated value of the diffusion coefficient. The exact values of the translational and rotational diffusion coefficients of the sphere with the radius of 25 Å in 20 °C water obtained from analytical solutions [19] are,

$$D_{\mathrm{t}}^{\mathrm{exact}} = \frac{k_B T}{6\pi\mu r_{in}} = 8.57 \times 10^{-3} \ \frac{\text{Å}^2}{\text{psec}}$$

$$D_{\mathrm{r}}^{\mathrm{exact}} = \frac{k_B T}{8\pi\mu {r_{in}}^3} = 1.03 \times 10^{-5} \ \frac{1}{\text{psec}}$$

Additionally, changing the ratio of r_{out} to r_{in}, from 5 to 400, significantly decreases the errors in the calculated diffusion coefficients, while the errors remain almost constant for the ratios greater than 400 (Fig. 2-4). Since the ratios greater than 400 substantially increase the cost of computation without a significant decrease in the errors, we choose the ratio to be 400 for the FE simulations.

Setting r_{free} to 0.2 and the ratio of r_{out} to r_{in} to 400, we can check the convergence of the calculated diffusion coefficients by changing the element sizes. The element

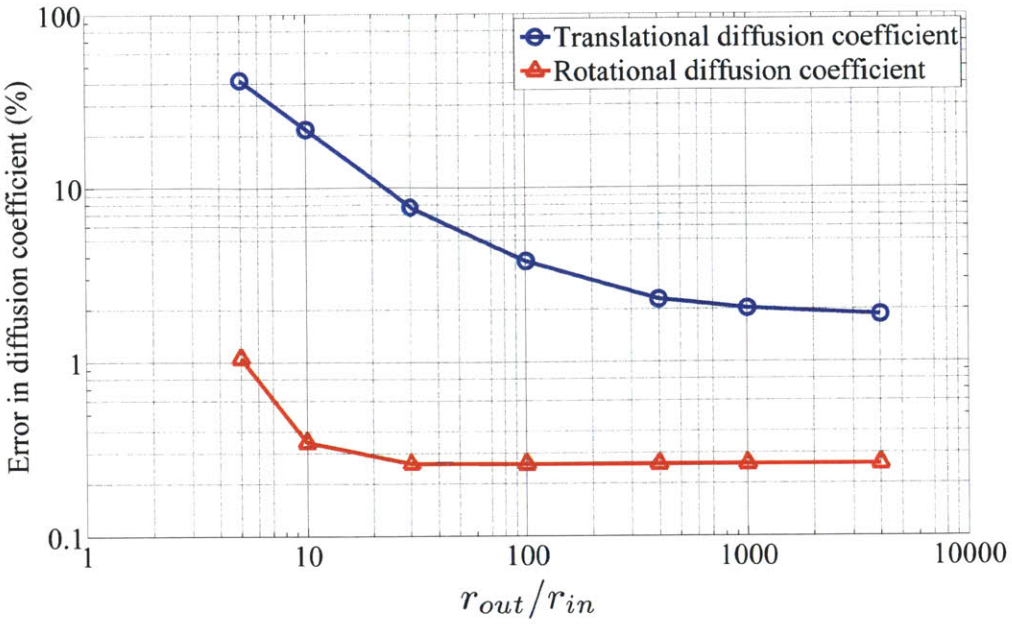

Figure 2-4 – Error in the calculated translational and rotational diffusion coefficients of the inner sphere versus the ratio of r_{out} to r_{in}. Here r_{free} is set to 0.2 and the ratio of the radius of the inner sphere, r_{in}, to the element size near the inner sphere surface, h, is chosen equal to 11.43.

sizes of all the layers are defined based on the element size near the inner sphere surface, h. When we change the ratio of r_{in} to h from 1.429 to 11.43, the error in the translational diffusion coefficient reduces from 46.31% to 2.28%, and that of the translational diffusion coefficient decreases from 29.16% to 0.26% (Fig. 2-5). Although the cost of computation for the ratio equal to 11.43 is reasonable (the number of elements used in the FE model is about 100,000), the errors are remarkably small. We use ratios similar to 11.43 for the FE simulations used for globular proteins.

As seen above, setting r_{free} to 0.2, the ratio of r_{out} to r_{in} to 400, and the ratio of r_{in} to h to 11.43, with a reasonable cost of computation, we can obtain accurate results for the diffusion coefficients of a sphere. Since globular proteins have spherical shapes, we can choose similar parameters to obtain accurate friction matrices for the proteins.

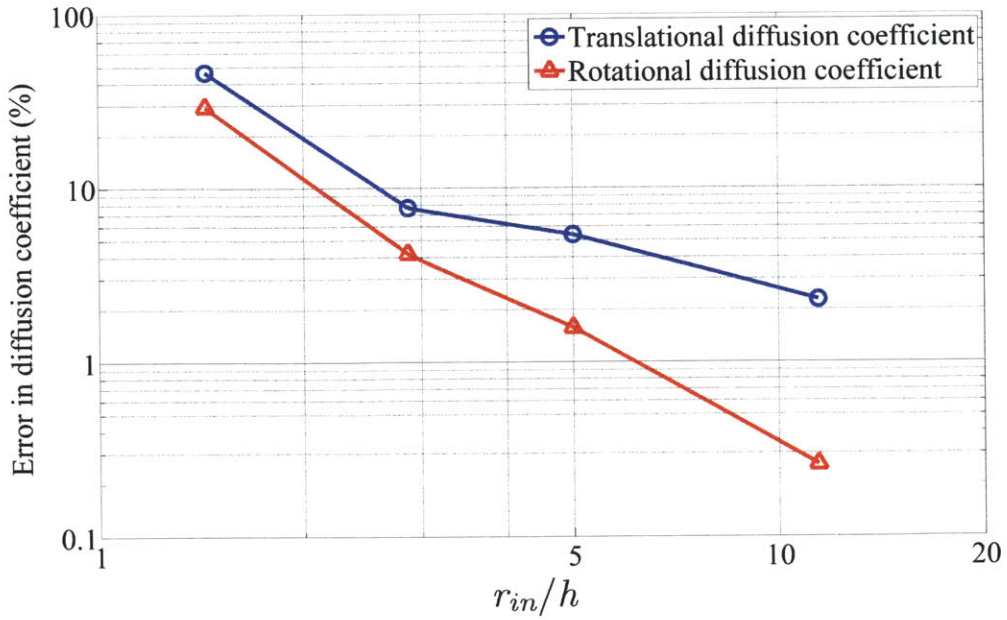

Figure 2-5 – Error in the calculated translational and rotational diffusion coefficients of the inner sphere versus the ratio of r_{in} to h. Here r_{free} is set to 0.2 and the ratio of r_{out} to r_{in} is chosen equal to 400.

2.2.2 Diffusion coefficients of proteins

We have calculated the rotational and translational diffusion coefficients of 10 different proteins (Table 2.1) using the FEM. To calculate the coefficients, as explained before, the SES of the protein is embedded in an incompressible solid sphere with a diameter of 400 times the largest dimension of the protein. The mechanical properties of the solid are the same as in the previous section. According to the results in the previous section, we mesh the whole intervening space between the SES and the sphere surface with approximately 100,000 8/1 elements and set r_{free} to 0.2. The diffusion coefficients are calculated for the hydration layer thicknesses of 0 and 1 Å (Table 2.2). To compute the SES of a protein with a hydration layer thickness of 1 Å, we need to increase the original van der Waals radii of atoms by 1 Å, before creating the corresponding SES.

After calculating the coefficients for the two different layer thicknesses, we calibrate the hydration layer thickness of the protein to the experimental diffusion coefficients of the protein, by performing interpolations and extrapolations on the data

Table 2.1 – Experimental values of the translational and rotational diffusion coefficients of 10 different proteins. All the reported diffusion coefficients are obtained from Ref. [77], except for those of myoglobin and hemoglobin that are obtained from Ref. [12].

Protein name	PDB ID	Structure weight (kDa)	Translational diffusion coefficient ($10^{-2} \times \frac{\text{Å}^2}{\text{psec}}$)	Rotational diffusion coefficient ($10^{-5} \times \frac{1}{\text{psec}}$)
BPTI (q)	4pti	7	1.29	4.17
Ribonuclease A	1rbx	14	1.07	–
Lysozyme	193l	14	1.09	2.6
Myoglobin	1dwr	18	1.02	1.72
Chymotrypsinogen	2cga	26	0.93	–
β-Lactoglobulin	1beb	37	0.782	0.75
Hemoglobin	2dn2	65	0.69	0.442
GPD (r)	4gpd	143	0.5	–
Adolase	1ado	158	0.445	–
Nitrogenase MoFe	2min	233	0.4	–

Table 2.2 – Calculated values of the translational and rotational diffusion coefficients of 10 different proteins for the hydration layer thicknesses of 0 and 1 Å.

Protein name	Translational diffusion coefficient $(10^{-2} \times \frac{\text{Å}^2}{\text{psec}})$	Rotational diffusion coefficient $(10^{-5} \times \frac{1}{\text{psec}})$	Translational diffusion coefficient $(10^{-2} \times \frac{\text{Å}^2}{\text{psec}})$	Rotational diffusion coefficient $(10^{-5} \times \frac{1}{\text{psec}})$
	Hydration layer thickness = 0 Å		Hydration layer thickness = 1 Å	
BPTI (q)	1.49	5.30	1.35	4.03
Ribonuclease A	1.17	2.56	1.08	2.05
Lysozyme	1.17	2.64	1.09	2.12
Myoglobin	1.08	1.87	1.01	1.52
Chymotrypsinogen	0.950	1.44	0.889	1.19
β-Lactoglobulin	0.829	0.926	0.768	0.741
Hemoglobin	0.697	0.524	0.660	0.448
GPD (r)	0.508	0.217	0.483	0.188
Adolase	0.474	0.170	0.454	0.151
Nitrogenase MoFe	0.446	0.147	0.429	0.131

Table 2.3 – Calculated values of the optimal hydration layer thicknesses and the errors in the translational and rotational diffusion coefficients of 10 different proteins. Errors cannot be calculated for cases for which there is no experimental data reported in Table 2.1.

Protein name	Optimal hydration layer thickness (Å)	Error in the calculated translational diffusion coefficient (%)	Error in the calculated rotational diffusion coefficient (%)
BPTI (q)	0.888	5.5	1.2
Ribonuclease A	1.09	0.1	–
Lysozyme	0.070	6.5	1
Myoglobin	0.447	2.2	1.6
Chymotrypsinogen	0.333	0.3	–
β-Lactoglobulin	0.954	1.5	0.4
Hemoglobin	1.08	4.6	0.5
GPD (r)	0.334	0.5	–
Adolase	1.44	0.7	–
Nitrogenase MoFe	2.69	1.7	–

given in Tables 2.1 and 2.2. Then the determined optimal thickness is added to the van der Waals radii of atoms and the diffusion coefficients are recalculated. All the experimental and calculated diffusion coefficients are given for the viscosity of water at 20 °C, except for the rotational diffusion coefficients of myoglobin and hemoglobin that are for the viscosity of 1.1cP (equivalent to 66.24 $\frac{\text{Da}}{(\text{psec}\times\text{Å})}$). As seen, the errors in the calculated diffusion coefficients are small (Table 2.3). Hence, the FEM can be used to obtain the friction matrix and the diffusion coefficients of proteins accurately.

2.2.3 Langevin modes of crambin

The initial structure of crambin, a small protein with 46 amino acids, is obtained from the work of Bang et al. [97] (Protein Data Bank ID 2FD7). In CHARMM version 35b1 [46] using the implicit solvation model EEF1 [47], steepest descent minimization followed by adopted-basis Newton-Raphson minimization is performed in the presence of successively reduced harmonic constraints on backbone atoms to achieve a final root-mean-square (RMS) energy gradient of 4×10^{-4} $\frac{\text{kcal}}{(\text{mol} \times \text{Å})}$ with corresponding RMS deviation between the X-ray and energy minimized structures of 1.1 Å (Fig. 2-1-A).

Then the SES of the energy-minimized molecular structure is obtained, as explained before, and imported into ADINA ver. 8.7.1 (Fig. 2-1-B). The SES is embedded in an incompressible solid sphere with Poisson's ratio of 0.4999 and shear modulus of 54.31 $\frac{\text{Da}}{(\text{psec} \times \text{Å})}$, which is 0.9 times the viscosity of water at 20 °C, μ_{20}. As in the previous section, the diameter of the sphere is approximately 400 times the largest dimension of the protein, the whole intervening space between the SES and the sphere surface is meshed with approximately 100,000 8/1 elements, and r_{free} is set to 0.2. Here we set the hydration layer thickness to zero and calculate the friction matrix $\mathbf{Z}$ using the FE model. The translational and rotational diffusion coefficients calculated from the $\mathbf{Z}$ matrix are, respectively, 1.88×10^{-2} $\frac{\text{Å}^2}{\text{psec}}$ and 8.56×10^{-5} $\frac{1}{\text{psec}}$. Note that for cases in which we know the experimental diffusion coefficients of a protein, we can calculate the $\mathbf{Z}$ matrix with the optimal hydration layer thickness. Otherwise, we set the thickness to zero.

For the sake of comparison, we also calculate the $\mathbf{Z}$ matrix from a shell-type model, as explained before. The hydrodynamic radius of beads located at the nodes on the SES is chosen such that the diffusion coefficients calculated from the shell model are the closest to those obtained using the FEM. For crambin, the hydrodynamic radius is found to be 0.4 Å. The errors in the translational and rotational diffusion coefficients calculated from the shell model in comparison with those of the FEM are, respectively, 1.7 % and 0.1 %. Hence, both of the friction matrices, obtained using the FEM and the shell model, give almost the same diffusion coefficients.

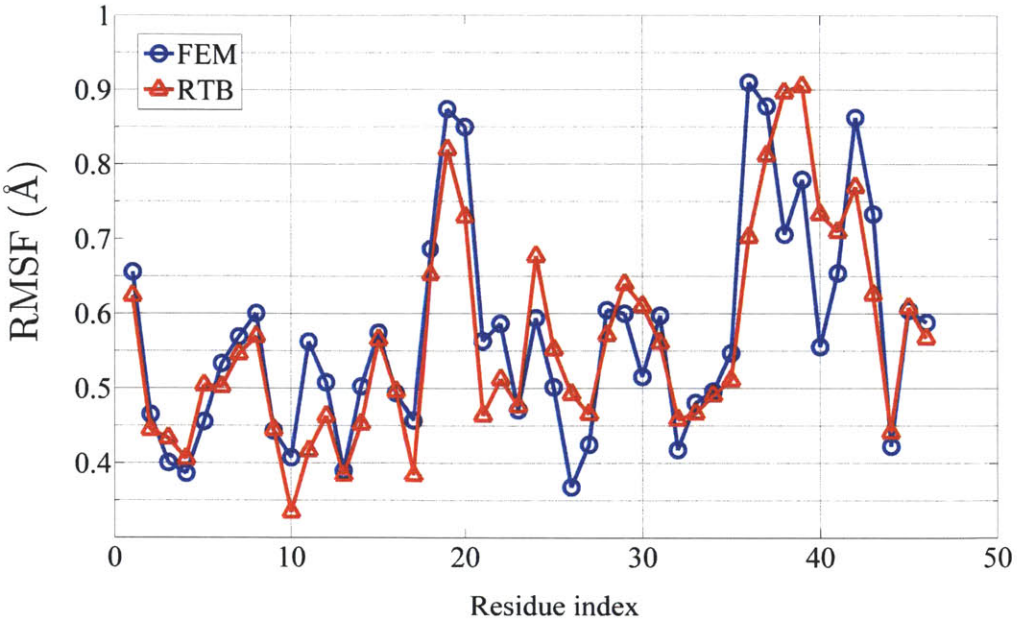

Figure 2-6 – Root-mean-square fluctuations of α-carbons of crambin obtained using the FEM and the RTB procedure.

To obtain the stiffness and mass matrices for crambin, we import the SES of crambin into ADINA ver. 8.7.1 and mesh the protein volume with 3D 4-node tetrahedral elements, where 2449 nodes are located in the model. The material of the protein is assumed homogeneous, isotropic, and linear elastic, where the mass density is 1525 $\frac{\text{kg}}{\text{m}^3}$ (equivalent to 0.918 $\frac{\text{Da}}{\text{Å}^3}$), Young's modulus is 7.8 GPa (equivalent to 467 $\frac{\text{Da}}{(\text{psec}^2 \times \text{Å})}$), and Poisson's ratio is 0.3. The mass density is obtained from the molecular weight and molecular volume of the protein. Young's modulus is obtained by fitting root-mean-square fluctuations of α-carbon atoms in the FE model to those obtained using the Rotation Translation Block procedure [57, 58] at room temperature in CHARMM (Fig. 2-6), where one block per residue and the implicit solvation model EEF1 [47] are employed.

Having the stiffness and mass matrices using the FEM and the friction matrix using the FEM or the shell model, we can solve for the Langevin modes of crambin. Additionally, we can calculate the normal modes of the protein in a vacuum by use of the stiffness and mass matrices (Eq. 1.1), where the friction matrix is totally neglected. Each of the calculated Langevin modes and vacuum normal modes contain $6N$ and $3N$

components, respectively, where N is equal to 2449. To compare a vacuum normal mode with a critically damped or over-damped Langevin mode, first, we take the upper $3N$ components of the Langevin mode and renormalize them. Then the dot product of the normalized vacuum normal mode and the normalized, upper half of the Langevin mode is calculated as an overlap score between the two modes [12].

The overlap score between the critically damped or over-damped Langevin mode that correlates most with a vacuum normal mode and the normal mode is the highest in comparison with all other critically damped or over-damped Langevin modes. To compare the two sets of critically damped or over-damped Langevin modes obtained by use of the two friction matrices, we can calculate their highest overlap scores with the first 10 non-zero vacuum normal modes (Table 2.4). The values of the highest overlap scores between the Langevin modes and the vacuum normal modes decrease with the vacuum normal mode number. This observation shows that only the lowest few vacuum normal modes of crambin can be characterized by individual critically damped or over-damped Langevin modes. Additionally, the two sets of the relaxation times corresponding to the first 10 vacuum normal modes obtained using the FEM and the shell model are in good agreement for vacuum normal modes 1, 2, 3, 5, 6, and 8, while, unlike the FEM results, the relaxation times calculated using the shell model for the other vacuum normal modes are substantially small (see also Ref. [12]). The short relaxation times calculated by the shell model would be expected to be far higher, since the lowest vacuum normal modes are collective motions of the protein.

At the solvent viscosity of zero, all the Langevin modes of crambin, except for the 12 zero-eigenvalue modes corresponding to the purely translational and rotational motions of the protein, are under-damped, i.e., the imaginary parts of their eigenvalues are non-zero. However, as the solvent viscosity increases, the number of under-damped modes reduces and the Langevin modes become critically damped and then over-damped (Table 2.5). Additionally, we expect the relaxation times of the Langevin modes to increase with the solvent viscosity (Fig. 2-7). Note that none of the relaxation times of the critically damped or over-damped Langevin modes of crambin calculated here are longer than the relaxation time corresponding to the

Table 2.4 – Highest overlap scores and corresponding critically damped or over-damped Langevin modes and relaxation times for the 10 lowest non-zero vacuum normal modes of crambin. The friction matrix **Z** is calculated using the FEM and the shell model, while the stiffness and mass matrices are calculated using only the FEM. Here the solvent viscosity is 0.9 times μ_{20}.

	The FEM			The shell model		
Vacuum normal mode	Highest overlap scores	Langevin mode	Relaxation time (psec)	Highest overlap	Langevin mode	Relaxation time (psec)
1	0.9066	29	13.61	0.9052	7	14.00
2	0.9598	31	11.46	0.9576	8	11.82
3	0.6885	33	9.95	0.6027	9	9.84
4	0.5194	54	4.70	0.6166	8775	0.03
5	0.5420	35	8.38	0.4891	10	8.50
6	0.5504	41	6.93	0.5078	12	7.25
7	0.3770	520	0.92	0.4905	3841	0.07
8	0.4023	50	5.00	0.4185	15	5.92
9	0.3475	472	1.02	0.5421	3976	0.06
10	0.4946	55	4.56	0.4607	10633	0.02

Table 2.5 – Number of critically damped or over-damped Langevin modes of crambin at different solvent viscosities. Here the friction matrix $\mathbf{Z}$ is calculated using only the FEM.

Solvent viscosity ($\times\mu_{20}$)	Number of critically damped or over-damped Langevin modes
0	12
0.1	1556
0.3	3530
0.5	4966
0.9	6262
1	6438

rotational diffusion coefficient. That is because all of the SES is involved in rotational diffusion while, in a Langevin mode, only part of the protein surface undergoes significant motion [12].

2.3 Concluding remarks

The solvent friction matrices of proteins need to be calculated and coupled with the stiffness and mass matrices of the proteins to incorporate the effects of solvent-damping into protein NMA. Several methods such as the FEM and the RTB procedure have already proven successful in calculating the stiffness and mass matrices of proteins [18]. However, none of the approaches that are currently used to compute the friction matrices for the LMA of proteins [12, 81] are expected to capture the true solvent-damping effects on proteins, due to their problematic aspects and unrealistic approximations [82–84].

The objective of this chapter was to present an accurate algorithm for calculating the Langevin modes of proteins. To this end, we employed the well-established FEM to calculate the solvent friction matrices of proteins. The FEM is well suited to the calculation of the friction matrices because, using the Stokes equations, it almost

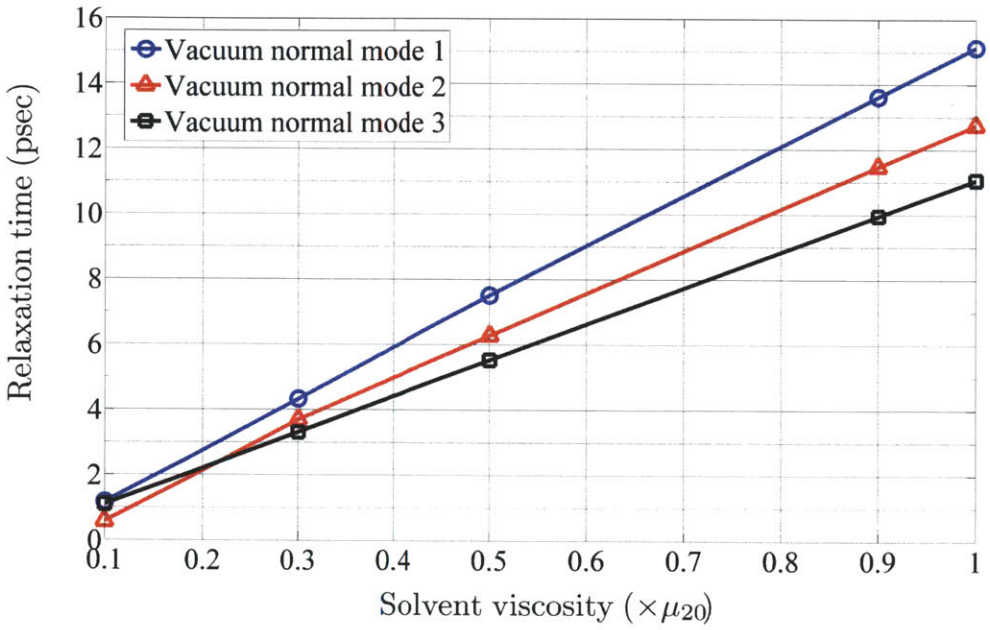

Figure 2-7 – Relaxation times of the critically damped or over-damped Langevin modes of crambin calculated for different solvent viscosities that heavily correlate with non-zero vacuum normal modes 1–3 of crambin.

models the true solvent drag on the protein surface, where the friction matrix is expected to converge to the exact solution as the mesh element size goes to zero [13]. Also, since the Stokes equations are employed here, only one FE solid simulation with the commercial finite element software programs is enough to calculate the whole friction matrix of a protein, which renders the computational efficiency of the FEM, for the LMA of proteins, high.

