SNAPSHOTS AND SPRINGS:
ANALYZING AND REPRODUCING THE MOTIONS OF MOLECULES

by

David Morris
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Master of Science
Computer Science

Committee:

_____	Dr. Amarda Shehu, Thesis Director
_____	Dr. Zoran Duric, Committee Member
_____	Dr. Kevin Molloy, Committee Member
_____	Dr. Sanjeev Setia, Chairman, Department of Computer Science
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date: _____	Fall Semester 2017 George Mason University Fairfax, VA

Snapshots and Springs: Analyzing and Reproducing the Motions of Molecules

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

David Morris
Bachelor of Science
George Mason University, 2015

Director: Dr. Amarda Shehu, Professor
Department of Computer Science

Fall Semester 2017
George Mason University
Fairfax, VA

Dedication

I dedicate this thesis to norepenepherine and its friends.

Acknowledgments

I would like to thank the following people who made this possible: Prof. Amarda Shehu, my advisor, Prof. Erion Plaku, who has contributed to the research effort associated with this thesis, Dr. Tatiana Maximova, who has patiently helped me understand and debug software, and Rudy Clausen, who wrote the foundations this research was written on, and explained it years later.

Table of Contents

List of Figures

Abstract

SNAPSHOTS AND SPRINGS: ANALYZING AND REPRODUCING THE MOTIONS OF MOLECULES

David Morris

George Mason University, 2017

Thesis Director: Dr. Amarda Shehu

Nearly all cellular processes involve proteins structurally rearranging to accommodate molecular partners. The energy landscape underscores the inherent nature of proteins as dynamic molecules interconverting between structures with varying energies. Reconstructing a proteins energy landscape holds the key to characterizing the structural dynamics and its regulation of protein function. In practice, the disparate spatio-temporal scales spanned by the slow dynamics challenge wet and dry laboratories. The growing number of deposited structures for proteins central to human biology presents an opportunity to infer the relevant dynamics. Recent computational efforts using extrinsic modes of motion as variables have successfully reconstructed detailed energy landscapes of several medium-size proteins. Here we investigate the extent to which one can reconstruct the energy landscape of a protein in the absence of sufficient, wet-laboratory structural data. We do so by integrating intrinsic modes of motion extracted off a single structure in a stochastic optimization framework that supports the plug-and-play of different variable selection strategies. We demonstrate that, while knowledge of more wet-laboratory structures yields better-reconstructed landscapes, precious information can be obtained even when one structural model is available. The presented work opens up interesting venues of research on structure-based inference of

dynamics. Added data with a second protein suggests that the findings are not specific to the molecules analyzed.

Chapter 1: Background and Related Work

1.1 Structural Plasticity of Molecules

Proteins go through both small fast movements, and slow large movements. Small fast movements reflect thermal fluctuations which all molecules will go through, while slow large movements represent structural changes which are often biologically relevant. Physics-based simulations, called Molecular Dynamics (or MD), are good at showing the small fast motions. However, the times required to simulate structural rearrangements with MD often makes it impractical. Wet laboratories (laboratories which deal with actual samples of the proteins they study) are capable of determining long-lived structures, which represent the functional results of the structural rearrangements, but they do not determine many structures, so they do not report detailed information about the space of structures a given protein can take.

My thesis focuses on structural plasticity of proteins. In between the MD generated fast movements, and the frozen still snapshots supplied by wet labs, there is a need for detailed reconstruction of the structure space *in silico*. To do that, the structure space can be thought of as an energy landscape: the energy estimate of a structure is the "height," and the ways you can vary the shape give the (numerous) other dimensions. Ridges in this energy landscape represent boundaries that must be overcome for a protein to change shape, and pits represent stable, long-lived structures. In this way, the energy landscape organizes the vast space of structures available to a protein by potential energy.