Since there are analytical solutions for the diffusional behavior of a sphere, to check the accuracy of the algorithm, the FEM was first applied to a sphere and the solvent friction matrix of the sphere was computed. Results show that the diffusion coefficients calculated from the matrix match well with the analytical solutions. Hence, since globular proteins have spherical shapes, the FEM is also expected to be useful in calculating accurate friction matrices for proteins.

The FEM was then employed to calculate the diffusion coefficients of 10 different proteins. Here it was shown that the error between the FEM-calculated and experimental diffusion coefficients of the proteins is small.

Finally, the Langevin modes of crambin were obtained using the FEM. As expected, all of the relaxation times of the critically damped or over-damped Langevin modes of crambin calculated using the FEM were longer than the relaxation time corresponding to the rotational diffusion coefficient. Additionally, it was demonstrated that only the first few non-zero vacuum normal modes of crambin could be well-characterized by individual critically damped or over-damped Langevin modes of the protein.

Chapter 3

Structure, evolutionary conservation, and conformational dynamics of *Homo sapiens* fascin-1, an F-actin crosslinking protein

The actin cytoskeleton of eukaryotic cells is centrally involved in a range of cellular functions including migration, endocytosis and division. In each case, the cell uses a host of accessory proteins to regulate actin dynamics spatiotemporally to achieve cellular function. In the case of migration, the leading edge of the cell consists predominantly of two types of protrusive filamentous actin (F-actin) structures: the dendritic sheetlike lamellipodium [98–102] and dynamic cortical spike-like filopodia [103–105]. Filopodia consist of compact ordered bundles of unipolar actin filaments that are nucleated at the leading edge by formins [106, 107] and crosslinked tightly into highly ordered bundles by the actin-binding protein fascin, which is both highly conserved evolutionarily and tissue- and cell-type-specific [104, 108]. While fascin dominates in cortical actin bundles such as filopodia, oocyte microvilli, and the den-

drites of dendritic cells, which play direct roles in cell migration, cell-matrix adhesion, and cell-cell interactions, fascin is also associated to a lesser extent with cytoplasmic actin bundles that participate in the maintenance of internal cell architecture [109]. Fascin is additionally known to be upregulated in a number of highly motile cell phenotypes including invasive cancer cells [110–114].

It is well established that the molecular size and conformational flexibility of actin crosslinking proteins are highly correlated with the morphological cytoskeletal structures with which they are associated [115–118]. For example, human filamin is a relatively large homodimer (approximately 160 nm molecular dimension) consisting of 24 tandem immunoglobulin repeats that is associated primarily with dendritic cytoskeletal networks [119], α-actinin is a smaller (35 nm) anti-parallel homodimer that is associated with both networks and bundles such as the contractile ring in dividing cells and stress fibers in adherent cells [120, 121], and the compact ABP fimbrin (10 nm) [122] (also called plastin) is found nearly exclusively in highly ordered unipolar actin bundles such as microvilli. The X-ray structures of several of these and related actin-crosslinking proteins have been determined. They consist of dual calponin homology actin-binding domains that are suggested to stabilize actin filaments [123–129]. Fimbrin additionally has two N-terminal calcium-binding EF-hand motifs that confer calcium regulation of its F-actin crosslinking activity in human isoforms [130, 131].

Homo sapiens fascin-1 also crosslinks actin filaments into compact unipolar bundles, but in a manner that is regulated by phosphorylation of serine 39 by protein kinase C [132], which inhibits its actin-bundling activity without affecting its localization to the filopodial tip complex [104]. *H. sapiens* fascin-1 is a compact, 55 kDa (493 residues) globular monomer with putative actin-binding domains that differ in primary sequence. Fascin was originally discovered in extracts from sea urchin eggs [133] and was later found in other invertebrates as well as vertebrates including *Drosophilia* [134], starfish sperm [135], *Xenopus laevis* [136], rodents [137] and humans [138]. *H. sapiens* fascin-1 is the original vertebrate fascin discovered, whereas retinal and testis fascins were discovered later and named fascin-2 and fascin-3, respectively [108]. Se-

quence alignment shows no similarity between fascins and other known actin-binding proteins in humans, but strong similarities within the fascin family itself [139]. Atomic models for actin-fascin bundles have been proposed on the basis of optical diffraction studies of negatively stained reconstituted material [140]. These suggest an 11 nm transverse banding pattern and uniformly polarized actin filaments organized in a hexagonal array with an interfilament distance of 11.5 nm. Assuming a single fascin monomer per crosslink, the predicted fascin-actin stoichiometry of 1:4.5 is in good agreement with experimentally measured ratios [141]. This maximal stoichiometry is a result of the helical twist of the actin filament that limits crosslinking sites between pairs of filaments to every fourth or fifth actin monomer. Fascin is additionally suggested to impart unusual mechanical stiffness to actin bundles in cells [104, 142–144] and reconstituted actin systems [121, 145, 146] as compared with fimbrin, despite similar actin-binding affinities and bundle structure [140, 147]. Despite its importance to cell function, the molecular basis for the unique localization and actin-bundling properties of fascin are not known. Here, we examine the packing, evolutionary sequence conservation and conformational flexibility of the 2.9 Å resolution crystal structure of full-length recombinant *H. sapiens* fascin-1 (PDB ID 1DFC). The results of normal mode analysis (NMA) of fascin suggest potential functional consequences of its distinct molecular structure and flexibility in crosslinking F-actin. Mutational and other experimental studies are needed to test these predictions.

3.1 Results

3.1.1 Overall structure

The two fascin molecules in the asymmetric unit are highly similar, with a root-mean-square deviation (RMSD) of 0.64 Å for 474 C_α atom pair equivalences. Fascin is composed of four tandem repeat β-trefoil domains (Fig. 3-1) with pseudo 3-fold symmetry, consisting of 12 β-strands that form the barrel (B) and cap (C) regions in the sequence: BCCBBCCBBCCB (Fig. 3-2-A). The domains pack to form a distorted

tetrahedron composed of two lobes: domains F1 (residues 8–139) and F2 (residues 140–260) form the first lobe and domains F3 (residues 261–381) and F4 (residues 382–493) form the second lobe. These two lobes can be superimposed with an RMSD of 2.3 Å for 224 C_α atom pair equivalences. The two lobes are related by an approximate 2-fold axis (rotation angle 163°, screw distance 8.1 Å). This axis passes between the two lobes and is approximately perpendicular to the plane of the image in Fig. 3-1-A. Despite the similarity of the two fascin lobes, the difference in relative internal orientations of their domains is 9°. Overall, 6676 $Å^2$ of accessible surface area is buried at domain-domain interfaces, with approximately one-third of this surface area buried in the interface between domains F1 and F2, one-quarter between domains F3 and F4, 18% between domains F2 and F4, 11% between domains F1 and F4 and 11% between domains F2 and F3 (Table C.1).

Within the first lobe, the β-trefoil domains F1 and F2 are related by an approximate pseudo 2-fold axis (rotation 170°, screw 6.2 Å) passing between these domains and lying approximately vertically in the plane of the image in Fig. 3-1-A, with the C-terminus of the first domain linked directly to the N-terminus of the second. Within the second lobe, the β-trefoil domains F3 and F4 are also related by an approximate 2-fold axis (rotation 169°, screw 7.3 Å) passing between these domains and lying approximately vertically in the plane of the image in Fig. 3-1-A. Domains F3 and F4 in the second lobe are also packed head-to-tail like domains F1 and F2 in the first lobe. The interface between the domains in each lobe is formed by interactions between the β-barrels, with their β-hairpin triplets at the extreme ends of the lobes. This association results in an almost coaxial alignment of the pseudo 3-fold axes of the two domains within each end-capped β-barrel lobe (Figs. 3-1-C and 3-1-D), with the pseudo 3-fold axis lying approximately perpendicularly to the inter-domain dyad.

3.1.2 β-Trefoil domain structure

Each β-trefoil domain consists of 12 β-strands that form three structurally homologous subunits (green, red and blue in Fig. 3-2-A). These three subunits are related by a pseudo 3-fold axis and each contains a βββ-loop-β motif [148] that forms two pairs

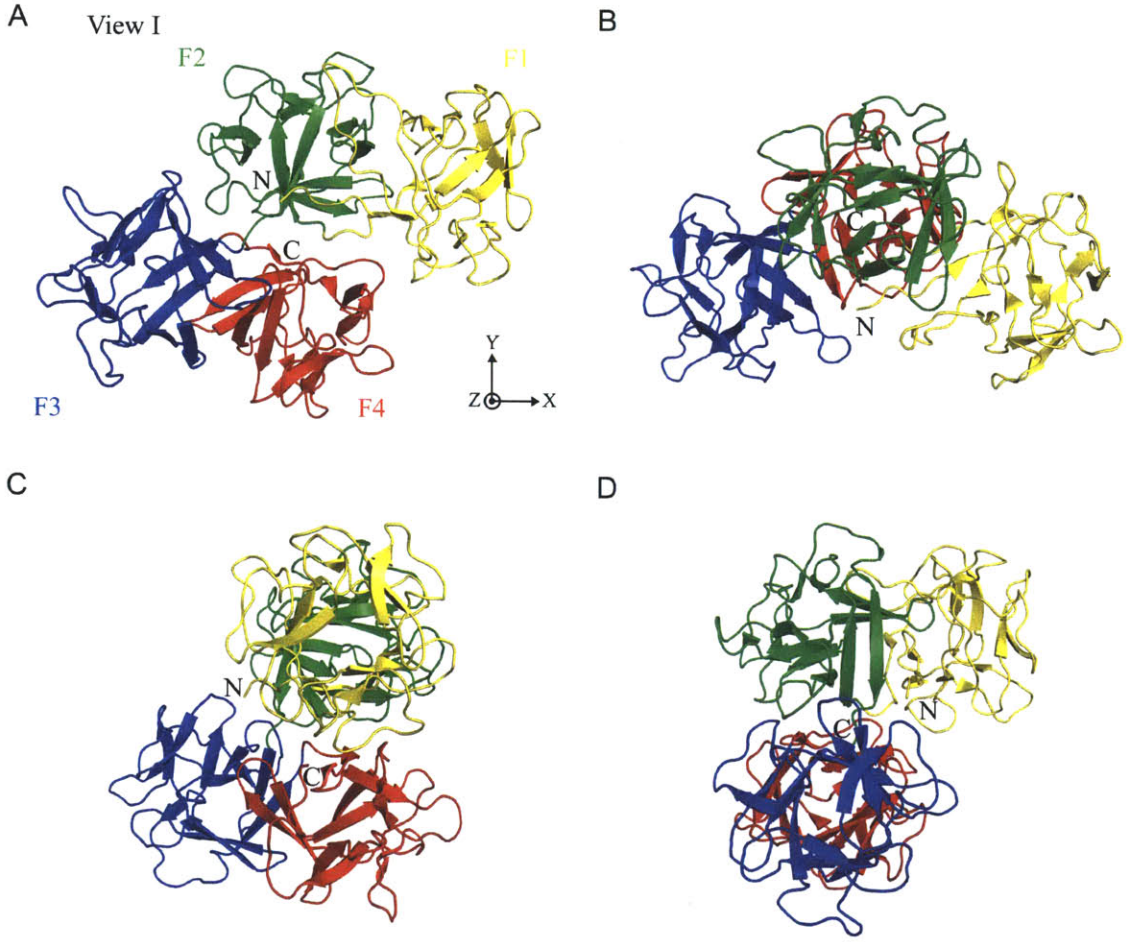

Figure 3-1 – Overall structure of *H. sapiens* fascin-1. (A) View I of *H. sapiens* fascin-1 showing lobe 1, which consists of β-trefoil domains F1 (residues 8−139, yellow) and F2 (residues 140−260, green) and lobe 2, which consists of β-trefoil domains F3 (residues 261−381, blue) and F4 (residues 382−493, red). The pseudo 2-fold axis is located between the lobes and is oriented approximately normal to the plane of the image. (B) View of fascin-1 from the top in comparison with view I in A, obtained by a 90° rotation of fascin about the X-axis in view I. The pseudo 2-fold axis is approximately vertical in the plane of the image. (C) View of fascin-1 along the pseudo 3-fold axis of lobe 1, obtained by rotation of fascin-1 in view I by 156° about the X-axis, −119° about the Y-axis, and −156° about the Z-axis. (D) View of fascin-1 along the pseudo 3-fold axis of lobe 2. This view is obtained by rotation of fascin-1 in view I by −150° about the X-axis, 125° about the Y-axis, and −151° about the Z-axis. All structural figures were generated with PyMOL [49].

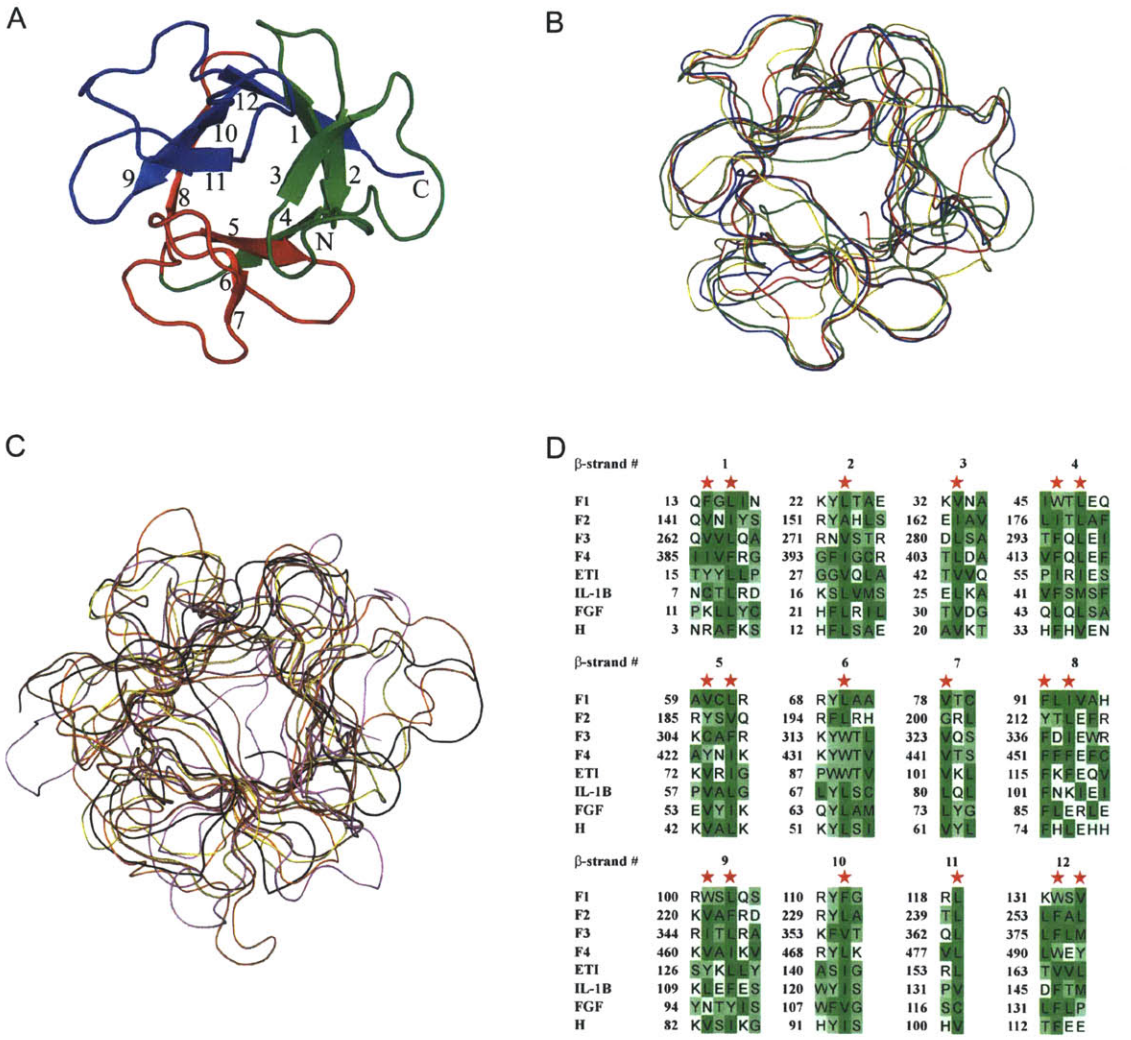

Figure 3-2 – Structure and sequence analyses of the β-trefoil fold. (A) View along the internal pseudo 3-fold axis of β-trefoil domain F2 of *H. sapiens* fascin-1, containing 12 β-strands. The hairpin triplet (β2, β3, β6, β7, β10 and β11) is located proximal and the six-stranded barrel (strands β1, β4, β5, β8, β9 and β12) distal. (B) Structural superposition of domains F1 (yellow), F2 (green), F3 (blue) and F4 (red) of fascin-1. (C) Structural superposition of homologous β-trefoil folds *Erythrina caffra* trypsin inhibitor (violet), *H. sapiens* interleukin-1β (orange), human acidic fibroblast growth factor (brown), *Dictyostelium discoideum* hisactophilin (black), and domain F1 (yellow) of fascin-1. (D) Structure-based sequence alignment of 59 homologous β-trefoil folds [148]: the four domains of *H. sapiens* fascin-1 (F1–F4), *E. caffra* trypsin inhibitor (ETI), *H. sapiens* interleukin-1β (IL-1B), human acidic fibroblast growth factor (FGF), and *D. discoideum* hisactophilin (H). Stars denote residues that stabilize the core of the β-trefoil fold. Green denotes hydrophobic and white denotes hydrophilic residues according to the hydrophobicity table described by Kyte et al. [149]. Secondary structure assignment is based on Ref. [147]. All multiple sequence alignment figures were generated using Jalview [150].

of antiparallel β-strands. The first and fourth β-strand of every unit coalesce to form a barrel structure (strands β1, β4, β5, β8, β9 and β12), with the strands inclined approximately 56° to the barrel axis. Three additional pairs of β-strands (β2, β3, β6, β7, β10 and β11) form three β-hairpins that cover the barrel from one side, with the exception of domains F1 and F4 based on the Dictionary of Protein Secondary Structure (DSSP) [151]. Structural superposition of the β-trefoil domains of fascin highlights their structural similarity (Fig. 3-2-B), with pair-wise RMSDs ranging from 1.3–2.2 Å (Table C.2).

The β-trefoil fold is present in over 40 distinct proteins deposited at present in the Protein Data Bank (PDB) [152] according to the structural classification of proteins (SCOP) database [153]. Members of the family include the mammalian cytokines interleukin-1α [154] and interleukin-1β [155], fibroblast growth factor [156], soybean trypsin inhibitor [157], the lectin B-chains from ricin [158] and abrin [159], the galactose-specific lectin amaranthin [160], and the actin-binding protein hisactophilin [161]. The β-trefoil domains of fascin exhibit a considerably higher degree of sequence identity and structural similarity to each other than to other members of the β-trefoil family, suggesting evolution of the fascin molecule via multiple gene duplication events from an ancestral β-trefoil fold (Fig. C-1-A and Tables C.3 and C.4) [156].

The core of the β-trefoil fold is stabilized by hydrophobic interactions between bulky hydrophobic residues contributed by each β-strand, extending nearly to the barrel axis to form a tightly packed hydrophobic core [148]. Residues that stabilize the core of the β-trefoil fold are identified from *Erythrina caffra* trypsin inhibitor (PDB ID 1TIE) [162], *H. sapiens* interleukin-1β (PDB ID 1I1B) [155], and *Bos taurus* acidic fibroblast growth factor (PDB ID 1JQZ)[1] [163] [148]. Structure-based multiple sequence alignment of 59 β-trefoil domains obtained from 45 distinct proteins available in the PDB demonstrates high variability at all amino acid positions other than these stabilizing hydrophobic core residues, which are highly conserved

[1] *H. sapiens* acidic fibroblast growth factor is used instead of *B. taurus* acidic fibroblast growth factor because structural coordinates for the latter are not currently available in the PDB.

across the β-trefoil domains analyzed (Figs. 3-2-D and C-2-A). Approximately 90% of the sequence pairs of these β-trefoil domains are $< 28\%$ identical, indicating their phylogenetic diversity (Fig. C-3-A). Residues responsible for hydrophobic packing in the β-hairpin triplet are mostly Leu, Val and Ile in all 59 β-trefoil domains, where each β-hairpin donates two residues to stabilize the triplet.

The stabilizing hydrophobic core residues present in the β-trefoil fold are also highly conserved across homologous fascin molecules, although they are somewhat more variable than among different β-trefoil proteins (30% versus 16% mean conservation grades) (Figs. 3-3, 3-4-A, and C-2-A). The conservation grade of the core-stabilizing hydrophobic residues in fascin and its homologues is reduced to 19% when residues are classified by their physicochemical properties using a seven-letter alphabet for the entropy-based conservation measure, suggesting that their hydrophobic nature is functionally more important than their specific amino acid type [164].

It is of interest to compare the single β-trefoil domain protein hisactophilin from the motile slime mold *D. discoideum* with fascin because it is also an actin-binding protein [161, 167]. Hisactophilin functions in osmoprotection of *D. discoideum*, enhancing the structural integrity of the cell cortex by crosslinking F-actin to the plasma membrane in a pH-dependent manner via protonation of its numerous histidine residues [168], which compose 31 out of its 118 amino acids. Interestingly, while eight analogous positions are occupied by the charged residues His, Arg and Lys in the first β-trefoil domain of fascin, seven in the second, four in the third, and three in the fourth, only three of these are His residues: the first domain has two His residues, the fourth domain has one, and the second and third domains do not have any. Additionally, the 109 residues in fascin-1 that are at analogous positions to the histidines in hisactophilin appear to be distributed randomly throughout the molecule, including 29 interfacial positions between β-trefoil domains (Fig. C-1-B). Thus, despite the structural and functional similarities of fascin to hisactophilin, regulation of its actin-binding via pH-dependent protonation of its histidine residues is unlikely to be present to the same extent as that in hisactophilin, if at all.