1.2 Obstacles to Detailed Reconstruction

Representations of shapes that a given protein can take are highly dimensional. One possible representation is recording the Cartesian coordinates of the atoms in the molecule. To save space, all but representative (C alpha) atoms can be omitted. There is only one C alpha atom per amino acid, so the dimensionality of this model is still three times the length of the protein. Another approach is to save bond angles. While bonds are spaced at fixed positions from the alpha carbon, the bonds can twist. This twisting, along with twisting of

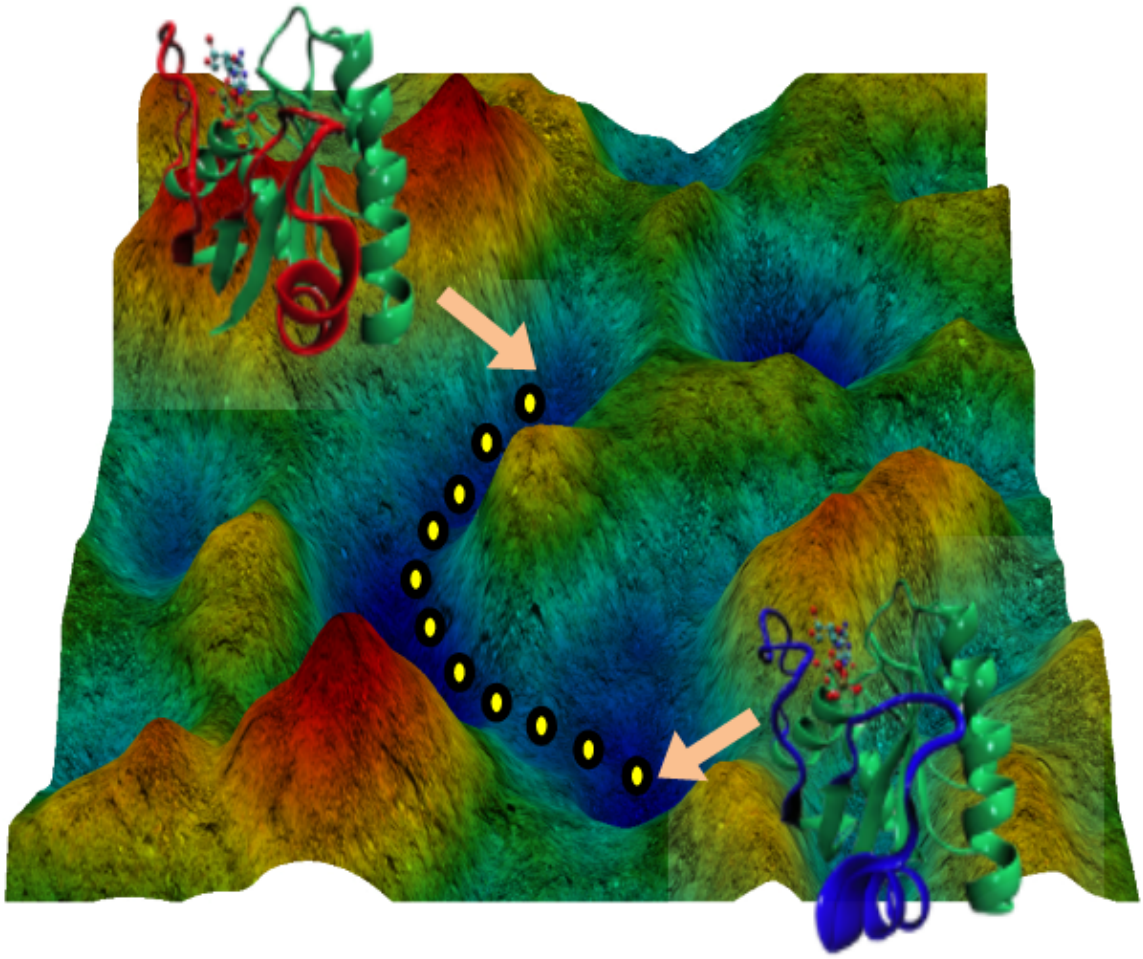


Figure 1.1: A simple illustration of an energy landscape.