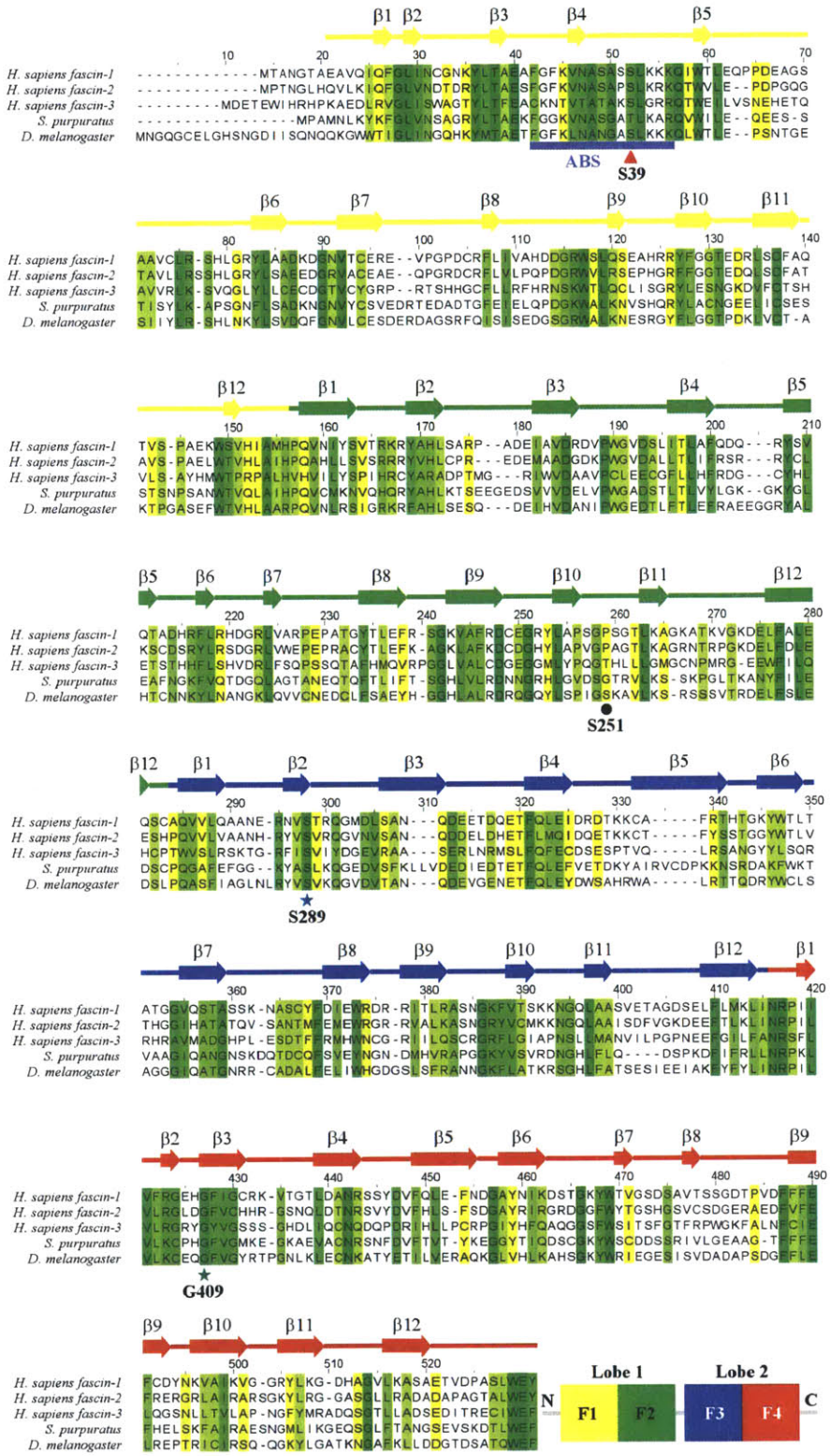

Figure 3-3 – Multiple sequence alignment of homologous fascins. Multiple sequence alignment of a subset of 61 homologous fascin sequences with NCBI accession codes and percentage sequence identity with *H. sapiens* fascin-1 provided in parentheses: *H. sapiens* fascin-1 (***NP_003079***; 100%), *H. sapiens* fascin-2 (***NP_036550***; 57%), *H. sapiens* fascin-3 (***NP_065102***; 28%), *Strongylocentrotus purpuratus* fascin (***NP_999701***; 37%) and *D. melanogaster* singed (isoform A) (***NP_727226***; 41%). Green denotes strictly conserved residues, yellow denotes strong conservation, and colors in between interpolate conservation grade linearly. A schematic of the fascin sequence is shown at bottom right. Point mutations in *D. melanogaster* fascin (the singed gene) that disrupt actin-bundling are S289N (S274 in *H. sapiens* fascin-1) (blue five-point star) and G409E (G393 in *H. sapiens* fascin-1) (green five-point star). S251F (P236 in *H. sapiens* fascin-1) (solid black circle) restores fascin function in the S289N mutant [165]. Mutation of S39 (also S39 in *H. sapiens* fascin-1) (red triangle) to alanine renders *in vitro* actin-binding by fascin-1 insensitive to regulation by phosphorylation by PKCα [132, 166]. See the text for details.

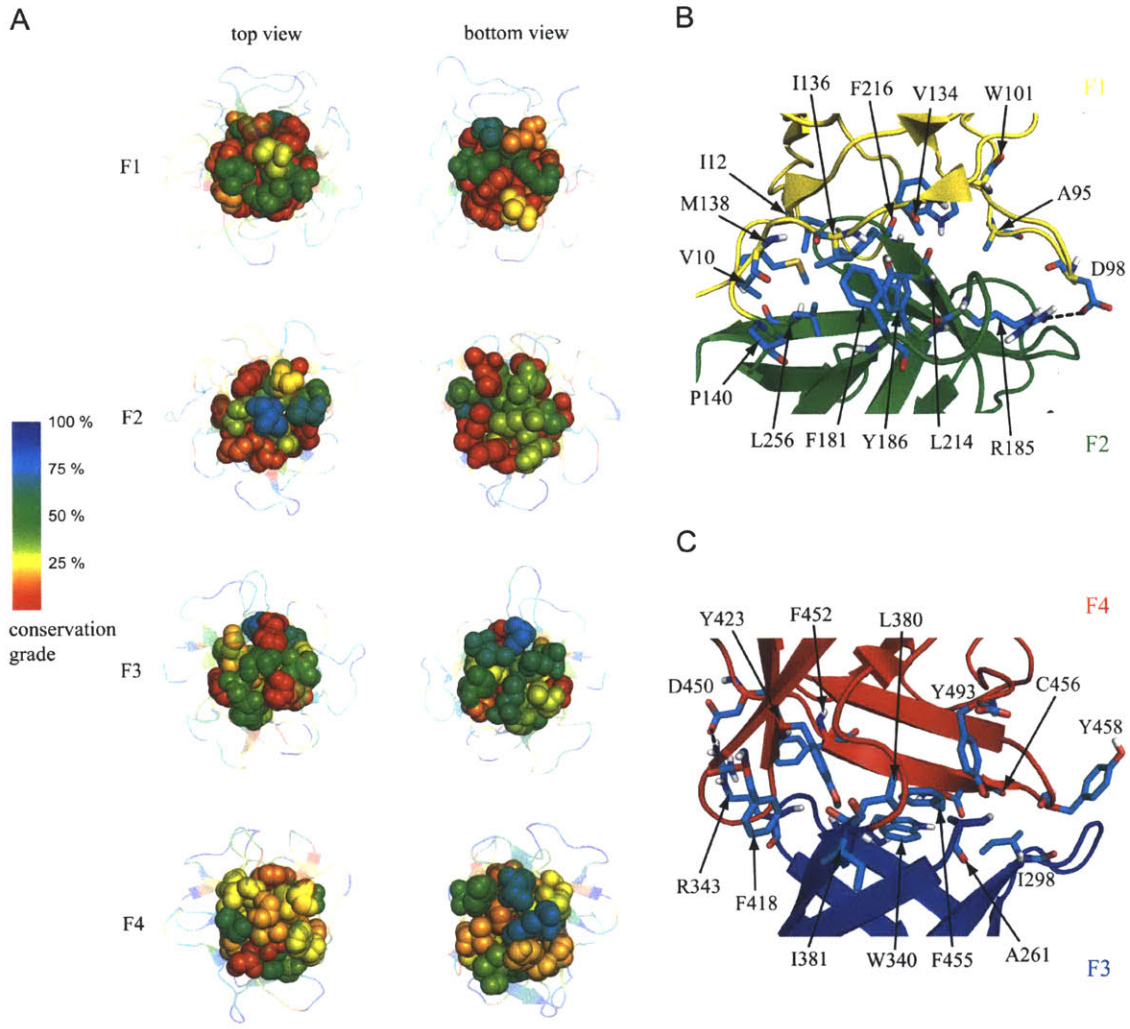

Figure 3-4 – Residues suggested to stabilize the β-trefoil cores and lobes of fascin-1. (A) Conservation grade and schematic of hydrophobic core residues (spheres) measured across homologous fascin molecules. Views are along the 3-fold axis of the β-trefoil domain, where "top" and "bottom" refer to viewing the domain from its cap and barrel, respectively. Conservation grades vary from blue to red, denoting highly variable and conserved residue positions, respectively. (B) Ionic and hydrophobic interactions between β-trefoil domains of lobe 1. (C) Ionic and hydrophobic interactions between the β-trefoil domains of lobe 2. Hydrophobic residues that lose more than 40 $Å^2$ of their solvent-accessible surface upon domain-domain association and residues participating in domain-domain salt bridges are labeled in B and C.

3.1.3 β-Trefoils associate to form two lobes in fascin

The β-trefoil domains of fascin associate at their exposed bases to form the first (F1 and F2) and second (F3 and F4) lobes of fascin with flanking polar interactions on the solvent-exposed surface of each lobe (Figs. 3-4-B, 3-4-C, and C-2-B). There is a single ion pair, D98-R185, between domains F1 and F2 in the first lobe (Fig. 3-4-B), and another, R343-D450, between domains F3 and F4 in the second lobe (Fig. 3-4-C)[2]. Two proline residues, P140 and P384, are highly conserved in the fascin family with respective conservation grades of 11% and 7%. The first induces a bend in the linker segment between domains 1 and 2 in the first lobe, and the second does the same in the linker between domains 3 and 4 in the second lobe. In total, 27% of all interfacial residues between domains F1 and F2 and between F3 and F4 reside in the first quartile of all residues in the molecule as ranked by conservation grade, suggesting that their inter-trefoil interactions might be important to the structural stability of each lobe. A detailed analysis of the relative contributions of electrostatic and hydrophobic interactions to lobe and fold stability would be of interest, but is beyond the scope of the present work.

3.1.4 Lobes associate to form the full-length fascin molecule

The full-length fascin molecule is formed by association of the lobes at a skew angle of approximately 56° (Fig. 3-1). The central region of fascin is stabilized by multiple polar interactions connecting all four β-trefoil domains together, where the residues H139, Q141, S259, R383 and R389 that contribute to these interactions are also highly conserved in the fascin family with respective conservation grades of 9, 8, 10, 0.5 and 17% (Figs. 3-3, 3-5, and 3-6). In total, 43% of all interfacial residues between the lobes of fascin reside in the first quartile of residues in the molecule as ranked by conservation grade, also suggesting their importance to the overall structural stability of the full-length molecule.

[2]These salt bridges are obtained for the oxygen-nitrogen distance cut-off of 3.2 Å. Increasing the cut-off to 4 Å results in one (D97-R224) and two (D342-K464 and R344-D420) more salt bridges between domains F1 and F2 and F3 and F4, respectively (see Section 3.3.2)

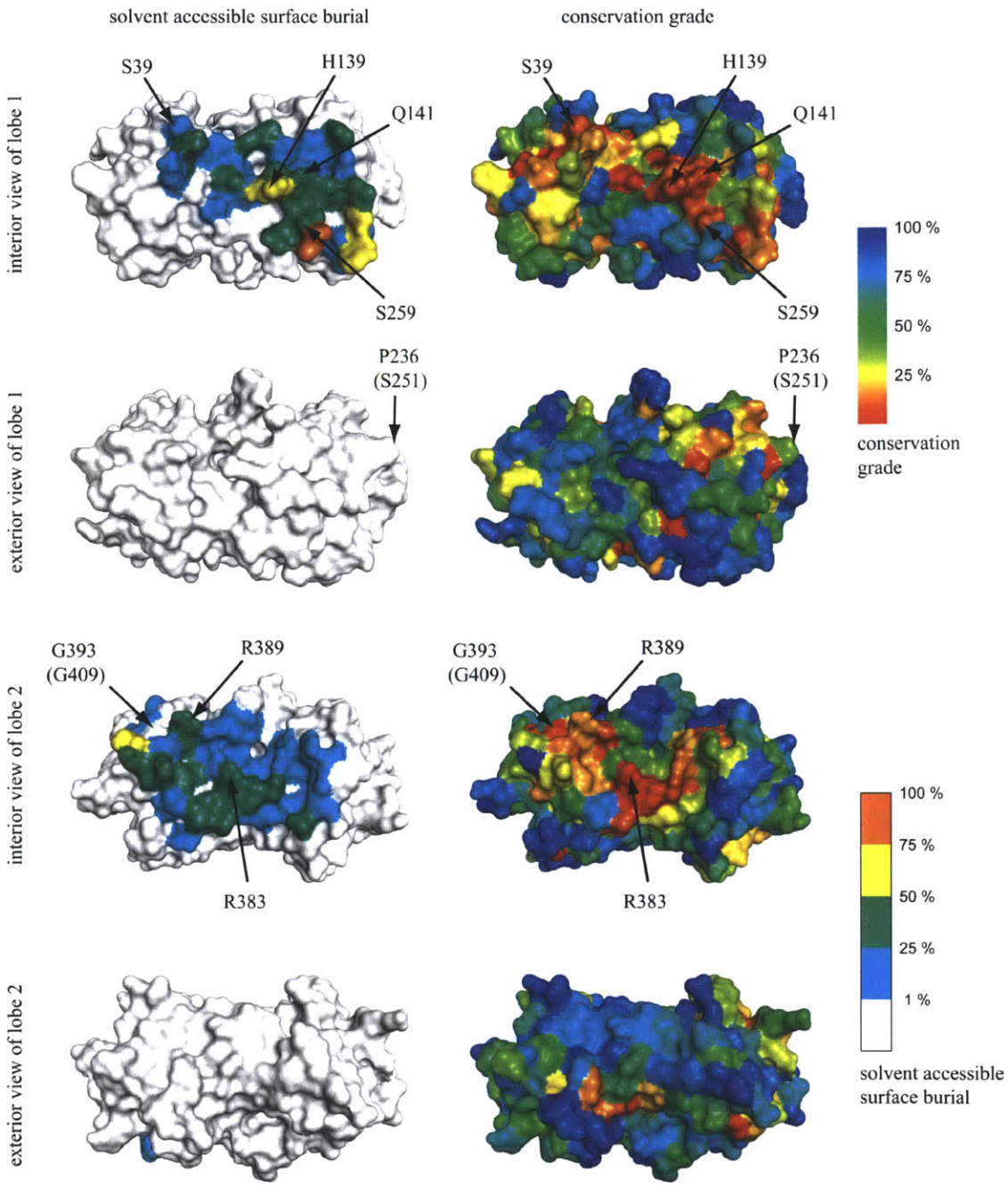

Figure 3-5 – Conservation grade and solvent-accessible surface burial of surface residues of the lobes of fascin-1. Interior and exterior views of the lobes of fascin-1 show residue conservation and solvent-accessible surface burial due to lobe-lobe association. White and orange denote low ($< 1\%$) and high ($> 75\%$) solvent-accessible surface burial, respectively. Conservation grades vary from blue to red, denoting highly variable and conserved residue positions, respectively. Interior and exterior lobe views are related by 180° rotations about the horizontal axis, where each lobe is aligned horizontally with the axis of its lowest principal moment of inertia. Point mutations affecting fascin function and residues H139, Q141, S259, R383 and R389 are indicated. Residue numbers in parentheses indicate the analogous positions in *D. melanogaster* fascin.

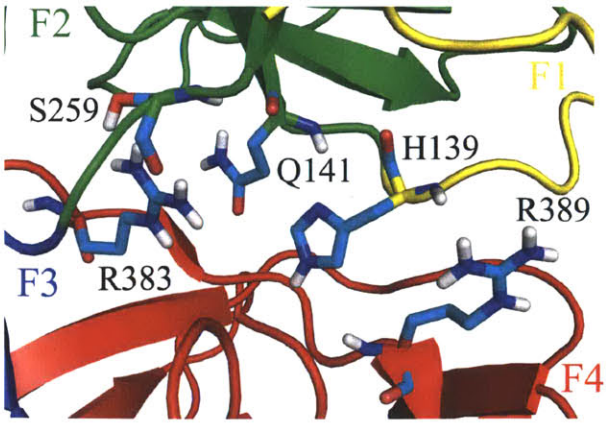

Figure 3-6 – Close-up view of highly conserved interfacial residues H139, Q141, S259, R383 and R389 in stick representation.

3.1.5 Putative actin-binding sites of fascin

In the absence of any structure of a fascin-F-actin complex, the precise actin-binding sites of fascin are not known. However, two point mutations in *Drosophila melanogaster* fascin (the singed gene) are known to impact bundling: G409E (equivalent to G393 in *H. sapiens* fascin-1) that leads to partial inactivation of fascin *in vivo* and S289N (equivalent to S274 in *H. sapiens* fascin-1) that inactivates fascin almost completely, suggesting that one actin-binding domain may be in the region near S274 that is also highly conserved (14% mean conservation grade) and solvent exposed (Figs. 3-3 and 3-7) [165]. Interestingly, mutation of S251 to phenylalanine (equivalent to P236 in *H. sapiens* fascin-1) partially restores the fascin function lost in the S289N mutation for reasons that are not understood [165].

A second actin-binding site is thought to be in the highly conserved protein kinase Cα (PKCα) substrate domain consisting of residues 29–43 in the first β-trefoil domain of fascin (Fig. 3-3). It has high sequence similarity to an actin-binding site of myristoylated alanine-rich C-kinase substrate (MARCKS) FGFKVNASASSLKKK (residues 29–43 in *H. sapiens* fascin-1) [108]. Mutation of S39 to alanine renders *in vitro* actin-binding by fascin-1 insensitive to regulation by phosphorylation by PKCα [132, 166], and is therefore a constitutively active mutant, as shown also in mouse B16F1 cells [104]. There, the constitutively active mutant fascin leads to a signif-

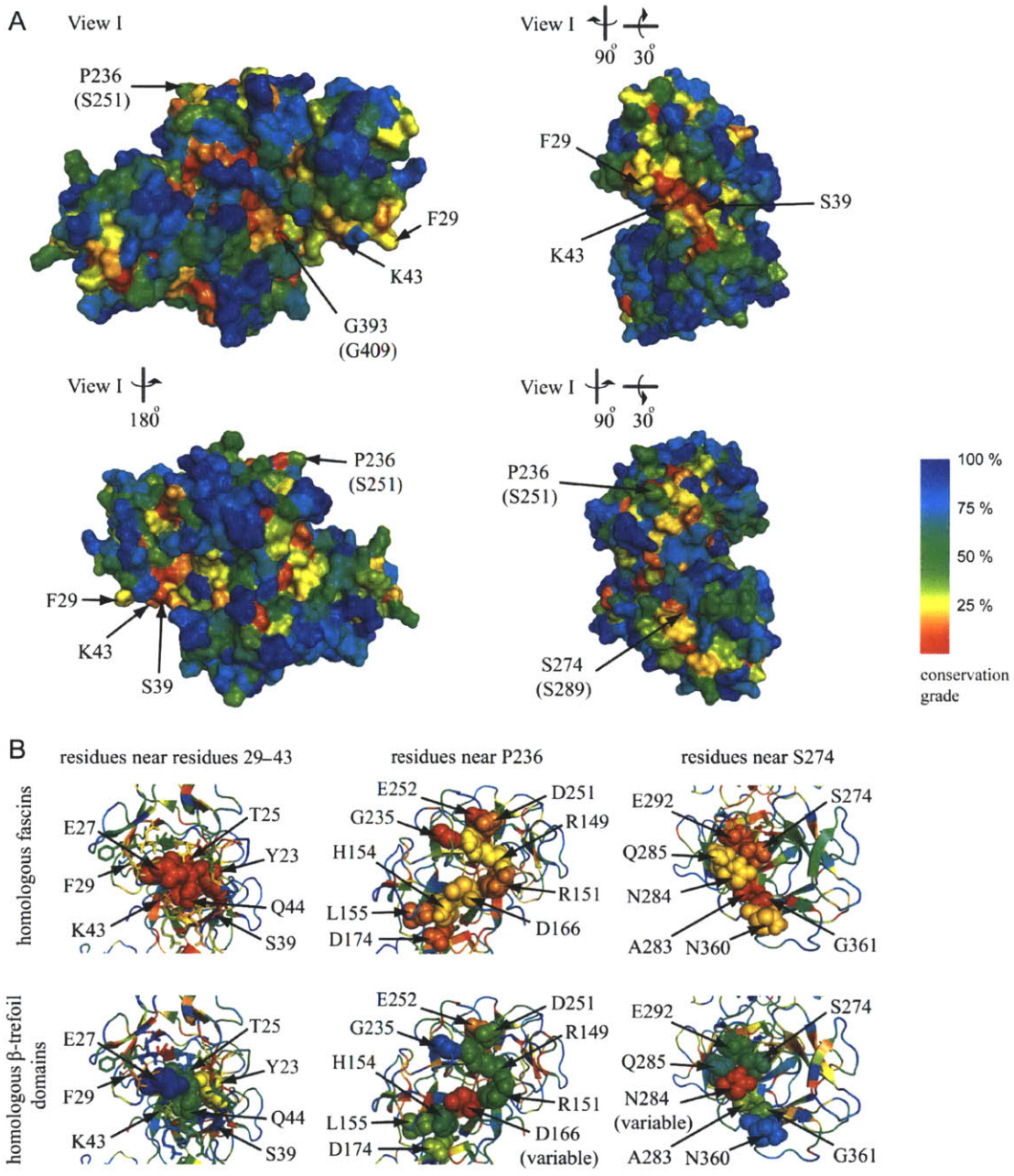

Figure 3-7 – Residue conservation near putative actin-binding sites of fascin-1. (A) Solvent-accessible surface residue conservation grades in fascin-1. Conservation grades vary from blue to red, denoting highly variable and conserved residue positions, respectively. Point mutations affecting fascin function and residues 29 and 43 are indicated. Residue numbers in parentheses indicate the analogous positions in *D. melanogaster* fascin. (B) Residues near residues 29–43 (putative actin-binding site), P236 and S274 are highly conserved across homologous fascins but not across homologous β-trefoils. Residues 23, 25, 27 and 44 near residues 29–43, residues 149, 151, 154–155, 166, 174, 235 and 251–252 near P236, and residues 274, 283–285, 292 and 360–361 near S274 are represented as spheres. Remaining residues are shown in cartoon representation. D166 and N284 are considered variable because the corresponding columns in the structure-based multiple sequence alignment of β-trefoil domains consist mostly of gaps (90%).

icant increase in the number and length of cortical filopodia, whereas the inactive phosphomimetic fascin mutant S39E is shown to reduce filopodial frequency [104].

The majority of the 123 highly conserved residues of fascin have functional importance that is rationalized on the basis of either contribution to fold/structural stability or to actin binding, accounting for 72 residues. Examples include the putative actin-binding site between residues 29 and 43, hydrophobic core stabilizing residues in the β-trefoil fold, interfacial residues between the lobes and, to a lesser extent, interfacial residues between β-trefoil domains within each lobe. This is supported by the distribution of conservation grades for each of these sub-sets of residues, which is biased towards high conservation when compared with all residues in the molecule (Figs. C-4 and C-5-A).

The remaining 51 highly conserved residues consist either of residues that are highly conserved both across fascin molecules and the β-trefoil fold, suggesting their functional importance to the fold but not necessarily to the specific biological function of fascin; glycines, which are likely to confer local flexibility to the molecule; and residues that are located proximal to the functional sites consisting of residues 29–43, S274, and P236, discussed above (Figs. 3-7-B and C-5 and Tables C.5 and C.6). The fact that these latter residues are highly conserved across homologous fascins but not generally across the β-trefoil fold suggests that they may have functional importance to the molecule, such as in binding to F-actin or stabilizing the fascin fold.

3.1.6 Conformational dynamics

Analysis of the conformational dynamics of fascin illustrates that end-association of the β-trefoil domains confers structural integrity within each cylindrical lobe that is not present across lobes (Fig. C-6). The generalized linear correlation coefficient (r_{LMI}) calculated for residue pairs using NMA may be used to infer the dynamical correlation between different regions of the protein (Fig. 3-8). Interestingly, correlations in the motions of domains F1 and F3 are the highest among all pairs of β-trefoil domains despite the fact that they are not in direct contact (Figs. 3-8-B and C-7 and Table 3.1). This suggests a potential allosteric coupling between these domains, which

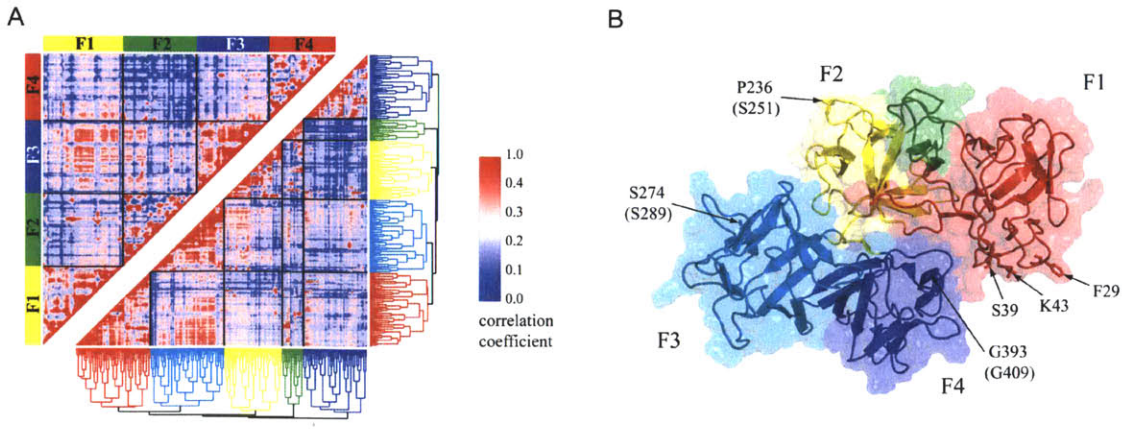

Figure 3-8 – Dynamically correlated domains of fascin-1. (A) Average-link hierarchical clustering is used to identify residue clusters that are highly correlated on the basis of the magnitude of their generalized linear mutual information coefficient. Upper and lower triangles show inter-residue correlations based on sequence position and after clustering, respectively. High dynamical correlations are found within individual β-trefoil domains, as expected, but also between distant domains F1 and F3 (see Table 3.1). (B) View I of fascin-1 colored according to the five clusters shown in A. Point mutations affecting fascin function and residues 29 and 43 are indicated. Residue numbers in parentheses indicate analogous positions in *D. melanogaster* fascin.

contain the putative actin-binding sites of the molecule. The structural origin of this long-range coupling is in the tight binding between the β-trefoil domains of fascin to form two β-barrel lobes that interact at an interfacial region that forms a hinge about which the lobes may bend (Fig. C-6). This lobe-structure is in stark contrast to the similar-sized crosslinking protein fimbrin, which consists of four calponin homology domains that each consists of an α-helix bundle [122]. Further analysis of fimbrin is required to understand potential functional implications of the distinct architectures of these molecules. Identification of the precise actin-binding sites of fascin via mutagenesis and structural studies are needed to elucidate implications of the observed structural and dynamical properties of the molecule on its actin-crosslinking function.