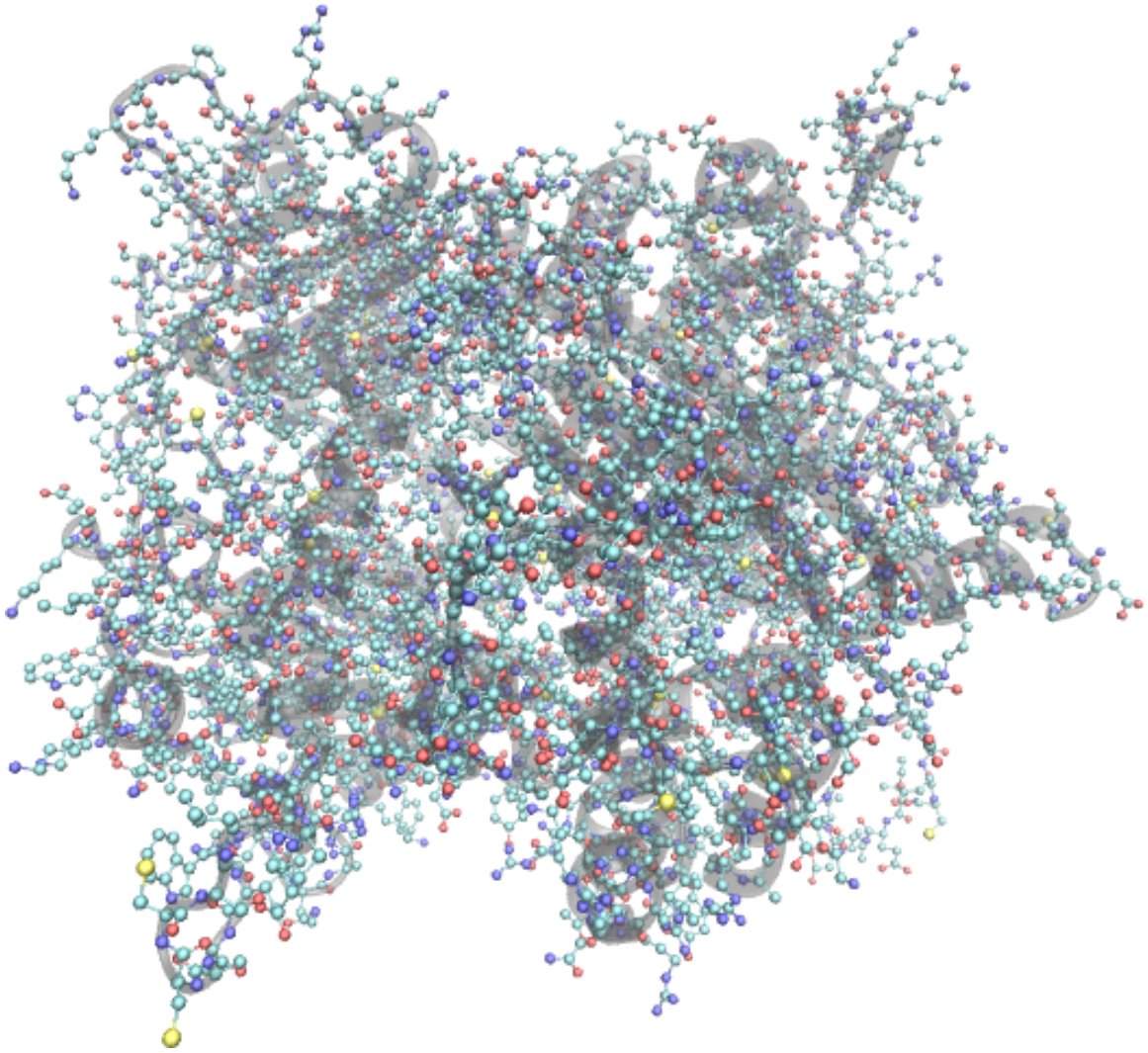


Figure 1.2: An example protein showing its atoms and bonds. This is to illustrate the complexity of representing a protein by either its atom coordinates or bond locations.

bonds connecting to and inside the part of the amino acid which varies based on the amino acid (the side chain), describes the movements which proteins go through. The side chain angles can be omitted, since they can be easily re-estimated.