3.2 Discussion

The crystal structure of *H. sapiens* fascin-1 reveals that its β-trefoil domains associate via internal hydrophobic interactions and external ionic interactions at their bases to

Table 3.1 – Average generalized linear mutual information coefficient and fraction of residues that are in contact (% in parentheses) for the five clusters in fascin-1 shown in Fig. 3-8.

	Yellow	Red	Cyan	Green	Blue
Yellow	–	0.26 (11)	0.29 (11)	0.26 (31)	0.21 (16)
Red	–	–	0.31 (0)	0.26 (23)	0.27 (9)
Cyan	–	–	–	0.25 (0)	0.27 (13)
Green	–	–	–	–	0.17 (0)
Blue	–	–	–	–	–

Clusters (1) red-cyan, (2) cyan-green and (3) green-blue have mean generalized linear correlation coefficients in the 73^{rd}, 46^{th}, and 13^{th} percentiles, respectively, excluding intra-domain contributions and contributions from residues in direct contact (i.e., residues that have at least two heavy atoms within 5 Å of one another [169]).

form cylindrical β-barrel lobes. NMA indicates that these interactions confer structural integrity to each cylindrical lobe that is maintained in the full-length molecule. Conservation analysis confirms the functional importance of the cores of the β-trefoil domains, which contain bulky hydrophobic residues that are highly conserved across fascins and the entire β-trefoil family of proteins, which are otherwise divergent in sequence and biological function. Indeed, the majority of highly conserved residues in fascin are suggested either to contribute to its structural stability or to binding to F-actin. These residues are located in and near the putative actin-binding site between residues 29 and 43 and the previously identified functional residues S274 and P236, hydrophobic core-stabilizing residues of the β-trefoil fold, interfacial residues between the lobes and, to a lesser extent, interfacial residues between β-trefoil domains within each lobe (Figs. C-4 and C-5 and Table C.6).

Association of the cylindrical lobes of fascin at their largely polar interface results in a hinge that facilitates large-scale collective motions of each lobe, whereby the opposing putative actin-binding domains of fascin are correlated in their dynamical

motions. This distinct molecular design of fascin might have important implications on its biological function via a lever-like mechanism: action at one end of the molecule is transmitted in a lever-like fashion to the other end. Indeed, fascin binding to F-actin at one end (e.g. between F1 and F4) might induce a direct conformational change at the opposing end of the molecule (between F2 and F3) that could alter its binding affinity for F-actin, resulting in cooperative binding that could explain its unique ability to form tightly packed and ordered actin bundles *in vitro* and *in vivo*. Future mutational studies guided, in part, by the results of this work, together with cryo-EM-based reconstructions of the actin-bound fascin crosslink, will eventually facilitate a detailed understanding of the molecular basis for the unique structural and mechanical properties that these ubiquitous and highly conserved actin-binding proteins confer to actin bundles.

3.3 Computational procedures

3.3.1 Sequence analysis

We used three independent procedures for computing evolutionary sequence conservation: the conservation surface mapping method (ConSurf) [170, 171], the real-valued evolutionary trace method (ET) [172], and a simple entropy-based method using a 21 letter alphabet [164]. The correlation coefficient between each of these three conservation analysis procedures is greater than 80%. In the absence of any information regarding the superiority of any one approach, the grades were weighted such that each contributed equally to the final conservation grade, except for positions that were missing residues in the original PDB file, for which only the entropy 21-based grade was used. The "conservation grade" presented in Results is defined as percentage relative to all residues in the protein. For example, a residue conservation grade of 5% means that 5% of residues in the protein are at least as conserved as this residue, and 95% are less conserved. Highly conserved residues are defined as residues residing in the first quartile of residues in the molecule as ranked by conservation grade.

ConSurf was applied on the set of homologous *H. sapiens* fascin-1 sequences stored in the ConSurf data bank (ConSurf-DB) [171]. ConSurf-DB stores results of ConSurf analysis of all structures in the PDB [152]. Conservation grades were computed using the Rate4Site algorithm [173] on 22 homologous sequences selected by the ConSurf-DB protocol for *H. sapiens* fascin-1 [171].

The real-valued ET procedure is a hybrid method that incorporates an entropy grade into a phylogenetic analysis of a multiple sequence alignment [172]. The phylogenetic tree was created using hierarchical clustering based on the unweighted pair group method with arithmetic mean [174]. The tree was divided into distinct subfamilies using different partitionings, where an entropy grade was calculated for each position in a subfamily-specific manner [172]. Conservation analysis was performed on *H. sapiens* fascin-1 (SwissProt ID Q16658) using 31 sequences extracted from the HSSP database [175].

In the third conservation analysis procedure, a simple entropy measure with a 21 letter alphabet was used. Sequences with a high degree of homology to *H. sapiens* fascin-1 were retrieved from the NCBI [4] using PSI-BLAST [176] with an E-value cutoff of 10^{-6} and one iteration. Sequences that differed by more than 5% in length from the query sequence were removed together with redundant sequences, resulting in 61 sequences (Table C.7). Sequences were aligned using ClustalX with default parameters [177]. The entropy of each position in the sequence alignment, S_i, was computed using the standard measure for entropy:

$$S_i = -\sum_{k=1}^{21} p_{ik} \log_{21} p_{ik} \tag{3.1}$$

where p_{ik} is the probability of occurrence of amino acid type k at sequence position i. The structure-based sequence alignment of β-trefoil domains was done with STAMP [178].

3.3.2 Physical property analysis

Interfacial residues between fascin domains are identified by the change in their relative solvent accessibility upon domain-domain association [179], where relative solvent accessibility is defined as the ratio of the actual solvent accessibility of the residue to its solvent accessibility in the extended Gly-x-Gly tripeptide [180]. Interfacial residues are defined as residues that change their relative solvent accessibility by more than 1% between the dissociated and associated states [179]. Salt bridges were found using the VMD 1.8.6 salt bridge plug-in [181] with default parameters. Increasing the oxygen-nitrogen distance cut-off of 3.2 Å (VMD default value) to 4 Å was also considered [182].

Conformational dynamics of fascin were calculated using NMA [183, 184]. The crystal structures of chains A and B of fascin (PDB ID 1DFC) were analyzed with missing loops built using Swiss-PDB viewer [185]. All-atom energy minimization and subsequent NMA were done with CHARMM version 33a1 [46] using the implicit solvent force-field EEF1 [47]. Hydrogen atoms and amino acid side chains were first minimized sequentially with all remaining protein atoms fixed. Repeated rounds of steepest descent minimization followed by adopted basis Newton-Raphson minimization were subsequently done in the presence of successively reduced harmonic constraints applied to backbone chain atoms. The final energy-minimized structures of chains A and B had RMS energy gradients of 4×10^{-4} and 2.5×10^{-4} $\frac{\text{kcal}}{\text{mol} \times \text{Å}}$, respectively, with RMS coordinate differences between the crystal and energy-minimized structures of 1.9 Å and 1.7 Å. The Block Normal Mode method [57, 58] was used with one residue per block to compute the 207 lowest frequency normal modes. The mean-square thermal fluctuation of α-carbon atom i is computed as:

$$\langle \Delta {r_i}^2 \rangle = k_B T \sum_k \frac{{y_{ik}}^2}{\lambda_k m_i} \tag{3.2}$$

where m_i is atomic mass, $k_B T$ is thermal energy, λ_k is the eigenvalue corresponding to normal mode k, and ${y_{ik}}^2$ is the squared magnitude of the projections of the k^{th} normal mode on the Cartesian components of the displacement vector of the i^{th} atom

[21].

Correlations in atomic fluctuations were computed using the generalized linear correlation coefficient:

$$r_{\mathrm{LMI}}\left[\mathbf{x}_i, \mathbf{x}_j\right] = \left(1 - \exp\left(-\frac{2I_{\mathrm{lin}}\left[\mathbf{x}_i, \mathbf{x}_j\right]}{3}\right)\right)^{1/2} \tag{3.3}$$

based on the linear mutual information metric:

$$I_{\mathrm{lin}}\left[\mathbf{x}_i, \mathbf{x}_j\right] = \frac{1}{2}\left[\ln \det\mathbf{C}_{(i)} + \ln \det\mathbf{C}_{(j)} - \ln \det\mathbf{C}_{(ij)}\right] \tag{3.4}$$

where $\mathbf{C}_{(i)} = \langle \mathbf{x}_i{}^{\mathrm{T}}\mathbf{x}_i \rangle$ and $\mathbf{C}_{(ij)} = \langle (\mathbf{x}_i, \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i, \mathbf{x}_j) \rangle$ are marginal covariances of atom i and the pair-covariance matrix of atoms i and j, respectively (see Appendix C) [186]. The results were compared with the more commonly used Pearson correlation coefficient:

$$C_{ij} = \frac{\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle}{\left(\langle x_i{}^2 \rangle \langle x_j{}^2 \rangle\right)^{1/2}} \tag{3.5}$$

where $\mathbf{x}_i$ denotes displacement of atom i in Fig. C-8 [187].

Conclusions

This thesis contributes mainly to protein normal mode analysis (NMA) by both developing a computationally efficient and robust eigenvalue solver and incorporating the effects of solvent-damping into the analysis. Additionally, it examines comprehensively the structure, evolutionary sequence conservation, and conformational flexibility of *Homo sapiens* fascin-1 and gives insight into its functional mechanism.

In the present work, first, we presented an algorithm to optimize the number of iteration vectors used in the subspace iteration (SSI) method, a widely used eigenvalue solver in engineering problems, for protein NMA. We demonstrated that the algorithm improves substantially the effectiveness of the SSI method for proteins. With this algorithm, the computational effort scales linearly with the number of eigenpairs computed, as demonstrated for G-actin and pertussis toxin. The SSI method is well suited to the analysis of protein conformational change pathways, where numerous analyses may be performed in nearby conformations, which have relatively similar eigensolutions. In such cases, the eigensolution calculated for the previous conformation provides an excellent set of initial iteration vectors for the current solution, as demonstrated for the open-to-close conformational change of adenylate kinase.

Second, we developed an algorithm to accurately calculate the effects of solvent-damping on proteins as a solvent friction matrix. In this algorithm, the whole solvent friction matrix is obtained using only one finite element solid simulation with the commercial finite element software program ADINA, which renders the algorithm significantly efficient. The finite element method (FEM) proved successful in calculating the diffusion coefficients of a sphere and 10 proteins with various molecular weights. Subsequently, we coupled the friction matrix and the stiffness and mass

matrices of crambin calculated using the FEM to obtain the Langevin modes of the protein. As expected, all the relaxation times of the critically damped or over-damped Langevin modes of crambin are longer than the one corresponding to its rotational diffusion coefficient. It was also demonstrated that only the first few non-zero vacuum normal modes of crambin can be well-characterized by individual critically damped or over-damped Langevin modes.

Finally, we examined and described the structure of *H. sapiens* fascin-1 in detail. The structure is comprised of four tandem β-trefoil domains that associate via internal hydrophobic and external ionic interactions at their bases to form two β-barrel lobes. Sequence conservation analysis confirms that the bulky hydrophobic residues in the cores of β-trefoil domains are responsible for stabilizing the β-trefoil fold. Additionally, the interfacial residues between lobes and, to a lesser extent, the interfacial residues between β-trefoil domains within each lobe were suggested to play a central role in stabilizing the overall structure of fascin. An important observation is that conformational dynamics analysis suggests an allosteric mechanism between the putative actin-binding domains of fascin, which contain highly conserved surface patches. Hence, fascin binding to F-actin at one actin-binding domain might cause a conformational change at the other one that could alter its binding affinity for F-actin, resulting in cooperative binding that could explain its unique ability to form tightly packed and ordered actin bundles *in vitro* and *in vivo*. The results of this work can guide future experimental studies to provide a detailed understanding of the molecular basis for the unique structural and mechanical properties that these ubiquitous and highly conserved actin-binding proteins confer to actin bundles.

Appendix A

Calculation of the conformational change pathway of adenylate kinase

The open-to-closed transition of adenylate kinase is calculated using the elastic-based finite element approach of M. Bathe [18]. For each conformation along the pathway, starting with the open conformer (PDB ID 4AKE [50]), we calculate the finite element eigenvectors, $\boldsymbol{\varphi}_i{}^k$, where i and k denote the eigenvector and conformation numbers, respectively. Finite element eigenvectors for each conformation k are interpolated to the α-carbon positions of conformation k using the finite element interpolation functions h_l, where l denotes the element node number. The projection equations are $u_{ij}{}^k = \sum_{l=1}^{q} h_l u_{ijl}{}^k$, $v_{ij}{}^k = \sum_{l=1}^{q} h_l v_{ijl}{}^k$, and $w_{ij}{}^k = \sum_{l=1}^{q} h_l w_{ijl}{}^k$. Here $u_{ij}{}^k$, $v_{ij}{}^k$, and $w_{ij}{}^k$ are the x-, y-, and z-components, respectively, of the α-carbon eigenvectors $\mathbf{C}_i{}^k$ corresponding to α-carbon j. Also, $u_{ijl}{}^k$, $v_{ijl}{}^k$, and $w_{ijl}{}^k$ denote the x-, y-, and z-components, respectively, of $\boldsymbol{\varphi}_i{}^k$ corresponding to node l of the finite element enclosing α-carbon j [13]. The summations are performed over the q nodes of the finite element enclosing α-carbon j.

To generate the $(k+1)^{\text{th}}$ conformation along the trajectory, first, the α-carbon eigenvectors $\mathbf{C}_i{}^k$ are multiplied by the coefficients $d_i{}^k$ ($\mathbf{C}_i{}^{k\prime} = d_i{}^k \times \mathbf{C}_i{}^k$) such that the resultant eigenvectors are normalized ($\left\|\mathbf{C}_i{}^{k\prime}\right\| = 1$, where prime denotes the normalized set of eigenvectors.). Since $\mathbf{C}_i{}^k$ are obtained directly from the original $\boldsymbol{\varphi}_i{}^k$ via the FE interpolation functions, $\boldsymbol{\varphi}_i{}^k$ must also be scaled by $d_i{}^k$ as $\boldsymbol{\varphi}_i{}^{k\prime} = d_i{}^k \times \boldsymbol{\varphi}_i{}^k$.

In calculating the conformational change pathway, we use the eigenvectors denoted by primes. However, from here on, primes are dropped for simplicity of notation. Hence, $\mathbf{C}_i{}^k$ and $\boldsymbol{\varphi}_i{}^k$ will be the normalized α-carbon eigenvectors and their corresponding FE eigenvectors, respectively.

Next, the difference vector between α-carbon positions in the k^{th} conformation and those in the closed conformer (PDB ID 1AKE [4]), $\boldsymbol{\Delta}\mathbf{r}^k$, is projected onto the eigenvectors corresponding to the α-carbons in order to calculate the contribution of each α-carbon eigenvector, $c_i{}^k$, to the displacement of the α-carbons towards their new positions in the $(k+1)^{\text{th}}$ conformation. The projection equation is $c_i{}^k = \beta^k \boldsymbol{\Delta}\mathbf{r}^k \cdot \mathbf{C}_i{}^k$ [23, 59], where β^k is a parameter that is chosen to be between 0 and 1 in order to maintain an approximately constant step-size ($\Delta RMSD^k = \text{constant}$). $\Delta RMSD^k$ is defined as the difference between the k^{th} and $(k+1)^{\text{th}}$ root-mean-square differences (RMSD) ($\Delta RMSD^k = RMSD^k - RMSD^{k+1}$), where $RMSD^k$ refers to the RMSD between the α-carbon positions in the k^{th} conformation and those in the closed conformer.

Given $c_i{}^k$, new finite element nodal positions can be calculated as $\mathbf{R}^{k+1} = \mathbf{R}^k + \sum_{i=1}^{n} c_i{}^k \boldsymbol{\varphi}_i{}^k$, where $\mathbf{R}^k$ is a vector of nodal positions in the k^{th} conformation and n is the number of eigenvectors used in the conformational change pathway analysis. Atomic positions are subsequently computed using the FE interpolation functions.

The above procedure is used to calculate an initial conformational change pathway consisting of 1843 conformations (Fig. A-1), where a total of 2000 conformations were desired (the appropriate β^k that are needed to generate a conformational change pathway consisting of 2000 conformations are unknown *a priori* due to the nonlinear nature of the conformational change pathway). Subsets of 1001, 101, 11, and one conformation(s) are subsequently selected from this original set such that the step-size is approximately constant along the pathway (Fig. A-2). $\Delta RMSD^k$ varies considerably more for the 1001-conformation pathway than for the 101- and 11-conformation pathways because of the limited number (1843) of conformations available in the original conformational change pathway (Fig. A-2).

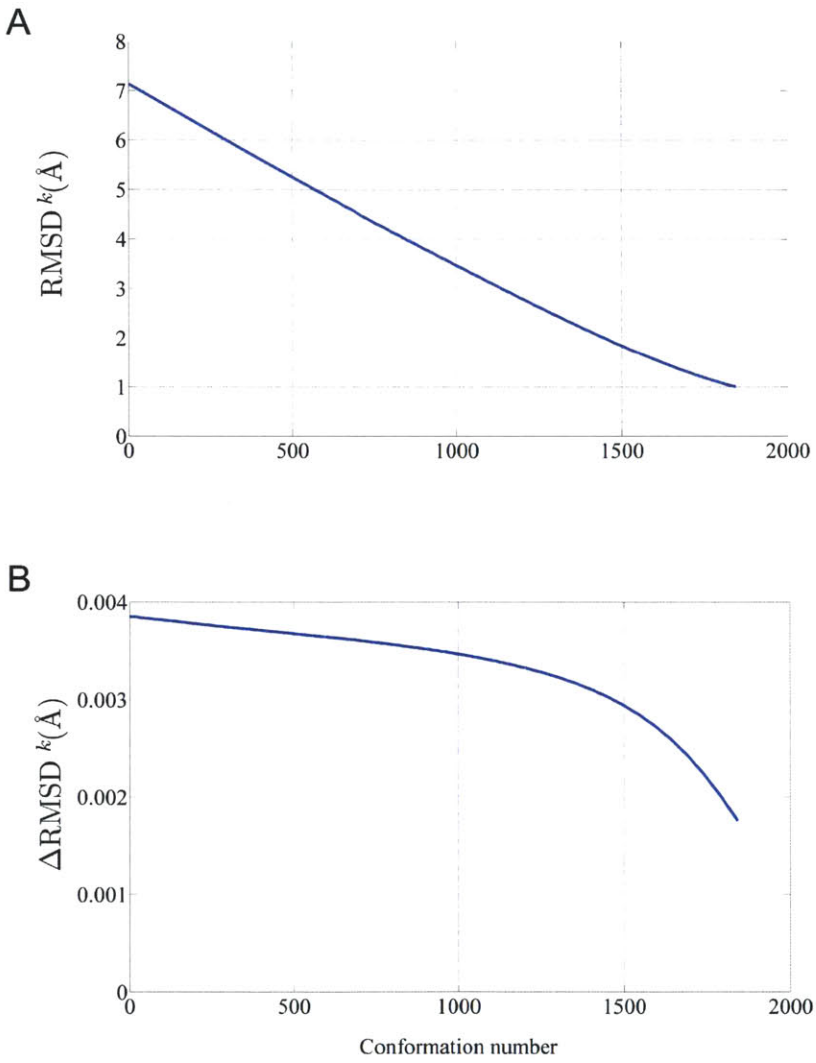

Figure A-1 – (A) $RMSD^k$ and (B) $\Delta RMSD^k$ versus conformation number for the 1843-conformation pathway.

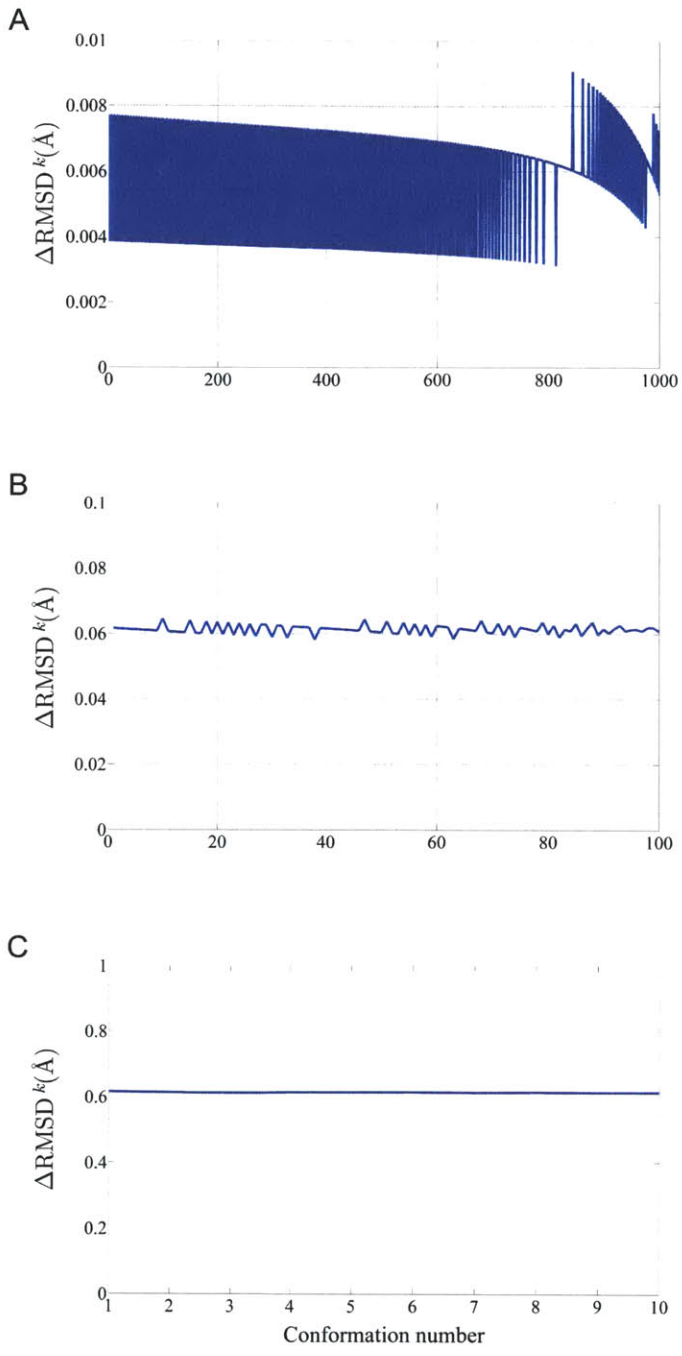

Figure A-2 – $\Delta RMSD^k$ versus conformation number for the (A) 1001-, (B) 101-, and (C) 11-conformation pathways.

Appendix B

Calculation of the effective material properties of adenylate kinase

The local mass density of adenylate kinase, $\rho = 1420\ \frac{\text{kg}}{\text{m}^3}$, is assumed to be homogeneous and equal to its molecular weight (23.6 kDa) divided by the molecular volume (27,567 Å^3) [52] of the energy-minimized structure of the open conformer (PDB ID 4AKE [50]). Energy minimization is performed in CHARMM version 33a1 [46] using the implicit solvation force-field EEF1 [47]. Steepest descent minimization followed by adopted-basis Newton-Raphson is performed in the presence of successively reduced harmonic constraints on all backbone atoms to achieve a final root-mean-square (RMS) energy gradient of $4 \times 10^{-4}\ \frac{\text{kcal}}{(\text{mol}\times\text{Å})}$ with corresponding RMSD between the X-ray and energy-minimized structures of 2.1 Å. Energy minimization is required by the Rotational Translational Block (RTB) procedure [57, 58], which is used as a reference to calculate the effective material properties of adenylate kinase.

An effective Young's modulus of adenylate kinase, E= 4.9 GPa, is chosen such that equilibrium thermal fluctuations of α-carbons in the finite element model are equal to those computed using the RTB procedure, and Poisson's ratio is assumed to be 0.3 [18]. Equilibrium thermal fluctuations of α-carbons are calculated using the FEM by first computing the eigenvectors corresponding to the finite element nodes, $\boldsymbol{\varphi}_i$, for the

energy-minimized structure of the open conformer using the commercially available finite element software program ADINA ver. 8.5 (Watertown, MA), where i denotes eigenvector number. A total of 46,902 four-node tetrahedral finite elements with linear interpolation functions are used to discretize the molecular volume of the protein. The eigenvectors corresponding to the α-carbons, $\mathbf{C}_i$, of the energy-minimized structure of the open conformer are calculated, as described in Appendix A, by projecting the FE eigenvectors onto the α-carbon positions: $u_{ij} = \sum_{l=1}^{q} h_l u_{ijl}$, $v_{ij} = \sum_{l=1}^{q} h_l v_{ijl}$, and $w_{ij} = \sum_{l=1}^{q} h_l w_{ijl}$, where u_{ij}, v_{ij}, and w_{ij} are the x-, y-, and z-components, respectively, of the α-carbon eigenvectors $\mathbf{C}_i$ corresponding to α-carbon j; and u_{ijl}, v_{ijl}, and w_{ijl} denote the x-, y-, and z-components, respectively, of $\boldsymbol{\varphi}_i$ corresponding to node l of the finite element enclosing α-carbon j [13]. Here the summations are performed over the q nodes of the finite element enclosing α-carbon j.

The normalized vector $\mathbf{y}_i = \frac{\mathbf{C}_i}{\|\mathbf{C}_i\|}$ is subsequently defined such that $\mathbf{y}_i \cdot \mathbf{y}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta. Equilibrium thermal fluctuations of α-carbon o is then given by $\langle \Delta {r_o}^2 \rangle = \sum_k \langle \Delta {r_{ok}}^2 \rangle$ [18], where $\langle \Delta {r_{ok}}^2 \rangle = \left(\frac{k_B T}{m_o \lambda_k} \right) {y_{ok}}^2$ is the mean-square fluctuation of α-carbon o due to mode k, $k_B T$ is thermal energy, and m_o is the mass of amino acid o (here amino acid o is represented by one pseudo-atom at the position of α-carbon o) [21]. Also, y_{ok} is the magnitude of a vector made of the three components of $\mathbf{y}_k$ corresponding to α-carbon o. The lowest 201 non-rigid-body normal modes are used to calculate thermal fluctuations of adenylate kinase using the RTB and FE-based approaches, where every α-carbon atom is considered (Fig. B-1).

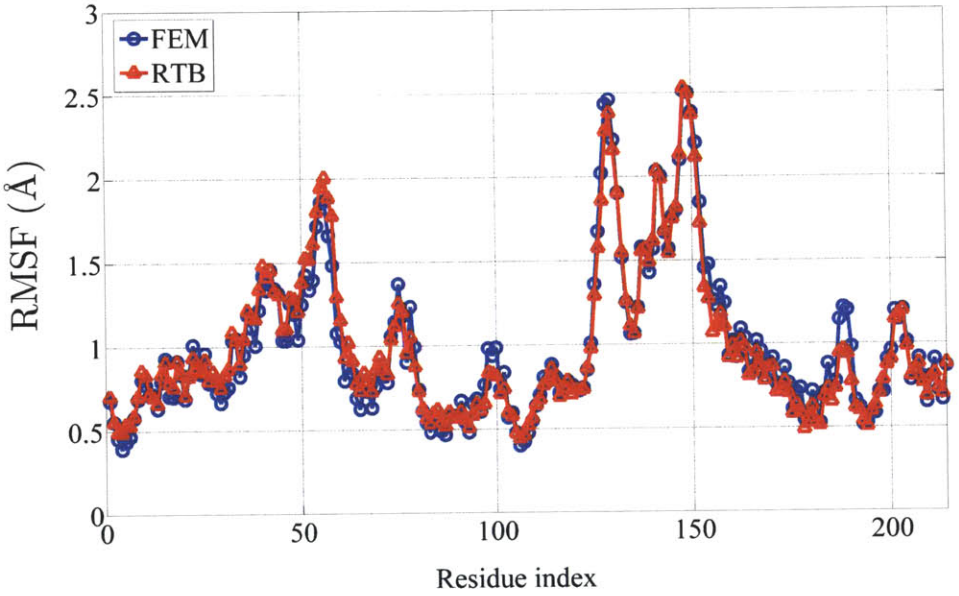

Figure B-1 – Root-mean-square fluctuations of α-carbons obtained using the FEM and the RTB procedure.

Appendix C

Supplementary materials for Chapter 3

C.1 Supplementary figures

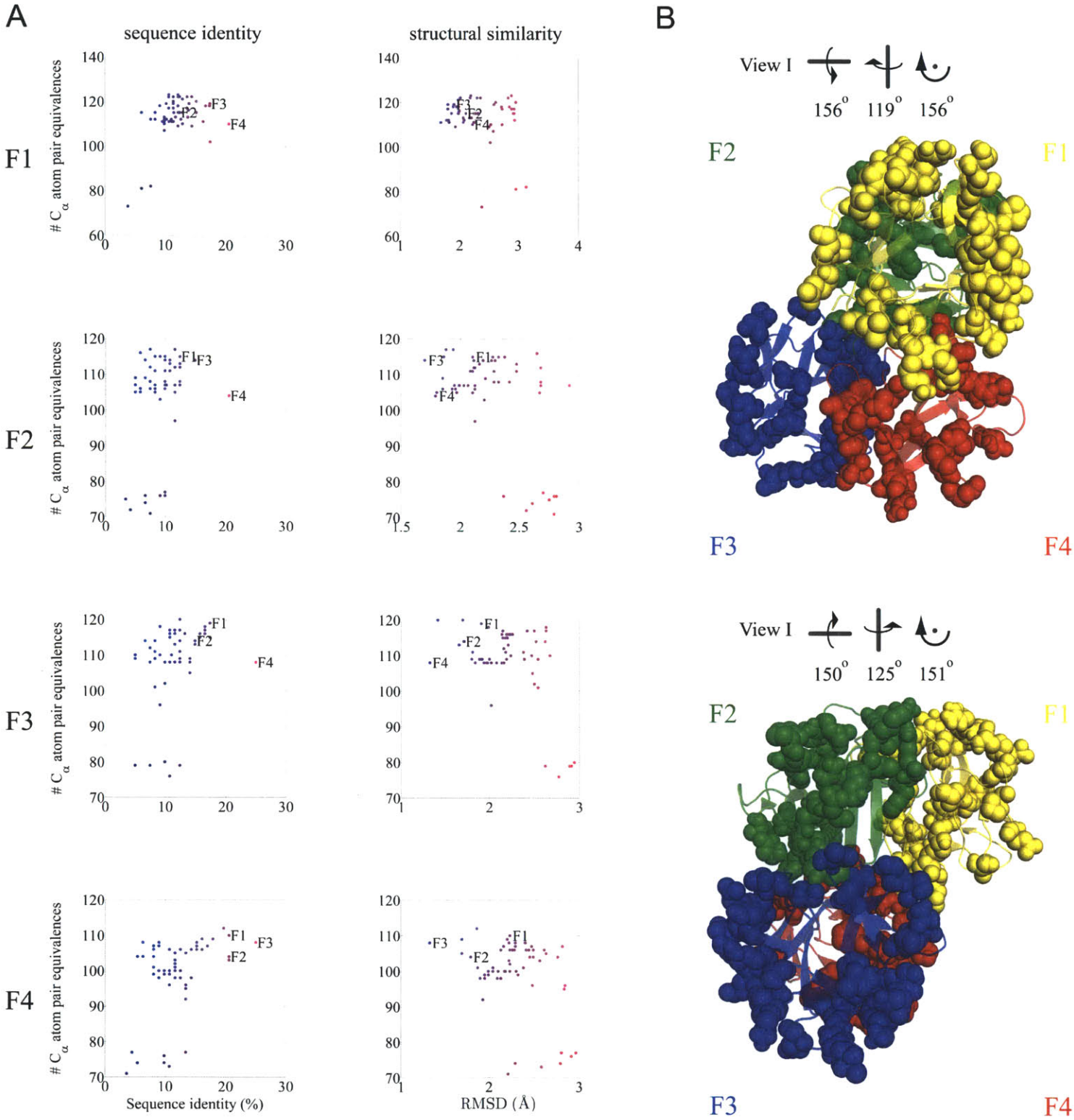

Figure C-1 – Analysis of structural alignments of fascin-1 domains with other β-trefoil fold domains. (A) Sequence identity and structural similarity of domains of fascin-1 to β-trefoil domains available in the Protein Data Bank (PDB) [152] (see also Section C.3). Each data point denotes one β-trefoil domain, with fascin-1 domains labeled. Domains of fascin-1 exhibit higher sequence identity and structural similarity to each other than they do to other β-trefoil domains. (B) General views of fascin-1 along the pseudo 3-fold axes of the first and second lobes. Fascin-1 residues analogous to histidines in hisactophilin [161] are represented as spheres. Remaining residues are shown in transparent cartoon representation. The 31 histidines in hisactophilin have 109 analogous positions in fascin. The 109 residues are distributed randomly throughout fascin-1 (i.e., interfacial domains, solvent-exposed domains, etc.). All structural figures were generated with PyMOL [49].

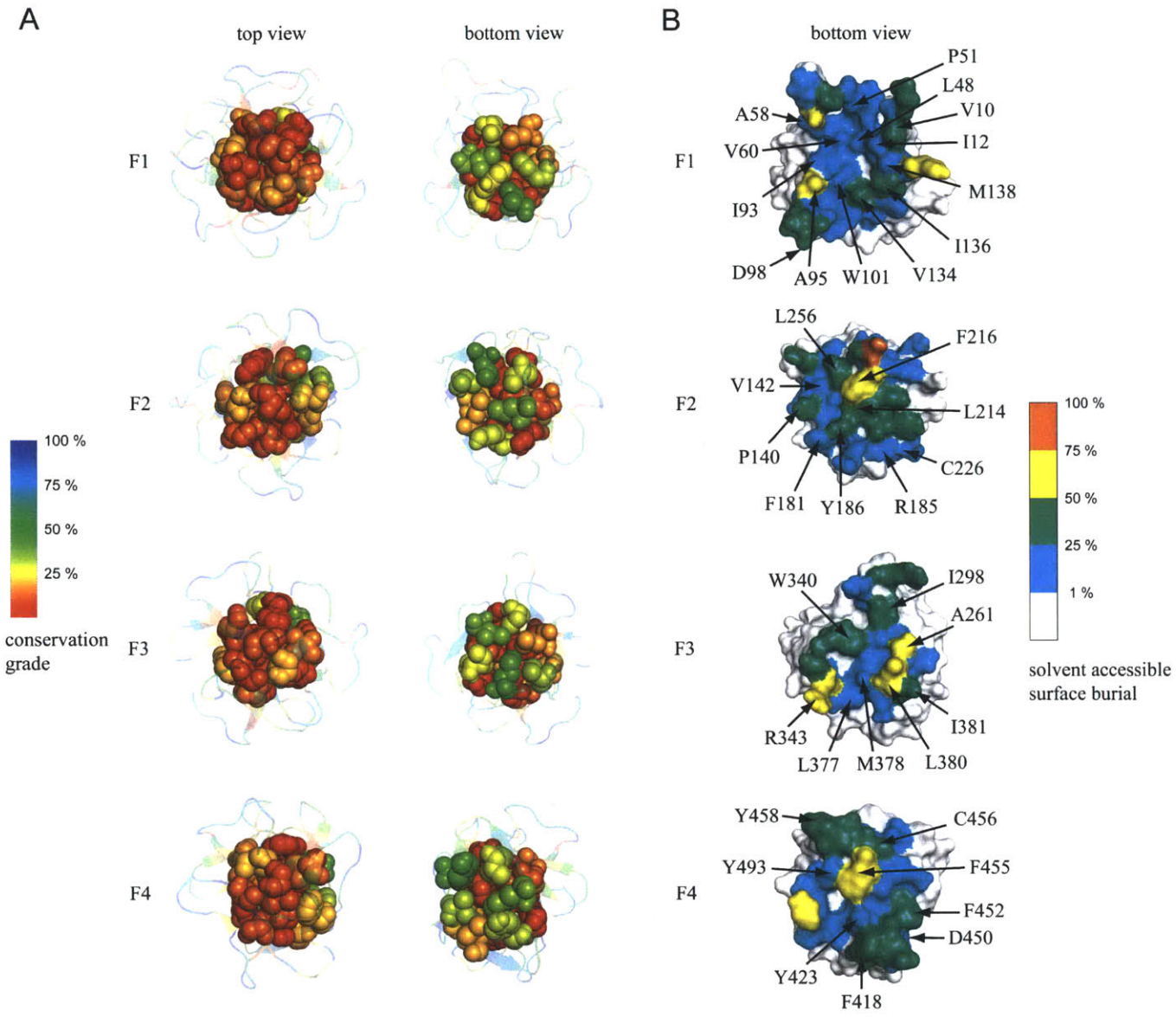

Figure C-2 – Conservation of residues suggested to stabilize the β-trefoil core and solvent accessible surface burial upon β-trefoil domain-domain association within each lobe of fascin-1. (A) Conservation of hydrophobic core residues measured across β-trefoil domains available in the PDB (see also Section C.3). Views are along the pseudo 3-fold axis of the β-trefoil domain, where "top" and "bottom" refer to viewing the domain from its cap and barrel, respectively. Conserved hydrophobic residues that stabilize the core of the domain are drawn as spheres. Remaining residues are shown in transparent cartoon representation. Red denotes highly conserved and blue denotes highly variable residue positions. Most hydrophobic residues that stabilize the β-trefoil domain cores in fascin-1 are highly conserved over the set of β-trefoil domains present in the PDB. This observation implies that these residues are important generally to the stability of the β-trefoil domain structure. (B) Solvent accessible surface burial upon β-trefoil domain-domain association within each lobe of fascin-1. White and orange denote low ($< 1\%$) and high ($> 75\%$) solvent accessible surface burial, respectively. Interfacial residues between the β-trefoil domains of fascin-1 within each lobe are defined as residues that have a solvent accessible surface burial that is greater than 1%. Hydrophobic residues losing more than 20 $Å^2$ of solvent accessible surface upon domain-domain association and residues participating in domain-domain salt bridges are labeled.

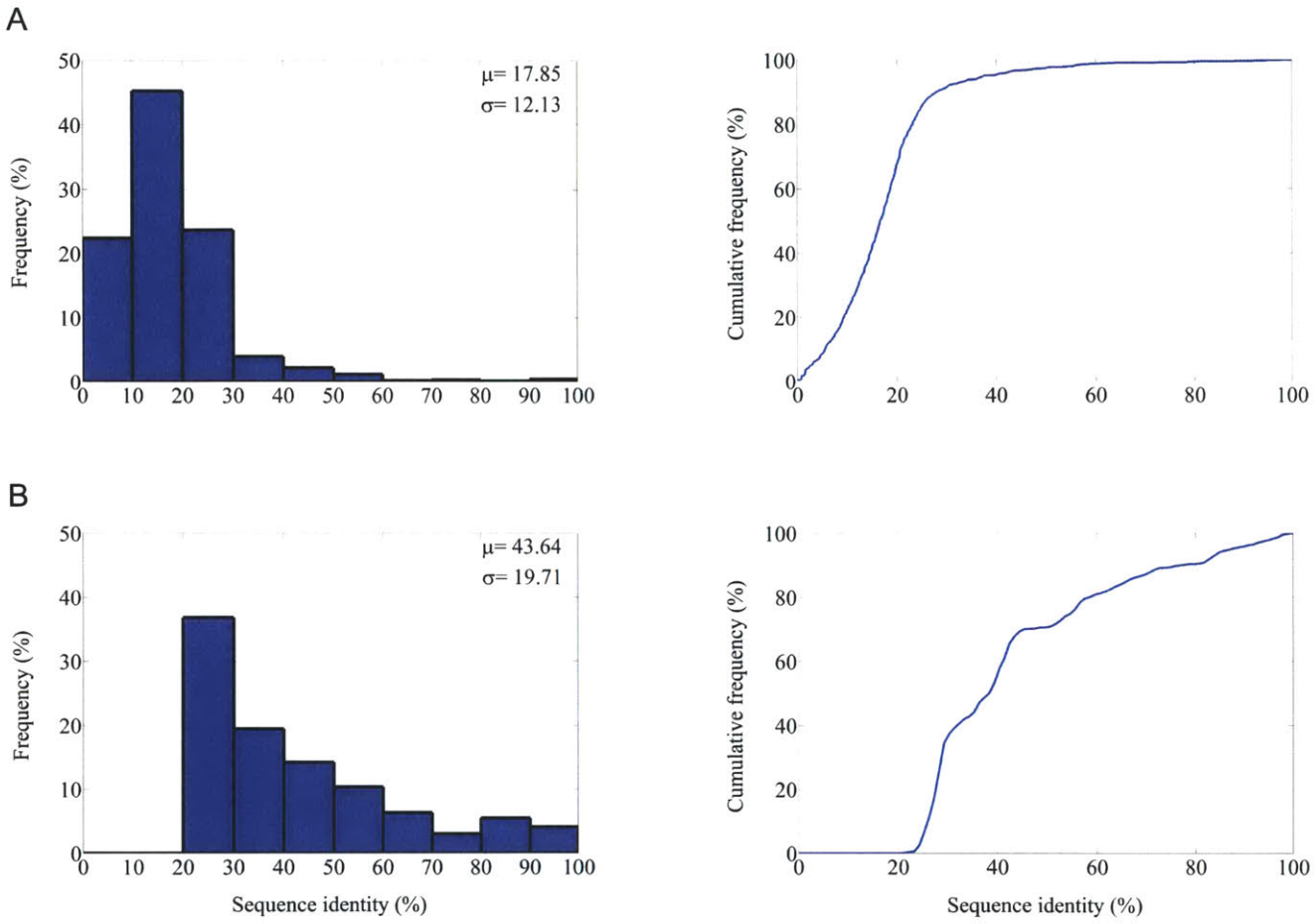

Figure C-3 – Distributions of pair-wise sequence identities of β-trefoil domains and homologous fascins. Histograms and cumulative distributions of pair-wise sequence identities between (A) 59 β-trefoil domains available in the PDB and (B) 61 homologous fascins given in Table C.7. Mean values (μ) and standard deviations (σ) are shown. Pair-wise sequence alignments are performed using the Needleman-Wunsch global alignment algorithm [188] implemented in EMBOSS [189]. Sequence identity is calculated as the total number of identical residue pairs between the two aligned sequences divided by the length of the shorter sequence. 90% of β-trefoil domain pairs and 90% of homologous fascin pairs are less than 28% and 78% identical, respectively. This observation implies significant phylogenetic diversity among these homologous sequences.

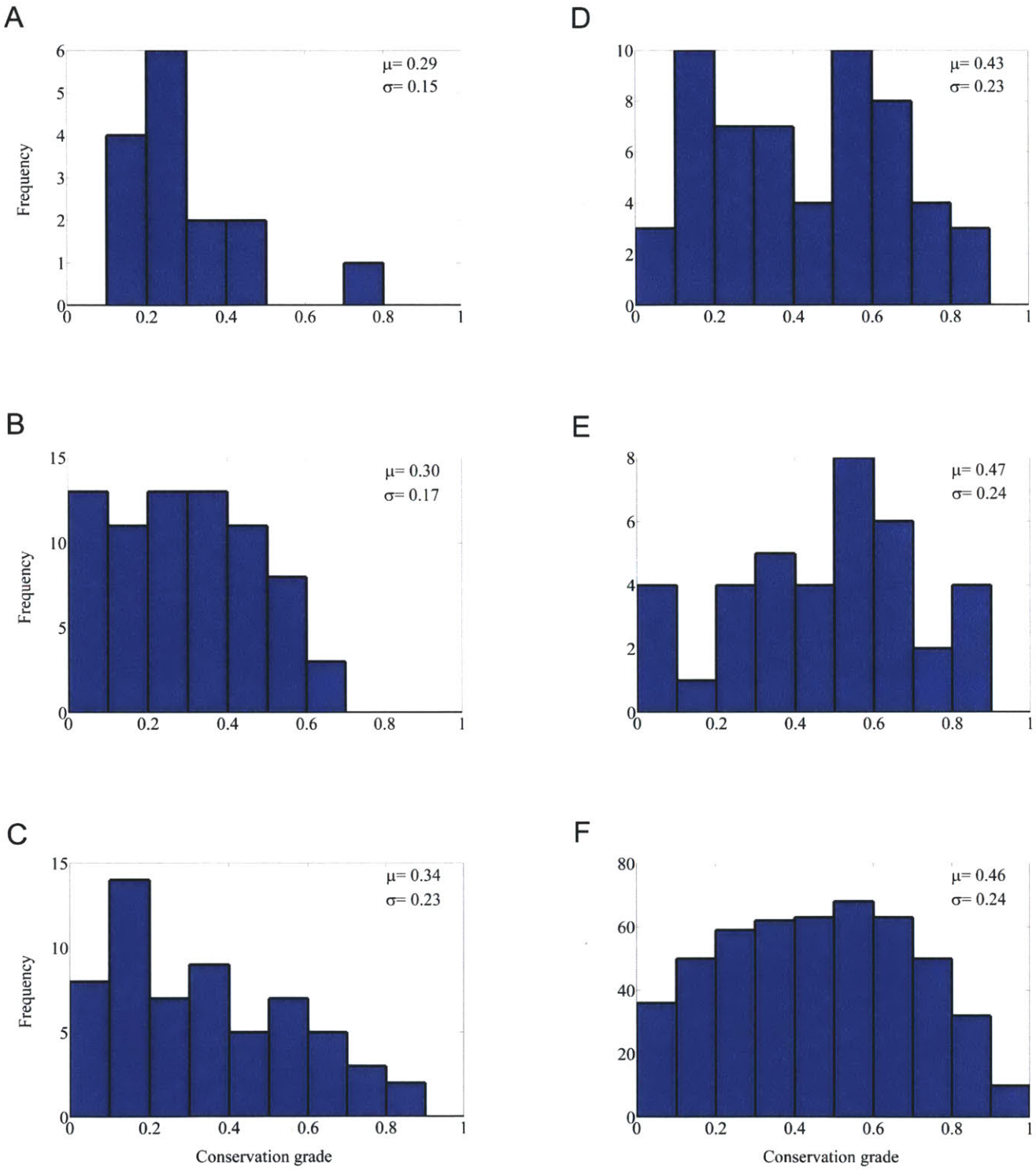

Figure C-4 – Histograms of conservation grades across homologous fascins. Histograms of conservation grades measured across homologous fascin molecules for (A) residues 29–43, (B) hydrophobic core stabilizing residues, (C) interfacial residues between lobes, (D) interfacial residues between domains F1 and F2, (E) interfacial residues between domains F3 and F4, and (F) all fascin-1 residues. Zero denotes maximal conservation and one indicates maximal variability. Mean values (μ) and standard deviations (σ) are shown. Residues 29–43 (a putative actin-binding site), hydrophobic core stabilizing residues, interfacial residues between lobes, and interfacial residues between domains within each lobe are expected to be conserved due to their importance to F-actin binding, β-trefoil fold stability, overall fascin-1 stability, and structural integrity of the lobes, respectively. Residues 29–43, hydrophobic core stabilizing residues, and interfacial residues between lobes are in general conserved, as expected.