However, for robotics applications, we would prefer to have still fewer degrees of freedom. This thesis explores sampling protein states using one way of using fewer degrees of freedom to represent the movements of a protein, and compares it to an existing method.

It should also be noted that one limitation to our understanding of the energy landscape is that our energy estimation functions are imperfect, however, that is a problem for physicists and computational chemists.

1.3 Selecting Variables to Represent Proteins

The way to reduce the dimensionality of a representation of a protein is to find variables which represent the deformations which are relevant to the functioning of the protein. If appropriate variables are parameterized, the energy landscape can be reconstructed in detail via sampling. Most existing techniques for reconstructing energy landscapes are not time-effective, since their sampling is not directed along functionally relevant movements. A recent insight from Clausen et al. and Maximova et al. in Prof. Amarda Shehu's lab has been to extract relevant variables as statistical descriptions of variance in known structures of a protein. The statistical analysis of a protein's known states gives useful information about the flexibility of the protein.

1.4 The SoPriM Sampler

SoPriM is an existing sampling strategy by Maximova et al. designed for proteins, inspired by robotics. SoPriM extracts variables relevant for slow motions of a protein by performing Principal Component Analysis of many known structures. A number of Principal Components are then chosen to act as a subspace to sample in by iterating selection and variation (and transformation) operators. The selection operator selects a known good structure,

represented in the variable space. It is designed to choose non-redundant local minimums of energy. The variation operator varies the selected structure to obtain a new one. It is necessary to vary instead of creating a new sample from scratch because most protein structures are not plausible (and so sampling from scratch would waste time). The transformation operator accepts a structure described with the extracted variables, and generates the coordinates of the protein that would correspond to. It is necessary to do this both to evaluate the energy of the structure, which is used by the selection operator, and to replace the structure with a more physically plausible one which is similar (a process called minimization).

Chapter 2: Conference Paper Produced by Work

2.1 Introduction

Wet and dry laboratories have demonstrated that proteins switch between three-dimensional (3d) structures to accommodate molecular partners in different cellular processes [1]. In particular, the structural rearrangements that a protein molecule undergoes under physiological conditions (at equilibrium) are both fast (and small) and slow (and large). Slow rearrangements occur on the nanosecond-to-millisecond time scale and allow a protein to access different functionally-relevant substates (often several Å apart). In the energy landscape that organizes the vast space of structures available to a protein by potential energies, slow structural rearrangements constitute paths that connect energy basins corresponding to different substates [2].

Characterizing the equilibrium structural dynamics of a protein is key to elucidating how structure modulates function [3]. Due to the diffusion time scales involved, it is not possible to probe all stable and semi-stable structural states or to reveal the detailed structure-by-structure rearrangements a protein uses to diffuse among such states in the wet laboratory. In principle, these issues can be addressed via a detailed reconstruction of the energy landscape *in silico* [2]. In practice, due to the disparate spatio-temporal scales involved, neither wet nor dry laboratories can reconstruct the energy landscape of any protein of interest [4]. Nonetheless, the challenges continue to spur computational research [3].

Two main challenges have been recognized *in-silico*. The first relates to the high dimensionality of the search space, which limits sampling capability. The second relates to inaccuracies in molecular mechanics-based energy functions that evaluate atomic interactions in a structure and is known as the local minima (or ruggedness) issue.

While it remains challenging to reconstruct the energy landscape of a medium-size protein ($100-300$ amino acids long) that utilizes slow structural rearrangements to access different functionally-relevant substates, progress has been made. This has been due to the realization that limited sampling capability is principally a variable selection issue [5].