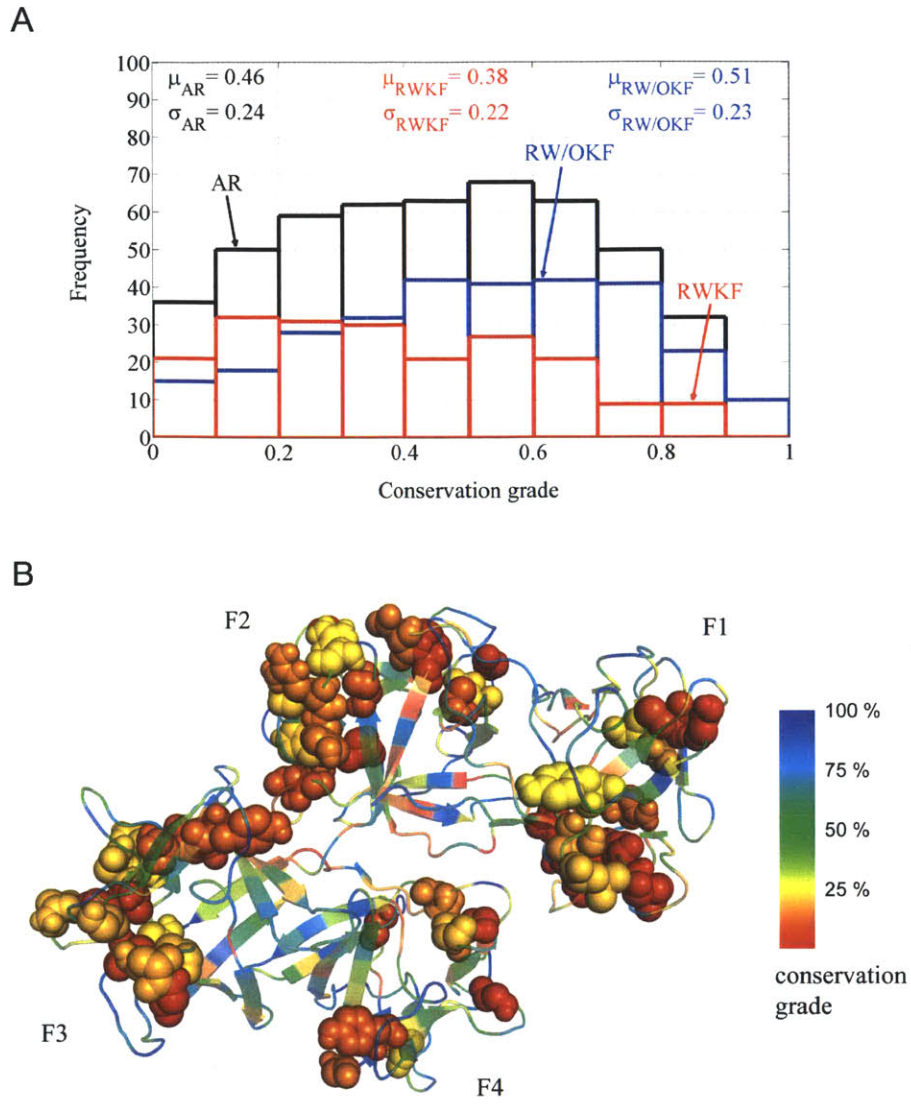

Figure C-5 – Functional analysis of residues of fascin-1. (A) Histograms of conservation grades for all residues (AR) of fascin-1, residues with known function (RWKF) that include hydrophobic core stabilizing residues, interfacial residues and residues 29–43, and residues without known function (RW/OKF) that include all residues of fascin-1 except for hydrophobic core stabilizing residues, interfacial residues, and residues 29–43. Mean values (μ) and standard deviations (σ) are shown. The distribution of conservation grades for RWKF is biased towards high conservation. This bias is in contrast to the distribution of conservation grades for RW/OKF, which is biased towards low conservation. 59% and 41% of highly conserved residues across all homologous fascins are comprised of RWKF and RW/OKF, respectively (also see Tables C.5 and C.6). (B) View I (see Chapter 3) of fascin-1 colored according to conservation grade measured across homologous fascins. Conservation grade varies from blue to red, denoting highly variable and highly conserved positions, respectively. Highly conserved RW/OKF are represented as spheres and remaining residues are drawn as a transparent cartoon. Residues that are highly conserved across fascins but not generally across β-trefoils include residues 23, 25, 27, and 44, which are located near residues 29–43, a putative actin-binding site (Fig. 3-7-B and Table C.6); residues 274, 283–285, 292, and 360–361, which are located near S274; and residues 149, 151, 154–155, 166, 174, 235, and 251, which are located near P236 (Fig. 3-7-B). These residues constitute approximately two-fifths of highly conserved RW/OKF, with another approximately one-third consisting of glycines, mostly located in loops (Table C.6).

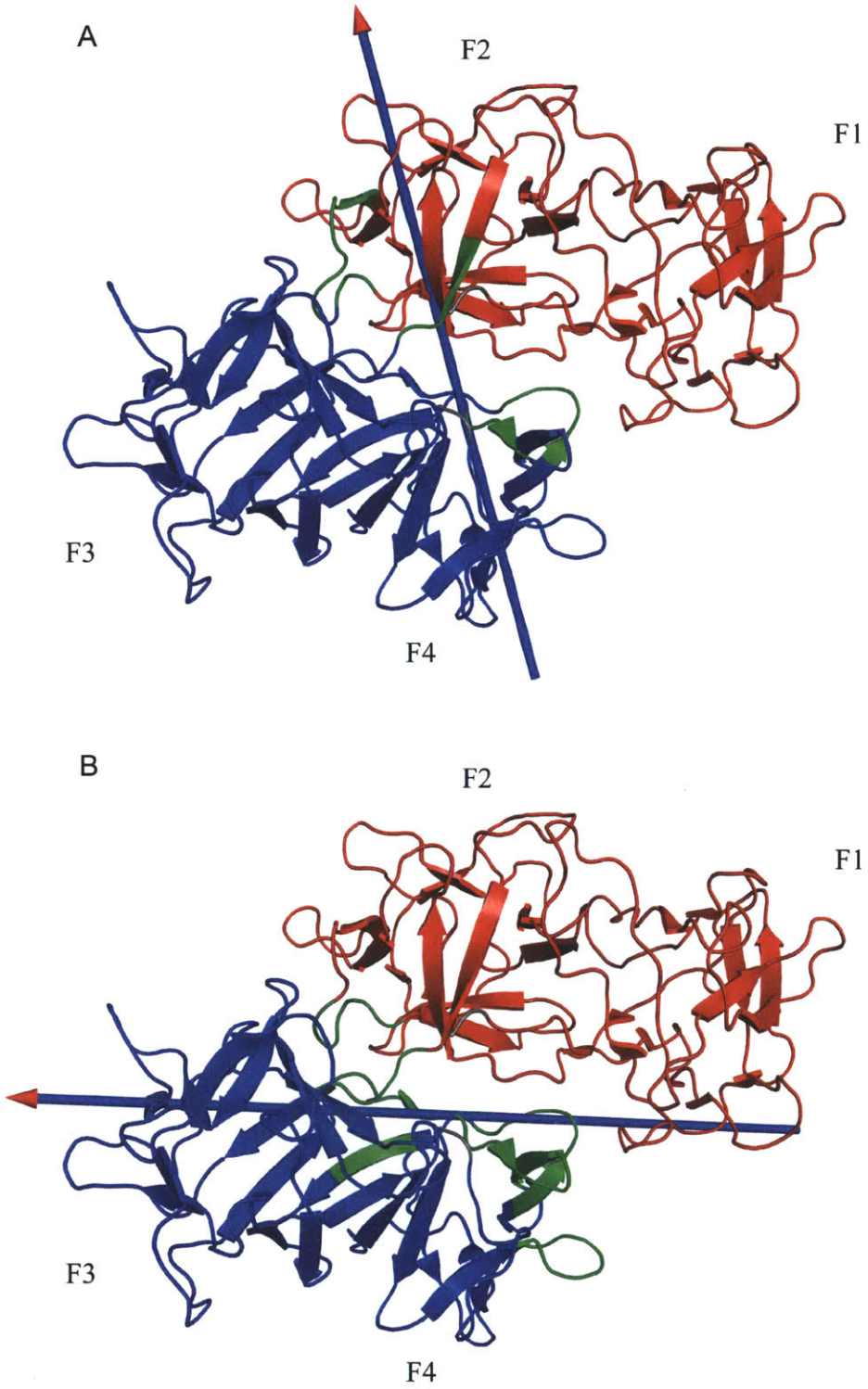

Figure C-6 – The two lowest normal modes of fascin-1. (A) The lowest normal mode is a hinge-like motion in which the β-barrel lobes twist as rigid bodies relative to each other in a scissor-like motion. The axis of rotation indicated by the arrow is generated using DynDom [190] with the maximum partition-size that generates a hinge when analyzing two conformational substates in the direction of the normal mode under analysis. Hinge-domains are colored red and blue, and hinge-bending residues are colored green. (B) The second lowest normal mode is a hinge-like motion in which each lobe rolls along its long axis with respect to the other one as a rigid body.

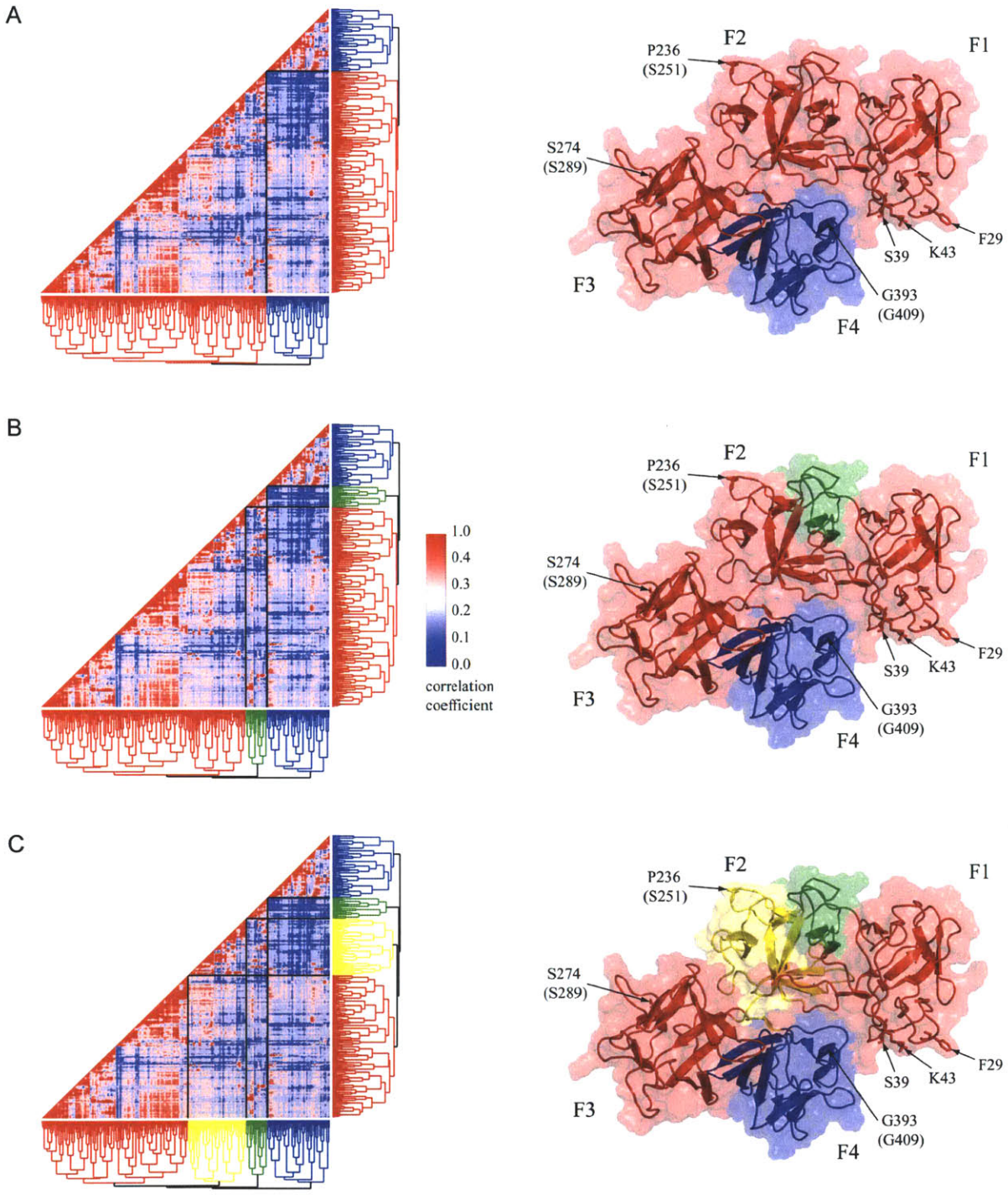

Figure C-7 – Correlated dynamical motions of fascin-1. Average-link hierarchical clustering is used to identify residue clusters that are highly correlated based on the magnitude of their generalized linear mutual information coefficient. The triangle shows inter-residue correlations after clustering. View I of fascin-1 is colored according to the different clusters shown in the dendrogram. Point mutations affecting fascin-1 function and residues 29 and 43 are indicated. Residue numbers in parentheses indicate the analogous positions in *D. melanogaster* fascin. Fascin-1 is clustered into (A) two, (B) three, and (C) four clusters. Domains F1 and F3 remain in the same cluster in each partitioning, despite the fact that the domains are not in direct physical contact. This dynamical coupling suggests a potential allosteric mechanism involving these two distant domains.

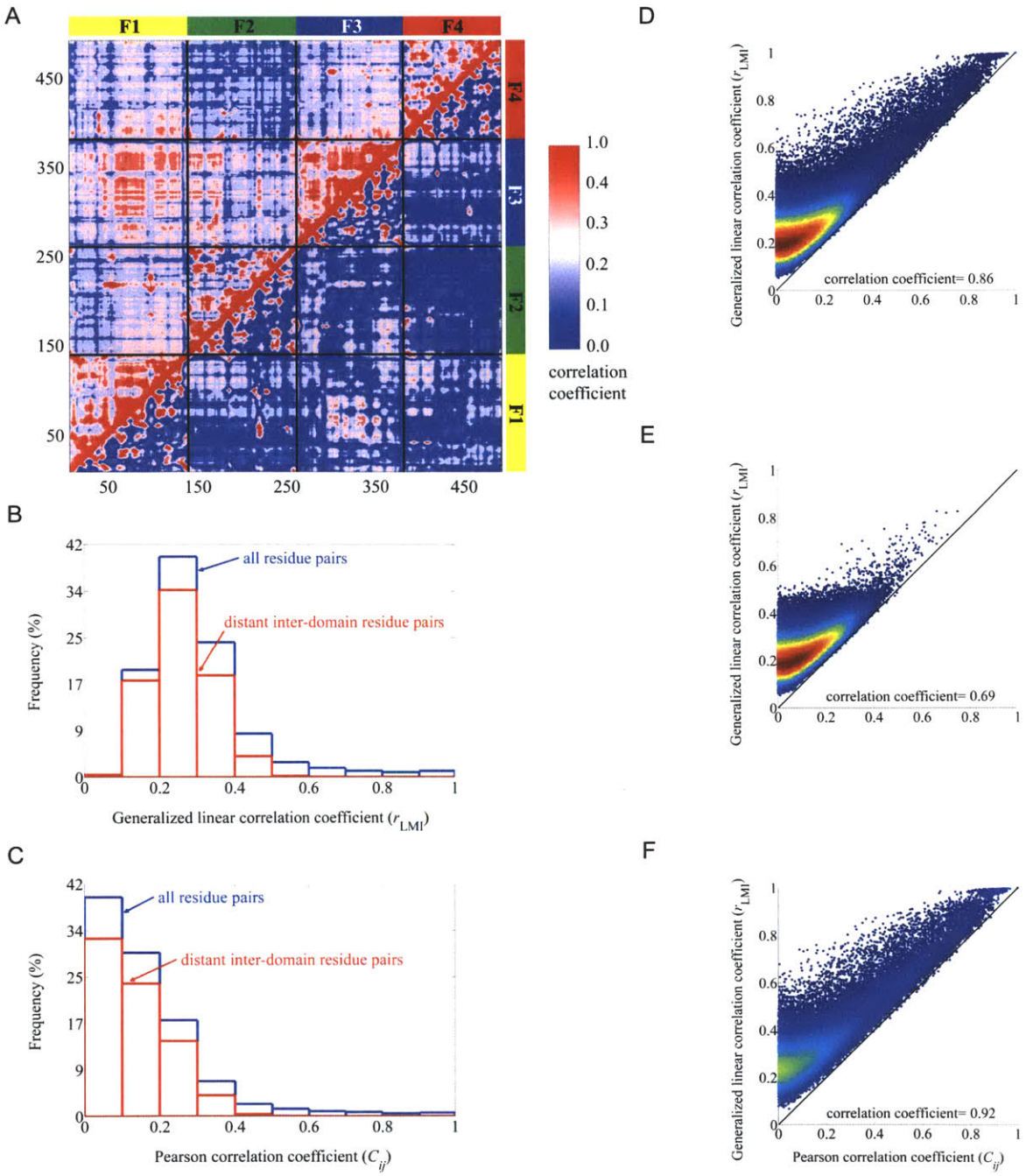

Figure C-8 – Analysis of correlation coefficients between C_α atom thermal fluctuations in fascin-1. (A) Dynamical correlation matrix for conformational fluctuations of fascin-1. The upper triangle shows the generalized linear mutual information correlation coefficient and the lower triangle shows the Pearson correlation coefficient. Strong coupling between distant (see Table C.1) β-trefoil domains F1 and F3 is evident using both measures, although the correlation is more apparent in the mutual information metric. Histograms of correlation coefficients between all residue pairs (blue) and between distant inter-domain residue pairs (red) calculated using (B) mutual information and (C) Pearson correlation. Two residues are defined to be in contact if any of their two heavy atoms are within 5 Å of one another [169]. Similarly, a residue pair is defined to be "distant" if the minimum distance between all of their heavy atoms is greater than 5 Å. Thermal fluctuations of residue pairs that are close spatially and/or in the same β-trefoil domain are generally more correlated than distant inter-domain residue pairs due to physical contact and structural integrity of the β-trefoil fold. Scatter plots between Pearson correlation coefficients (C_{ij}) and generalized linear correlation coefficients (r_{LMI}) of (D) all residue pairs, (E) distant inter-domain residue pairs, and (F) close and/or intra-domain residue pairs. Solid lines denote a slope of one. As seen, the generalized linear correlation coefficient is at least as large as the Pearson correlation coefficient because the latter neglects non-colinear correlated atomic motions. Warmer colors denote a higher density of points. Notwithstanding, the two metrics are highly correlated with one another, as shown.

C.2 Supplementary tables

Table C.1 – Solvent-accessible surface area ($Å^2$) buried between β-trefoil domain-domain interfaces in fascin-1.

	F1	F2	F3	F4
F1	–	2,206	0	702
F2	–	–	756	1,225
F3	–	–	–	1,787

Table C.2 – RMSDs between the pair-wise aligned β-trefoil domains of fascin-1 (F1–F4) given in Å for each pair of domains. The number of C_α atom pair equivalences is shown in parentheses.

	F1	F2	F3	F4
F1	–	2.11 (115)	1.91 (119)	2.23 (110)
F2	–	–	1.70 (114)	1.79 (104)
F3	–	–	–	1.32 (108)

Table C.3 – Sequence identity between domains of fascin-1 and other β-trefoil domains available in the PDB.

	Mean sequence identity between a fascin-1 β-trefoil domain and other β-trefoil domains (%)	Standard deviation of sequence identities between a fascin-1 β-trefoil domain and other β-trefoil domains (%)	Mean sequence identity between a fascin-1 β-trefoil domain and other fascin-1 domains (%)
F1	12	3	17
F2	9	3	16
F3	12	4	19
F4	12	4	22

Table C.4 – Structural similarity between fascin-1 domains and other β-trefoil domains available in the PDB.

	Mean RMSD between a fascin-1 β-trefoil domain and other β-trefoil domains (Å)	Standard deviation of RMSDs between a fascin-1 β-trefoil domain and other β-trefoil domains (Å)	Mean RMSD between a fascin-1 β-trefoil domain and other fascin-1 domains (Å)
F1	2.31	0.40	2.08
F2	2.26	0.30	1.87
F3	2.20	0.35	1.64
F4	2.28	0.33	1.78

Table C.5 – Residue type, number of residues of specific residue type, fraction of residues of specific residue type (in parentheses) and residue numbers of the fifty-one highly conserved residues across homologous fascin molecules that are not included in the set of hydrophobic core stabilizing residues, interfacial residues, and residues 29–43 (see also Fig. C-5).

Residue type	Number of residues	Residue number(s)
ALA	6 (11.8%)	222, 267, 283, 349, 365, and 480
ARG	2 (3.9%)	149 and 151
ASN	4 (7.8%)	18, 284, 351, and 360
ASP	5 (9.8%)	166, 174, 225, 251, and 412
CYS	1 (2.0%)	121
GLN	2 (3.9%)	44, and 285
GLU	4 (7.8%)	27, 130, 252, and 292
GLY	14 (27.5%)	15, 76, 113, 228, 235, 321, 352, 361, 390, 393, 396, 430, 467, and 476
HIS	1 (2.0%)	154
ILE	0 (0.0%)	–
LEU	3 (5.9%)	66, 155, and 202
LYS	0 (0.0%)	–
MET	0 (0.0%)	–
PHE	0 (0.0%)	–

Continued on Next Page...

Table C.5 – Continued

Residue type	Number of residues	Residue number(s)
PRO	0 (0.0%)	–
SER	3 (5.9%)	64, 146, and 274
THR	1 (2.0%)	25
TRP	0 (0.0%)	–
TYR	5 (9.8%)	23, 69, 152, 314, and 469
VAL	0 (0.0%)	–

Table C.6 – Residue type, residue number, and conservation grades across β-trefoil domains (CGTD) available in the PDB, conservation grades across homologous fascin (CGHF) molecules, fraction of corresponding column which is of type "gap" (FCCTG) in the structure-based sequence alignment of the 59 β-trefoil domains available in the PDB, and potential functional reason for conservation of the fifty-one highly conserved residues across homologous fascin molecules that are not included in the set of hydrophobic core stabilizing residues, interfacial residues, and residues 29–43 (see also Fig. C-5). According to conservation grade, Asp166 and Asn284 are highly conserved across 59 β-trefoil domains; however, because their corresponding columns in the structure-based sequence alignment of the 59 β-trefoil domains consist mostly of gaps (90%), these two residues are considered as variable across the 59 β-trefoil domains and their corresponding rows in this table are colored red.

Residue type	Residue number	CGTD (%)	CGHF (%)	FCCTG (%)	Potential functional reason for conservation
ALA	222	22.8	11.2	1.7	β-trefoil stability/function
ALA	267	13.2	24.3	1.7	β-trefoil stability/function

Continued on Next Page...