Recent efforts have demonstrated that insight on variables underlying the slow dynamics is key to defining a low-dimensional space amenable to exploration and effective variation operators obtaining samples (new structures) under the umbrella of stochastic optimization [6–12]. These algorithms leverage the growing number of structures deposited in public databases for healthy/wildtype (WT) and diseased/mutated forms of a protein. They extract the *extrinsic modes of motion* via Principal Component Analysis (PCA) of atomic displacements compiled from known structures of a protein. The extracted principal components (PCs) are utilized as variables/axes of the variable space then explored via iterative applications of selection (to select an existing sample) and variation (to obtain a new one) operators [10].

Proteins at the center of proteinopathies (such as many human cancers and neurological disorders), are avidly studied by many wet laboratories that report on stable and semi-stable states of healthy and diseased variants. The growing number of structures on such proteins has presented an opportunity to make inferences on equilibrium structural dynamics that recent successful efforts have leveraged to define relevant, low-dimensional variable spaces amenable to exploration. While this line of work has revealed precious insights on known and novel functionally-relevant states, the rearrangements between states, and the mechanisms via which mutations alter dynamics to cause dysfunction [6, 8, 12, 13], the demand on sufficient prior structure data to define relevant variables limits broader applicability to proteins that are not as well studied in wet laboratories.

The key issue addressed in this paper is whether it is possible and to what extent one can reconstruct the energy landscape of a protein in the absence of sufficient, experimentally-available structural data. A complementary line of work in characterizing the slow dynamics presents an opportunity. Since the late 90s, normal mode (NM) analysis (NMA) has been established as an expedient technique via which to extract the *intrinsic modes of motion* (NMs) from a single structure [14, 15]. The low-frequency eigenvectors (slow modes) have been utilized to connect two structures (e.g., open/unbound and closed/bound) of a protein in algorithms seeking to elucidate a specific structural rearrangement between two known

structures [16–19].

Here, we assess the extent to which the slow modes allow to reconstruct the energy landscape of a protein (effectively, obtain many structures out of one). We utilize a stochastic optimization framework, SoPriM [6], which allows plugging different variables of interest. While in prior work we have assessed the effectiveness of PCs as variables, here we assess the employment of the slow (NMA-extracted) modes. We refer to the former algorithmic realization as SoPriM-PCA [6] and to the latter one, described and evaluated in this paper, as SoPriM-NMA. The objective is to assess in a controlled environment (on a protein that has been well studied by us and others) the landscape reconstructed when exploiting the dynamics encoded in only one structure (of the protein under investigation) versus the landscape that can be reconstructed when exploiting the dynamics encoded in a set of structures (caught for various forms of the protein under investigation).

We describe the proposed SoPriM-NMA in Section 2.2, after summarizing the main algorithmic components of SoPriM (and SoPriM-PCA). We present a detailed evaluation in Section 2.3 and conclude the paper with a summary and discussion of future directions of work in Section 2.4.

2.2 Methods

2.2.1 SoPriM

The input to SoPriM is a set Ω_S of known structures of a protein and a matrix $U_{3k\times 3k}$ encoding the variable space (each column encodes an axis, and k encodes the number of amino acids in the protein under investigation); Ω_S contains many structures, as in SoPriM-PCA, or a single structure, as in SoPriM-NMA. The structure(s) in Ω_S are projected onto the employed axes to obtain an initial population Ω_C of conformations, with each conformation being a point in the selected variable space. Ω_C initializes the desired population C of conformations. The SoPriM framework adds onto C via iterative application of a selection and a variation operator for a user-defined number of iterations (with iterations corresponding to the desired size of C).

At every iteration, the selection operator selects a conformation from C. The selection penalizes selecting conformations from over-populated or high-energy regions per a defined weighting function (over conformations and cells of a grid over two selected variables, as detailed in Ref. [6]). The selected conformation is then subjected to a variation operator that utilizes the variable axes (described below for the two different realizations SoPriM-PCA and SoPriM-NMA). Prior to adding a conformation resulting from an application of the variation operator to C, the conformation is transformed into an all-atom structure. The transformation occurs over various scales. First, the conformation is converted to a CA trace (CA atoms), then to a backbone trace, then side chains are packed, and finally the resulting all-atom structure is minimized via the sander protocol with the Amber ff14SB force field. Details of this transformation protocol are available in Ref. [6]. The resulting structure is projected back into the variable axes to obtain the improved conformation for addition to the growing population C.