Table C.6 – Continued

Residue type	Residue number	CGTD (%)	CGHF (%)	FCCTG (%)	Potential functional reason for conservation
ALA	283	35.6	5.1	5.1	fascin-1 stability/function (near S274)
ALA	349	21.6	20.9	1.7	β-trefoil stability/function
ALA	365	19.3	7.7	3.4	β-trefoil stability/function
ALA	480	20.4	23.3	3.4	β-trefoil stability/function
ARG	149	45.7	23.9	17	fascin-1 stability/function (near P236)
ARG	151	43.4	15.2	1.7	fascin-1 stability/function (near P236)
ASN	18	13	14.2	1.7	β-trefoil stability/function
ASN	284	4.7	21.7	89.8	fascin-1 stability/function (near S274)
ASN	351	62.1	18.9	5.1	unknown
ASN	360	81.7	20.7	10.2	fascin-1 stability/function (near S274)
ASP	166	4.9	21.5	89.8	fascin-1 stability/function (near P236)
ASP	174	33.5	10.8	27.1	fascin-1 stability/function (near P236)

Continued on Next Page...

Table C.6 – Continued

Residue type	Residue number	CGTD (%)	CGHF (%)	FCCTG (%)	Potential functional reason for conservation
ASP	225	21.8	22.1	1.7	β-trefoil stability/function
ASP	251	40.3	14.6	11.9	fascin-1 stability/function (near P236)
ASP	412	54.7	16.4	6.8	unknown
CYS	121	17.5	7.9	3.4	β-trefoil stability/function
GLU	27	100	4.7	1.7	actin-binding (near residues 29–43)
GLU	130	14.8	23.5	1.7	β-trefoil stability/function
GLU	252	15.8	4.9	1.7	fascin-1 stability/function (near P236)
GLU	292	48.1	10.3	5.1	fascin-1 stability/function (near S274)
GLN	44	47.5	3	5.1	actin-binding (near residues 29–43)
GLN	285	64	22.3	5.1	fascin-1 stability/function (near S274)
GLY	15	51.6	3.7	1.7	flexibility
GLY	76	73.5	0.8	13.6	flexibility
GLY	113	22.4	18.1	1.7	β-trefoil stability/function
GLY	228	55.4	2	1.7	flexibility

Continued on Next Page...

Table C.6 – Continued

Residue type	Residue number	CGTD (%)	CGHF (%)	FCCTG (%)	Potential functional reason for conservation
GLY	235	85.2	5.7	10.2	fascin-1 stability/function (near P236)
GLY	321	73.7	20.1	13.6	flexibility
GLY	352	55.6	1	1.7	flexibility
GLY	361	39.3	5.9	23.7	fascin-1 stability/function (near S274)
GLY	390	13.6	21.1	1.7	β-trefoil stability/function
GLY	393	43.6	1.2	1.7	flexibility
GLY	396	58.6	15.8	1.7	flexibility
GLY	430	38.3	8.5	0	flexibility
GLY	467	56	13.6	1.7	flexibility
GLY	476	39.7	2.8	23.7	flexibility
HIS	154	57.6	21.9	1.7	fascin-1 stability/function (near P236)
LEU	66	28.6	21.3	23.7	unknown
LEU	155	38.9	13.8	1.7	fascin-1 stability/function (near P236)
LEU	202	34.4	2.6	1.7	unknown

Continued on Next Page...

Table C.6 – Continued

Residue type	Residue number	CGTD (%)	CGHF (%)	FCCTG (%)	Potential functional reason for conservation
SER	64	29.2	17.9	0	unknown
SER	146	13.4	11	1.7	β-trefoil stability/function
SER	274	57.4	12.4	1.7	fascin-1 stability/function (near S274)
THR	25	56.2	1.8	1.7	actin-binding (near residues 29–43)
TYR	23	25.1	4.3	1.7	actin-binding (near residues 29–43)
TYR	69	14.6	25	0	β-trefoil stability/function
TYR	152	25.9	15	1.7	unknown
TYR	314	11.5	11.8	0	β-trefoil stability/function
TYR	469	32.7	9.9	1.7	unknown

Table C.7 – 61 sequences homologous to fascin-1 retrieved from the NCBI [4] and used for calculation of entropy grades.

1) gi|115494998|ref|NP_001070028.1| hypothetical protein LOC558271 [*Danio rerio*] Length=491

2) gi|130486462|ref|NP_001076338.1| hypothetical protein LOC570314 [*Danio rerio*] Length=494

3) gi|184186129|ref|NP_001116988.1| fascin 2A [*Danio rerio*] Length=488

Continued on Next Page...

Table C.7 – Continued

4) gi|183986779|ref|NP_957064.2| fascin homolog 1-like, actin-bundling protein [*Danio rerio*] Length=490

5) gi|126334482|ref|XP_001363553.1| PREDICTED: similar to molybdenum cofactor synthesis-step 1 protein [*Monodelphis domestica*] Length=494

6) gi|126308638|ref|XP_001370831.1| PREDICTED: similar to retinal fascin [*Monodelphis domestica*]Length=491

7) gi|126340773|ref|XP_001371599.1| PREDICTED: similar to fascin 3 [*Monodelphis domestica*] Length=500

8) gi|78045491|ref|NP_001030217.1| fascin homolog 1, actin-bundling protein [*Bos taurus*] Length=493

9) gi|28603742|ref|NP_788806.1| fascin 2 [*Bos taurus*] Length=492

10) gi|164451463|ref|NP_001069011.2| fascin homolog 3, actin-bundling protein, testicular [*Bos taurus*] Length=498

11) gi|88660673|gb|ABD48096.1| fascin-1 [*Xenopus tropicalis*] Length=484

12) gi|154147674|ref|NP_001093724.1| fascin homolog 2, actin-bundling protein, retinal [*Xenopus tropicalis*] Length=492

13) gi|147900999|ref|NP_001081581.1| fascin [*Xenopus laevis*] Length=484

14) gi|50603997|gb|AAH77847.1| FSCN1 protein [*Xenopus laevis*] Length=502

15) gi|2498360|sp|Q91837.1|FASC_XENLA RecName: Full=Fascin [*Xenopus laevis*] Length=483

16) gi|189217818|ref|NP_001121350.1| fascin 2 [*Xenopus laevis*] Length=492

Continued on Next Page...

Table C.7 – Continued

17) gi|4507115|ref|NP_003079.1| fascin 1 [*Homo sapiens*] Length=493

18) gi|6912626|ref|NP_036550.1| fascin 2 isoform 1 [*Homo sapiens*] Length=492

19) gi|9966791|ref|NP_065102.1| fascin 3 [*Homo sapiens*] Length=498

---

20) gi|73958049|ref|XP_546998.2| PREDICTED: similar to Fascin (Singed-like protein) (55 kDa actin bundling protein) (p55) isoform 1 [*Canis familiaris*] Length=477

21) gi|57099437|ref|XP_540481.1| PREDICTED: similar to Fascin 2 (Retinal fascin) [*Canis familiaris*] Length=492

---

22) gi|113680348|ref|NP_032010.2| fascin homolog 1, actin bundling protein [*Mus musculus*] Length=493

23) gi|80479179|gb|AAI09357.1| Fascin homolog 2, actin-bundling protein, retinal (*Strongylocentrotus purpuratus*) [*Mus musculus*] Length=492

24) gi|31982710|ref|NP_062515.2| fascin 3 [*Mus musculus*] Length=498

---

25) gi|201066380|ref|NP_001094276.1| fascin [*Rattus norvegicus*] Length=493

26) gi|157818957|ref|NP_001100542.1| fascin homolog 2, actin-bundling protein, retinal [*Rattus norvegicus*] Length=492

27) gi|51948446|ref|NP_001004232.1| fascin homolog 3, actin-bundling protein, testicular [*Rattus norvegicus*] Length=498

---

28) gi|167537922|ref|XP_001750628.1| predicted protein [*Monosiga brevicollis MX1*] Length=489

Continued on Next Page...

Table C.7 – Continued

29) gi|167526499|ref|XP_001747583.1| predicted protein [*Monosiga brevicollis MX1*] Length=499

30) gi|47551049|ref|NP_999701.1| fascin [*Strongylocentrotus purpuratus*] Length=496

31) gi|109119060|ref|XP_001110926.1| PREDICTED: fascin 2 [*Macaca mulatta*] Length=492

32) gi|109068075|ref|XP_001089987.1| PREDICTED: fascin 3 [*Macaca mulatta*] Length=498

33) gi|156394995|ref|XP_001636897.1| predicted protein [*Nematostella vectensis*] Length=488

34) gi|47219083|emb|CAG00222.1| unnamed protein product [*Tetraodon nigroviridis*] Length=475

35) gi|47226056|emb|CAG04430.1| unnamed protein product [*Tetraodon nigroviridis*] Length=490

36) gi|226372953|ref|NP_001139772.1| fascin homolog 1, actin-bundling protein [*Sus scrofa*] Length=493

37) gi|113205632|ref|NP_001038012.1| fascin 3 [*Sus scrofa*] Length=498

38) gi|194218724|ref|XP_001914942.1| PREDICTED: similar to fascin (Singed-like protein) (55 kDa actin-bundling protein) (p55) [*Equus caballus*] Length=478

Continued on Next Page...

Table C.7 – Continued

39) gi|194216553|ref|XP_001914983.1| PREDICTED: similar to retinal fascin [*Equus caballus*] Length=492 40) gi|149706195|ref|XP_001502608.1| PREDICTED: similar to fascin 3 [*Equus caballus*] Length=498
41) gi|198423660|ref|XP_002129293.1| PREDICTED: similar to Fascin (Singed-like protein) (55 kDa actin-bundling protein) (p55) [*Ciona intestinalis*] Length=487
42) gi|91089337|ref|XP_972494.1| PREDICTED: similar to fascin [*Tribolium castaneum*] Length=518
43) gi|195396945|ref|XP_002057089.1| GJ16540 [*Drosophila virilis*] Length=512
44) gi|194897139|ref|XP_001978598.1| GG19678 [*Drosophila erecta*] Length=512
45) gi|195132414|ref|XP_002010638.1| GI21600 [*Drosophila mojavensis*] Length=512
46) gi|157137463|ref|XP_001664000.1| fascin [*Aedes aegypti*] Length=511
47) gi|195448228|ref|XP_002071566.1| GK25072 [*Drosophila willistoni*] Length=512
48) gi|195045719|ref|XP_001992025.1| GH24440 [*Drosophila grimshawi*] Length=512

Continued on Next Page...

Table C.7 – Continued

49) gi|24640473|ref|NP_727226.1| singed, isoform A [*Drosophila melanogaster*] Length=512
50) gi|212506942|gb|EEB11002.1| protein singed, putative [*Pediculus humanus corporis*] Length=514
51) gi|156551822|ref|XP_001604095.1| PREDICTED: similar to ENSANGP00000010187 [*Nasonia vitripennis*] Length=517
52) gi|195356008|ref|XP_002044475.1| GM11990 [*Drosophila sechellia*] Length=512
53) gi|221127727|ref|XP_002166129.1| PREDICTED: similar to fascin homolog 1-like, actin-bundling protein [*Hydra magnipapillata*] Length=486 54) gi|221113999|ref|XP_002155122.1| PREDICTED: similar to fascin 2 [*Hydra magnipapillata*] Length=495
55) gi|209571732|gb|ACI62521.1| fascin 3 (predicted) [*Oryctolagus cuniculus*] Length=498
56) gi|223718854|gb|ACN22213.1| fascin 3 (predicted) [*Dasypus novemcinctus*] Length=498
57) gi|177771976|gb|ACB73265.1| fascin homolog 3, actin-bundling protein, testicular (predicted) [*Rhinolophus ferrumequinum*] Length=498
58) gi|169246066|gb|ACA51044.1| fascin 3 (predicted) [*Callicebus moloch*] Length=498

Continued on Next Page...

Table C.7 – Continued

59) gi|195977113|gb|ACG63662.1| fascin 3 (predicted) [*Otolemur garnettii*] Length=498
60) gi|167427285|gb|ABZ80263.1| fascin 3 (predicted) [*Callithrix jacchus*] Length=498
61) gi|114615791|ref|XP_519352.2| PREDICTED: fascin 3 [*Pan troglodytes*] Length=498

C.3 Supplementary computational procedures

C.3.1 Structural similarity and sequence identity of fascin-1 β-trefoil domains to β-trefoil domains available in the PDB

Each β-trefoil domain of fascin-1 is structurally aligned to each of the other 58 β-trefoil domains available in the PDB using STAMP [178] implemented in VMD 1.8.6 [181]. The β-trefoil domains are identified using the structural classification of proteins (SCOP) database [153]. Subsequently, sequence alignments are performed based on pair-wise structural alignments. Structural similarity and sequence identity between each fascin-1 β-trefoil domain and the other β-trefoil domains available in the PDB are calculated (Fig. C-1-A and Tables C.3 and C.4). Structural similarity is evaluated by RMSD of positions of C_α atom pair equivalences of the two β-trefoils after superposition of their structures. Sequence identity is calculated by division of the number of residue identities between two structurally aligned sequences by the shorter sequence length.

C.3.2 Conservation analysis over all β-trefoil domains available in the PDB

The set of 59 β-trefoil domains available in the PDB is structurally aligned using STAMP implemented in VMD 1.8.6. Conservation analysis is performed on the structure-based multiple sequence alignment of these β-trefoil domains. The conservation grade is the combination of three different methods: the conservation surface mapping method (ConSurf) [170, 191], the real-valued evolutionary trace method (ET) [172, 192], and a simple entropy-based method using a 21-letter alphabet [164]. Conservation grades from these three methods are weighted and combined such that each contributes equally to form the final grade. Highly conserved residues are defined as those residing in the first quartile of residues in the molecule as ranked by conservation grade.

C.3.3 Calculation of marginal-covariances and pair-covariance matrix of atoms

Marginal-covariances of atom i, $\mathbf{C}_{(i)} = \langle \mathbf{x}_i{}^{\mathrm{T}} \mathbf{x}_i \rangle$, and the pair-covariance matrix of atoms i and j , $\mathbf{C}_{(ij)} = \langle (\mathbf{x}_i, \mathbf{x}_j)^{\mathrm{T}} (\mathbf{x}_i, \mathbf{x}_j) \rangle$, are calculated using atomic fluctuations computed using NMA, given by $\langle x_{li} x_{oj} \rangle = k_B T \sum_k \frac{y_{lik} y_{ojk}}{\lambda_k \sqrt{m_i} \sqrt{m_j}}$, where $k_B T$ is thermal energy, m_i is the mass of atom i, λ_k is eigenvalue k, and y_{lik} is the component of mode k associated with the displacement component l of atom i, x_{li}.

Bibliography

[1] Matsumura M., Wozniak J. A., Daopin S., and Matthews B. W. Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *Journal of Biological Chemistry*, 264:16059–16066, 1989.

[2] Otterbein L. R., Graceffa P., and Dominguez R. The crystal structure of uncomplexed actin in the ADP state. *Science*, 293:708–711, 2001.

[3] Stein P. E., Boodhoo A., Armstrong G. D., Cockle S. A., Klein M. H., and Read R. J. The crystal structure of pertussis toxin. *Structure*, 2:45–57, 1994.

[4] Wheeler D. L., Church D. M., Federhen S., Lash A. E., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Sequeira E., Tatusova T. A., and Wagner L. Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, 31:28–33, 2003.

[5] Zheng W. J. and Brooks B. R. Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: Myosin versus kinesin. *Biophysical Journal*, 89:167–178, 2005.

[6] Cecchini M., Houdusse A., and Karplus M. Allosteric communication in myosin V: From small conformational changes to large directed movements. *PLoS Computational Biology*, 4:(e1000129)1–19, 2008.

[7] Zheng W. J., Brooks B. R., and Hummer G. Protein conformational transitions explored by mixed elastic network models. *Proteins: Structure, Function, and Bioinformatics*, 69:43–57, 2007.

[8] Keskin O., Durell S. R., Bahar I., Jernigan R. L., and Covell D. G. Relating molecular flexibility to function: A case study of tubulin. *Biophysical Journal*, 83:663–680, 2002.

[9] ben-Avraham D. and Tirion M. M. Dynamic and elastic properties of F-actin: A normal-modes analysis. *Biophysical Journal*, 68:1231–1245, 1995.

[10] Mouawad L. and Perahia D. Diagonalization in a mixed basis: A method to compute low-frequency normal modes for large macromolecules. *Biopolymers*, 33:599–611, 1993.

[11] Cui Q. and Bahar I., editors. *Normal mode analysis: Theory and applications to biological and chemical systems.* Chapman & Hall/CRC, Boca Raton, 2006.

[12] Miller B. T., Zheng W. J., Venable R. M., Pastor R. W., and Brooks B. R. Langevin network model of myosin. *Journal of Physical Chemistry B*, 112:6274–6281, 2008.

[13] Bathe K. J. *Finite element procedures.* Prentice Hall, Upper Saddle River, New Jersey, 1996.

[14] Sedeh R. S., Bathe M., and Bathe K. J. The subspace iteration method in protein normal mode analysis. *Journal of Computational Chemistry*, 31:66–74, 2010.

[15] Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45:255–282, 1950.

[16] Tama F. and Brooks C. L. Symmetry, form, and shape: Guiding principles for robustness in macromolecular machines. *Annual Review of Biophysics and Biomolecular Structure*, 35:115–133, 2006.

[17] Lamm G. and Szabo A. Langevin modes of macromolecules. *Journal of Chemical Physics*, 85:7334–7348, 1986.

[18] Bathe M. A finite element framework for computation of protein normal modes and mechanical response. *Proteins: Structure, Function, and Bioinformatics*, 70:1595–1609, 2008.

[19] Carrasco B. and Garcia de la Torre J. Hydrodynamic properties of rigid particles: Comparison of different modeling and computational procedures. *Biophysical Journal*, 76:3044–3057, 1999.

[20] Sedeh R. S., Fedorov A. A., Fedorov E. V., Ono S., Matsumura F., Almo S. C., and Bathe M. Structure, evolutionary conservation, and conformational dynamics of *Homo sapiens* fascin-1, an F-actin crosslinking protein. *Journal of Molecular Biology*, 400:589–604, 2010.

[21] Brooks B. R., Janezic D., and Karplus M. Harmonic analysis of large systems. I. Methodology. *Journal of Computational Chemistry*, 16:1522–1542, 1995.

[22] Brooks B. R. and Karplus M. Normal modes for specific motions of macromolecules: Application to the hinge-bending mode of lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 82:4995–4999, 1985.

[23] Tama F. and Brooks C. L. Unveiling molecular mechanisms of biological functions in large macromolecular assemblies using elastic network normal mode analysis. In Cui Q. and Bahar I., editors, *Normal mode analysis: Theory and*

applications to biological and chemical systems, pages 111–135. Chapman & Hall/CRC, Boca Raton, 2006.

[24] Bahar I., Atilgan A. R., and Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*, 2:173–181, 1997.

[25] Van Wynsberghe A., Li G. H., and Cui Q. Normal-mode analysis suggests protein flexibility modulation throughout RNA polymerase's functional cycle. *Biochemistry*, 43:13083–13096, 2004.

[26] Sherman W. and Tidor B. Novel method for probing the specificity binding profile of ligands: Applications to HIV protease. *Chemical Biology & Drug Design*, 71:387–407, 2008.

[27] Lee J., Natarajan M., Nashine V. C., Socolich M., Vo T., Russ W. P., Benkovic S. J., and Ranganathan R. Surface sites for engineering allosteric control in proteins. *Science*, 322:438–442, 2008.

[28] Bathe K. J. Solution methods for large generalized eigenvalue problems in structural engineering. Report ucsesm 71-20, Department of Civil Engineering, University of California, Berkeley, 1971.

[29] Bathe K. J. and Wilson E. L. Solution methods for eigenvalue problems in structural mechanics. *International Journal for Numerical Methods in Engineering*, 6:213–266, 1973.

[30] Bathe K. J. and Ramaswamy S. An accelerated subspace iteration method. *Computer Methods in Applied Mechanics and Engineering*, 23:313–331, 1980.

[31] Akl F. A., Dilger W. H., and Irons B. M. Over-relaxation and subspace iteration. *International Journal for Numerical Methods in Engineering*, 14:629–630, 1979.

[32] Akl F. A., Dilger W. H., and Irons B. M. Acceleration of subspace iteration. *International Journal for Numerical Methods in Engineering*, 18:583–589, 1982.

[33] Jung H. J., Kim M. C., and Lee I. W. An improved subspace iteration method with shifting. *Computers & Structures*, 70:625–633, 1999.

[34] Pradlwarter H. J., Schueller G. I., and Szekely G. S. Random eigenvalue problems for large systems. *Computers & Structures*, 80:2415–2424, 2002.

[35] Zhao Q. C., Chen P., Peng W. B., Gong Y. C., and Yuan M. W. Accelerated subspace iteration with aggressive shift. *Computers & Structures*, 85:1562–1578, 2007.

[36] Wang X. and Zhou J. An accelerated subspace iteration method for generalized eigenproblems. *Computers & Structures*, 71:293–301, 1999.

[37] Qian Y. Y. and Dhatt G. An accelerated subspace method for generalized eigenproblems. *Computers & Structures*, 54:1127–1134, 1995.

[38] Bauer F. L. Das Verfahren der Treppen-Iteration und Verwandte Verfahren zur Lösung Algebraischer Eigenwertprobleme. *Zeitschrift für Angewandte Mathematik und Physik*, 8:214–235, 1957.

[39] Rutishauser H. Computational aspects of F.L. Bauer's simultaneous iteration method. *Numerische Mathematik*, 13:4–13, 1969.

[40] Bathe K. J. Convergence of subspace iteration. In Bathe K. J., Oden J. T., and Wunderlich W., editors, *Formulations and Computational Algorithms in Finite Element Analysis*, pages 575–598. MIT Press, Cambridge, MA, 1977.

[41] Paige C. C. Computational variants of the Lanczos method for the eigenproblem. *IMA Journal of Applied Mathematics*, 10:373–381, 1972.

[42] Ericsson T. and Ruhe A. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Mathematics of Computation*, 35:1251–1268, 1980.

[43] Elber R. and Karplus M. Low-frequency modes in proteins: Use of the effective-medium approximation to interpret the fractal dimension observed in electron-spin relaxation measurements. *Physical Review Letters*, 56:394–397, 1986.

[44] ben-Avraham D. Vibrational normal-mode spectrum of globular proteins. *Physical Review B*, 47:14559–14560, 1993.

[45] Bathe K. J. The finite element method. In Wah B., editor, *Encyclopedia of Computer Science and Engineering*, pages 1253–1264. John Wiley & Sons, 2009.

[46] Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S., and Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[47] Lazaridis T. and Karplus M. Effective energy function for proteins in solution. *Proteins: Structure, Function, and Genetics*, 35:133–152, 1999.

[48] Kabsch W., Mannherz H. G., Suck D., Pai E. F., and Holmes K. C. Atomic structure of the actin:DNase I complex. *Nature*, 347:37–44, 1990.

[49] DeLano W. The PyMOL molecular graphics program (http://www.pymol.org). 2002.

[50] Muller C. W., Schlauderer G. J., Reinstein J., and Schulz G. E. Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding. *Structure*, 4:147–156, 1996.

[51] Muller C. W. and Schulz G. E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap_5A refined at 1.9 Å resolution: A model for a catalytic transition state. *Journal of Molecular Biology*, 224:159–177, 1992.

[52] Sanner M. F., Olson A. J., and Spehner J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38:305–320, 1996.

[53] Heckbert P. S. and Garland M. Optimal triangulation and quadric-based surface simplification. *Computational Geometry: Theory and Applications*, 14:49–65, 1999.

[54] Garland M. Quadric-based polygonal surface simplification. PhD thesis, Carnegie Mellon University, 1999.

[55] Garland M. and Heckbert P. S. Surface simplification using quadric error metrics. In *SIGGRAPH '97*, Los Angeles, CA, USA, August 1997.

[56] Cignoni P., Callieri M., Corsini M., Dellepiane M., Ganovelli F., and Ranzuglia G. MeshLab: An open-source mesh processing tool. In *Proceedings of Sixth Eurographics Italian Chapter Conference*, pages 129–136, Fisciano, Italy, 2008.

[57] Tama F., Gadea F. X., Marques O., and Sanejouand Y. H. Building-block approach for determining low frequency normal modes of macromolecules. *Proteins: Structure, Function, and Genetics*, 41:1–7, 2000.

[58] Li G. H. and Cui Q. A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca^{2+}-ATPase. *Biophysical Journal*, 83:2457–2474, 2002.

[59] Tama F., Miyashita O., and Brooks C. L. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *Journal of Molecular Biology*, 337:985–999, 2004.

[60] Maschho K. and Sorensen D. A portable implementation of ARPACK for distributed memory parallel architectures. In *Proceedings of Copper Mountain Conference on Iterative Methods*, Copper Mountain, CO, USA, 1996.

[61] Karplus M. and McCammon J. A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9:646–652, 2002.

[62] Dinner A. R., Sali A., Smith L. J., Dobson C. M., and Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in Biochemical Sciences*, 25:331–339, 2000.

[63] Shakhnovich E. I. Theoretical studies of protein-folding thermodynamics and kinetics. *Current Opinion in Structural Biology*, 7:29–40, 1997.

[64] Dill K. A. and Chan H. S. From Levinthal to pathways to funnels. *Nature Structural Biology*, 4:10–19, 1997.

[65] Isralewitz B., Gao M., and Schulten K. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology*, 11:224–230, 2001.

[66] Chu J. W. and Voth G. A. Coarse-grained free energy functions for studying protein conformational changes: A double-well network model. *Biophysical Journal*, 93:3860–3871, 2007.

[67] Best R. B., Chen Y. G., and Hummer G. Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor. *Structure*, 13:1755–1763, 2005.

[68] Maragakis P. and Karplus M. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. *Journal of Molecular Biology*, 352:807–822, 2005.

[69] Kneller G. R. Inelastic neutron scattering from damped collective vibrations of macromolecules. *Chemical Physics*, 261:1–24, 2000.

[70] Lange O. F. and Grubmuller H. Collective Langevin dynamics of conformational motions in proteins. *Journal of Chemical Physics*, 124:(214903)1–18, 2006.

[71] Erkip A. and Erman B. Dynamics of large-scale fluctuations in native proteins. Analysis based on harmonic inter-residue potentials and random external noise. *Polymer*, 45:641–648, 2004.

[72] Garcia de la Torre J., Carrasco B., and Harding S. E. SOLPRO: Theory and computer program for the prediction of SOLution PROperties of rigid macromolecules and bioparticles. *European Biophysics Journal with Biophysics Letters*, 25:361–372, 1997.

[73] Garcia de la Torre J., Navarro S., Martinez M. C. L., Diaz F. G., and Cascales J. J. L. HYDRO: A computer software for the prediction of hydrodynamic properties of macromolecules. *Biophysical Journal*, 67:530–531, 1994.

[74] Carrasco B., Harding S. E., and Garcia de la Torre J. Bead modeling using HYDRO and SOLPRO of the conformation of multisubunit proteins: Sunflower and rape-seed 11S globulins. *Biophysical Chemistry*, 74:127–133, 1998.

[75] Garcia de la Torre J. and Bloomfield V. A. Hydrodynamic properties of macromolecular complexes. I. Translation. *Biopolymers*, 16:1747–1763, 1977.

[76] Garcia de la Torre J., Echenique G. D., and Ortega A. Improved calculation of rotational diffusion and intrinsic viscosity of bead models for macromolecules and nanoparticles. *Journal of Physical Chemistry B*, 111:955–961, 2007.

[77] Garcia de la Torre J., Huertas M. L., and Carrasco B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophysical Journal*, 78:719–730, 2000.

[78] Garcia de la Torre J. and Bloomfield V. A. Hydrodynamic properties of complex, rigid, biological macromolecules: Theory and applications. *Quarterly Reviews of Biophysics*, 14:81–139, 1981.

[79] Bloomfield V. A., Dalton W. O., and Vanholde K. E. Frictional coefficients of multisubunit structures. I. Theory. *Biopolymers*, 5:135–148, 1967.

[80] Bloomfield V. A. Hydrodynamic studies of structure of biological macromolecules. *Science*, 161:1212–1219, 1968.

[81] Kottalam J. and Case D. A. Langevin modes of macromolecules: Applications to crambin and DNA hexamers. *Biopolymers*, 29:1409–1421, 1990.

[82] Carrasco B., Garcia de la Torre J., and Zipper P. Calculation of hydrodynamic properties of macromolecular bead models with overlapping spheres. *European Biophysics Journal with Biophysics Letters*, 28:510–515, 1999.

[83] Carrasco B. and Garcia de la Torre J. Improved hydrodynamic interaction in macromolecular bead models. *Journal of Chemical Physics*, 111:4817–4826, 1999.

[84] Garcia de la Torre J. and Carrasco B. Intrinsic viscosity and rotational diffusion of bead models for rigid macromolecules and bioparticles. *European Biophysics Journal with Biophysics Letters*, 27:549–557, 1998.

[85] Rugonyi S. and Bathe K. J. On finite element analysis of fluid flows fully coupled with structural interactions. *Computer Modeling in Engineering & Sciences*, 2:195–212, 2001.

[86] Bathe M. and Kamm R. D. A fluid-structure interaction finite element analysis of pulsatile blood flow through a compliant stenotic artery. *Journal of Biomechanical Engineering-Transactions of the ASME*, 121:361–369, 1999.

[87] Zienkiewicz O. C. and Taylor R. L. *The finite element method.* Butterworth-Heinemann, Boston, 2000.

[88] Regueiro R. A. and Ebrahimi D. Implicit dynamic three-dimensional finite element analysis of an inelastic biphasic mixture at finite strain. Part 1: Application to a simple geomaterial. *Computer Methods in Applied Mechanics and Engineering*, 199:2024–2049, 2010.

[89] Sedeh R. S., Ahmadian M. T., and Janabi-Sharifi F. Modeling, simulation, and optimal initiation planning for needle insertion into the liver. *Journal of Biomechanical Engineering-Transactions of the ASME*, 132:(041001)1–11, 2010.

[90] Garcia de la Torre J., Huertas M. L., and Carrasco B. HYDRONMR: Prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *Journal of Magnetic Resonance*, 147:138–146, 2000.

[91] Garcia de la Torre J., Sanchez H. E., Ortega A., Hernandez J. G., Fernandes M. X., Diaz F. G., and Martinez M. C. L. Calculation of the solution properties of flexible macromolecules: Methods and applications. *European Biophysics Journal with Biophysics Letters*, 32:477–486, 2003.

[92] Rotne J. and Prager S. Variational treatment of hydrodynamic interaction in polymers. *Journal of Chemical Physics*, 50:4831–4837, 1969.

[93] Yamakawa H. Transport properties of polymer chains in dilute solution: Hydrodynamic interaction. *Journal of Chemical Physics*, 53:436–443, 1970.

[94] Brune D. and Kim S. Predicting protein diffusion coefficients. *Proceedings of the National Academy of Sciences of the United States of America*, 90:3835–3839, 1993.

[95] Harvey S. C. and Garcia de la Torre J. Coordinate systems for modeling the hydrodynamic resistance and diffusion coefficients of irregularly shaped rigid macromolecules. *Macromolecules*, 13:960–964, 1980.

[96] Lu M. Y. and Ma J. P. The role of shape in determining molecular motions. *Biophysical Journal*, 89:2395–2401, 2005.

[97] Bang D., Tereshko V., Kossiakoff A. A., and Kent S. B. H. Role of a salt bridge in the model protein crambin explored by chemical protein synthesis: X-ray structure of a unique protein analogue, [V15A]crambin-α-carboxamide. *Molecular Biosystems*, 5:750–756, 2009.

[98] Abercrombie M. The Croonian lecture, 1978: The crawling movement of metazoan cells. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 207:129–147, 1980.

[99] Small J. V., Stradal T., Vignal E., and Rottner K. The lamellipodium: Where motility begins. *Trends in Cell Biology*, 12:112–120, 2002.

[100] Pollard T. D. and Borisy G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112:453–465, 2003.

[101] Pantaloni D., Le Clainche C., and Carlier M. F. Mechanism of actin-based motility. *Science*, 292:1502–1506, 2001.

[102] Matsudaira P. Actin crosslinking proteins at the leading edge. *Seminars in Cell Biology*, 5:165–174, 1994.

[103] Faix J. and Rottner K. The making of filopodia. *Current Opinion in Cell Biology*, 18:18–25, 2005.

[104] Vignjevic D., Kojima S., Aratyn Y., Danciu O., Svitkina T., and Borisy G. G. Role of fascin in filopodial protrusion. *Journal of Cell Biology*, 174:863–875, 2006.

[105] Mallavarapu A. and Mitchison T. Regulated actin cytoskeleton assembly at filopodium tips controls their extension and retraction. *Journal of Cell Biology*, 146:1097–1106, 1999.

[106] Romero S., Le Clainche C., Didry D., Egile C., Pantaloni D., and Carlier M. F. Formin is a processive motor that requires profilin to accelerate actin assembly and associated ATP hydrolysis. *Cell*, 119:419–429, 2004.

[107] Goode B. L. and Eck M. J. Mechanism and function of formins in the control of actin assembly. *Annual Review of Biochemistry*, 76:593–627, 2007.

[108] Kureishy N., Sapountzi V., Prag S., Anilkumar N., and Adams J. C. Fascins, and their roles in cell structure and function. *Bioessays*, 24:350–361, 2002.

[109] Adams J. C. Roles of fascin in cell adhesion and motility. *Current Opinion in Cell Biology*, 16:590–596, 2004.

[110] Hashimoto Y., Skacel M., and Adams J. C. Roles of fascin in human carcinoma motility and signaling: Prospects for a novel biomarker? *International Journal of Biochemistry & Cell Biology*, 37:1787–1804, 2005.

[111] Jawhari A. U., Buda A., Jenkins M., Shehzad K., Sarraf C., Noda M., Farthing M. J. G., Pignatelli M., and Adams J. C. Fascin, an actin-bundling protein, modulates colonic epithelial cell invasiveness and differentiation *in vitro*. *American Journal of Pathology*, 162:69–80, 2003.

[112] Vignjevic D., Schoumacher M., Gavert N., Janssen K. P., Jih G., Lae M., Louvard D., Ben-Ze'ev A., and Robine S. Fascin, a novel target of β-Catenin-TCF signaling, is expressed at the invasive front of human colon cancer. *Cancer Research*, 67:6844–6853, 2007.

[113] Peraud A., Mondal S., Hawkins C., Mastronardi M., Bailey K., and Rutka J. T. Expression of fascin, an actin-bundling protein, in astrocytomas of varying grades. *Brain Tumor Pathology*, 20:53–58, 2003.

[114] Hashimoto Y., Parsons M., and Adams J. C. Dual actin-bundling and protein kinase C-binding activities of fascin regulate carcinoma cell migration downstream of Rac and contribute to metastasis. *Molecular Biology of the Cell*, 18:4591–4602, 2007.

[115] Matsudaira P. Modular organization of actin cross-linking proteins. *Trends in Biochemical Sciences*, 16:87–92, 1991.

[116] Revenu C., Athman R., Robine S., and Louvard D. The co-workers of actin filaments: From cell structures to signals. *Nature Reviews Molecular Cell Biology*, 5:635–646, 2004.

[117] Lodish H., Berk A., Zipursky S. L., Matsudaira P., Baltimore D., and Darnell J. *Molecular cell biology.* W.H. Freeman and Company, New York, 1999.

[118] Bartles J. R. Parallel actin bundles and their multiple actin-bundling proteins. *Current Opinion in Cell Biology*, 12:72–78, 2000.

[119] Stossel T. P., Condeelis J., Cooley L., Hartwig J. H., Noegel A., Schleicher M., and Shapiro S. S. Filamins as integrators of cell mechanics and signalling. *Nature Reviews Molecular Cell Biology*, 2:138–145, 2001.

[120] Liu J., Taylor D. W., and Taylor K. A. A 3-D reconstruction of smooth muscle α-actinin by CryoEm reveals two different conformations at the actin-binding region. *Journal of Molecular Biology*, 338:115–125, 2004.

[121] Tseng Y., Fedorov E., McCaffery J. M., Almo S. C., and Wirtz D. Micromechanics and ultrastructure of actin filament networks crosslinked by human fascin: A comparison with α-actinin. *Journal of Molecular Biology*, 310:351–366, 2001.

[122] Klein M. G., Shi W. X., Ramagopal U., Tseng Y., Wirtz D., Kovar D. R., Staiger C. J., and Almo S. C. Structure of the actin crosslinking core of fimbrin. *Structure*, 12:999–1013, 2004.

[123] Carugo K. D., Banuelos S., and Saraste M. Crystal structure of a calponin homology domain. *Nature Structural Biology*, 4:175–179, 1997.

[124] Goldsmith S. C., Pokala M., Shen W. Y., Fedorov A. A., Matsudaira P., and Almo S. C. The structure of an actin-crosslinking domain from human fimbrin. *Nature Structural Biology*, 4:708–712, 1997.

[125] Keep N. H., Norwood F. L. M., Moores C. A., Winder S. J., and Kendrick-Jones J. The 2.0 Å structure of the second calponin homology domain from the actin-binding region of the dystrophin homologue utrophin. *Journal of Molecular Biology*, 285:1257–1264, 1999.

[126] Keep N. H., Winder S. J., Moores C. A., Walke S., Norwood F. L. M., and Kendrick-Jones J. Crystal structure of the actin-binding region of utrophin reveals a head-to-tail dimer. *Structure with Folding & Design*, 7:1539–1546, 1999.

[127] Fucini P., Renner C., Herberhold C., Noegel A. A., and Holak T. A. The repeating segments of the F-actin cross-linking gelation factor (ABP-120) have an immunoglobulin-like fold. *Nature Structural Biology*, 4:223–230, 1997.

[128] McCoy A. J., Fucini P., Noegel A. A., and Stewart M. Structural basis for dimerization of the *Dictyostelium* gelation factor (ABP120) rod. *Nature Structural Biology*, 6:836–841, 1999.

[129] Djinovic-Carugo K., Young P., Gautel M., and Saraste M. Structure of the α-actinin rod: Molecular basis for cross-linking of actin filaments. *Cell*, 98:537–546, 1999.

[130] Namba Y., Ito M., Zu Y. L., Shigesada K., and Maruyama K. Human T-cell L-plastin bundles actin-filaments in a calcium-dependent manner. *Journal of Biochemistry*, 112:503–507, 1992.

[131] Dearruda M. V., Watson S., Lin C. S., Leavitt J., and Matsudaira P. Fimbrin is a homolog of the cytoplasmic phosphoprotein plastin and has domains homologous with calmodulin and actin gelation proteins. *Journal of Cell Biology*, 111:1069–1079, 1990.

[132] Ono S., Yamakita Y., Yamashiro S., Matsudaira P. T., Gnarra J. R., Obinata T., and Matsumura F. Identification of an actin binding region and a protein kinase C phosphorylation site on human fascin. *Journal of Biological Chemistry*, 272:2527–2533, 1997.

[133] Kane R. E. Preparation and purification of polymerized actin from sea urchin egg extracts. *Journal of Cell Biology*, 66:305–315, 1975.

[134] Cant K., Knowles B. A., Mooseker M. S., and Cooley L. *Drosophila* singed, a fascin homolog, is required for actin bundle formation during oogenesis and bristle extension. *Journal of Cell Biology*, 125:369–380, 1994.

[135] Maekawa S., Endo S., and Sakai H. A protein in starfish sperm head which bundles actin-filaments *in vitro*: Purification and characterization. *Journal of Biochemistry*, 92:1959–1972, 1982.

[136] Holthuis J. C. M., Schoonderwoert V. T. G., and Martens G. J. M. A vertebrate homolog of the actin-bundling protein fascin. *Biochimica et Biophysica Acta-Gene Structure and Expression*, 1219:184–188, 1994.

[137] Edwards R. A., Herrerasosa H., Otto J., and Bryan J. Cloning and expression of a murine fascin homolog from mouse brain. *Journal of Biological Chemistry*, 270:10764–10770, 1995.

[138] Duh F. M., Latif F., Weng Y. K., Geil L., Modi W., Stackhouse T., Matsumura F., Duan D. R., Linehan W. M., Lerman M. I., and Gnarra J. R. cDNA cloning and expression of the human homolog of the sea urchin fascin and *Drosophila* singed genes which encodes an actin-bundling protein. *DNA and Cell Biology*, 13:821–827, 1994.

[139] Edwards R. A. and Bryan J. Fascins, a family of actin bundling proteins. *Cell Motility and the Cytoskeleton*, 32:1–9, 1995.

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

[141] Bryan J. and Kane R. E. Separation and interaction of major components of sea urchin actin gel. *Journal of Molecular Biology*, 125:207–224, 1978.

[142] Aratyn Y. S., Schaus T. E., Taylor E. W., and Borisy G. B. Intrinsic dynamic behavior of fascin in filopodia. *Molecular Biology of the Cell*, 18:3928–3940, 2007.

[143] Bathe M., Heussinger C., Claessens M. M. A. E., Bausch A. R., and Frey E. Cytoskeletal bundle mechanics. *Biophysical Journal*, 94:2955–2964, 2008.

[144] Howard J. Molecular mechanics of cells and tissues. *Cellular and Molecular Bioengineering*, 1:24–32, 2008.

[145] Claessens M. M. A. E., Bathe M., Frey E., and Bausch A. R. Actin-binding proteins sensitively mediate F-actin bundle stiffness. *Nature Materials*, 5:748–753, 2006.

[146] Honda M., Takiguchi K., Ishikawa S., and Hotani H. Morphogenesis of liposomes encapsulating actin depends on the type of actin-crosslinking. *Journal of Molecular Biology*, 287:293–300, 1999.

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

[148] Murzin A. G., Lesk A. M., and Chothia C. β-trefoil fold patterns of structure and sequence in the Kunitz inhibitors interleukins-1β and 1α and fibroblast growth factors. *Journal of Molecular Biology*, 223:531–543, 1992.

[149] Kyte J. and Doolittle R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.

[150] Clamp M., Cuff J., Searle S. M., and Barton G. J. The Jalview Java alignment editor. *Bioinformatics*, 20:426–427, 2004.

[151] Kabsch W. and Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[152] Bernstein F. C., Koetzle T. F., Williams G. J. B., Meyer E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T., and Tasumi M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542, 1977.

[153] Murzin A. G., Brenner S. E., Hubbard T., and Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[154] Graves B. J., Hatada M. H., Hendrickson W. A., Miller J. K., Madison V. S., and Satow Y. Structure of interleukin-1α at 2.7 Å resolution. *Biochemistry*, 29:2679–2684, 1990.

[155] Finzel B. C., Clancy L. L., Holland D. R., Muchmore S. W., Watenpaugh K. D., and Einspahr H. M. Crystal structure of recombinant human interleukin-1β at 2.0 Å resolution. *Journal of Molecular Biology*, 209:779–791, 1989.

[156] Ponting C. P. and Russell R. B. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins. *Journal of Molecular Biology*, 302:1041–1047, 2000.

[157] McLachlan A. D. 3-fold structural pattern in the soybean trypsin-inhibitor (Kunitz). *Journal of Molecular Biology*, 133:557–563, 1979.

[158] Rutenber E. and Robertus J. D. Structure of ricin B-chain at 2.5 Å resolution. *Proteins: Structure, Function, and Genetics*, 10:260–269, 1991.

[159] Tahirov T. H., Lu T. H., Liaw Y. C., Chen Y. L., and Lin J. Y. Crystal structure of abrin-a at 2.14 Å. *Journal of Molecular Biology*, 250:354–367, 1995.

[160] Transue T. R., Smith A. K., Mo H. Q., Goldstein I. J., and Saper M. A. Structure of benzyl T-antigen disaccharide bound to *Amaranthus caudatus* agglutinin. *Nature Structural Biology*, 4:779–783, 1997.

[161] Habazettl J., Gondol D., Wiltscheck R., Otlewski J., Schleicher M., and Holak T. A. Structure of hisactophilin is similar to interleukin-1β and fibroblast growth factor. *Nature*, 359:855–858, 1992.

[162] Onesti S., Brick P., and Blow D. M. Crystal-structure of a Kunitz-type trypsin-inhibitor from *Erythrina caffra* seeds. *Journal of Molecular Biology*, 217:153–176, 1991.

[163] Brych S. R., Blaber S. I., Logan T. M., and Blaber M. Structure and stability effects of mutations designed to increase the primary sequence symmetry within the core region of a β-trefoil. *Protein Science*, 10:2587–2599, 2001.

[164] Valdar W. S. J. Scoring residue conservation. *Proteins: Structure, Function, and Genetics*, 48:227–241, 2002.

[165] Cant K. and Cooley L. Single amino acid mutations in *Drosophila* fascin disrupt actin bundling function *in vivo*. *Genetics*, 143:249–258, 1996.

[166] Yamakita Y., Ono S., Matsumura F., and Yamashiro S. Phosphorylation of human fascin inhibits its actin binding and bundling activities. *Journal of Biological Chemistry*, 271:12632–12638, 1996.

[167] Scheel J., Ziegelbauer K., Kupke T., Humbel B. M., Noegel A. A., Gerisch G., and Schleicher M. Hisactophilin, a histidine-rich actin-binding protein from *Dictyostelium discoideum*. *Journal of Biological Chemistry*, 264:2832–2839, 1989.

[168] Hanakam F., Gerisch G., Lotz S., Alt T., and Seelig A. Binding of hisactophilin I and II to lipid membranes is controlled by a pH-dependent myristoyl-histidine switch. *Biochemistry*, 35:11036–11044, 1996.

[169] Grant B. J., McCammon J. A., Caves L. S. D., and Cross R. A. Multivariate analysis of conserved sequence-structure relationships in kinesins: Coupling of the active site and a tubulin-binding sub-domain. *Journal of Molecular Biology*, 368:1231–1248, 2007.

[170] Armon A., Graur D., and Ben-Tal N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology*, 307:447–463, 2001.

[171] Goldenberg O., Erez E., Nimrod G., and Ben-Tal N. The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Research*, 37:D323–D327, 2009.

[172] Mihalek I., Res I., and Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology*, 336:1265–1282, 2004.

[173] Mayrose I., Graur D., Ben-Tal N., and Pupko T. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Molecular Biology and Evolution*, 21:1781–1791, 2004.

[174] Lichtarge O., Bourne H. R., and Cohen F. E. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257:342–358, 1996.

[175] Schneider R., de Daruvar A., and Sander C. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Research*, 25:226–230, 1997.

[176] Altschul S. F., Madden T. L., Schaffer A. A., Zhang J. H., Zhang Z., Miller W., and Lipman D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[177] Thompson J. D., Higgins D. G., and Gibson T. J. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[178] Russell R. B. and Barton G. J. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins: Structure, Function, and Genetics*, 14:309–323, 1992.

[179] Caffrey D. R., Somaroo S., Hughes J. D., Mintseris J., and Huang E. S. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13:190–202, 2004.

[180] Miller S., Lesk A. M., Janin J., and Chothia C. The accessible surface area and stability of oligomeric proteins. *Nature*, 328:834–836, 1987.

[181] Humphrey W., Dalke A., and Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[182] Kumar S. and Nussinov R. Salt bridge stability in monomeric proteins. *Journal of Molecular Biology*, 293:1241–1255, 1999.

[183] Brooks B. and Karplus M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 80:6571–6575, 1983.

[184] Go N., Noguti T., and Nishikawa T. Dynamics of a small globular protein in terms of low frequency vibrational modes. *Proceedings of the National Academy of Sciences of the United States of America*, 80:3696–3700, 1983.

[185] Guex N. and Peitsch M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18:2714–2723, 1997.

[186] Lange O. F. and Grubmuller H. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 62:1053–1061, 2006.

[187] Ichiye T. and Karplus M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Genetics*, 11:205–217, 1991.

[188] Needleman S. B. and Wunsch C. D. A general method applicable to search for similarities in amino acid sequence of 2 proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[189] Rice P., Longden I., and Bleasby A. EMBOSS: The European molecular biology open software suite. *Trends in Genetics*, 16:276–277, 2000.

[190] Hayward S. and Lee R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *Journal of Molecular Graphics & Modelling*, 21:181–183, 2002.

[191] Landau M., Mayrose I., Rosenberg Y., Glaser F., Martz E., Pupko T., and Ben-Tal N. ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research*, 33:W299–W302, 2005.

[192] Morgan D. H., Kristensen D. M., Mittelman D., and Lichtarge O. ET viewer: An application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, 22:2049–2050, 2006.