2.2.2 SoPriM-PCA

The selected variables are PCs; $U_{3k\times 3k}$ is the set of eigenvectors obtained from a matrix A prepared as follows: Structures for the sequence under investigation (and variants no more than 3 mutations different) are collected from the PDB. The CA atoms are extracted from the n structures and stored in a matrix $A_{3k\times n}$ (we refer to a chain of CA atoms as a trace), and an average trace is computed. A is centered (by subtracting the average trace from each column of A) so that it encodes internal structural fluctuations rather than rigid-body motions in 3d. A singular value decomposition yields $1/\sqrt{n-1}\cdot A = U\cdot\Sigma\cdot V^T$. While further details can be found in Ref. [11], in summary, $U_{:,i}$ contains the coordinates of PC_i, and the singular values Σ_{ii} are square roots of eigenvalues e_i that measure the variance of the data (traces) when projected onto PC_i. The order of the PCs in U is from high-to-low corresponding eigenvalues. A cumulative variance analysis allows selecting the top m PCs that cumulatively capture a threshold of structural variance (typically, 80%) as coordinate/variable axes. For many proteins with multiple functional states, even the top two PCs capture more than 50% of the variance.

Given C as a point in the space of the top m PCs, the variation operator computes a new conformation $C_{new} = C + g$, where $g = g_1 \ldots g_m$ is a "global motion vector" that specifies displacements along each PC; $g_i = s_i \cdot \delta_i$, where s_i is sampled uniformly at random in $\{-1, +1\}$, δ_1 is a user-defined parameter, and $\delta_i = \delta_1 \cdot e_i/e_1$ (for each $i > 1$) to ensure that displacements are proportionate with the variations captured by each PC.

2.2.3 SoPriM-NMA

In this setting, the NMs extracted from an NMA off a single structure are selected as variables. The reader is directed to seminal work in [14] for background and foundations of NMA in statistical mechanics. In practice, we employ the utilities in Bio3D [20] to extract the matrix $U_{3k\times 3k}$ of the NMs off a single structure. Unlike PCA, the first 6 NMs capture rigid-body motions, so we discard them. From now on, NM7 through NM_{3k-6} are of interest

for variable selection, and they are ordered by their associated frequencies (low to high, with low corresponding to slow modes). Let us renumber and refer to these frequency-ordered NMs of interest as NM_1 through NM_d ($d = 3k - 6$). Prior to plugging them into SoPriM to obtain SoPriM-NMA, two questions need answering: (i) what $m << d$ to select as axes of the space; and (ii) how to utilize the selected m NMs to compute the global motion vector used by the variation operator. The first can be addressed by balancing between low dimensionality of the variable space and accurate reconstruction of known structures.

Suppose that many structures are available for a protein of interest (as is the case for an enzyme employed here for this analysis), even though the NMs are extracted off a single selected structure. The CA traces of all structures are projected onto NM_1, ..., NM_d to obtain a corresponding d-dimensional point/conformation C for each trace. For a given $i \in [d]$, for each of the conformations C, we can drop the other $d - i$ coordinates (thus arbitrarily reducing the dimensionality of the space) to obtain a "reduced" conformation C_i. For instance, if $i = 1$, C_1 contains only 1 coordinate (along NM_1 in a 1-dim variable space); if $i = d$, all coordinates are retained. The transformation operation described above then allows reconstructing a CA trace from a conformation C_i, and the least root-mean-squared-deviation (lrmsd) [21] between the reconstructed and the original trace can be recorded (for each of the structures). The mean and median lrmsds can then be reported for a given value of i, as Fig. 2.1 does over known structures of the H-Ras enzyme, as i varies from 1 to d on the x axis.

Fig. 2.1 shows that, as expected, the more NMs used, the lower the reconstruction error. This analysis also shows that the reconstruction error is less than 0.6Å even when less than 10 NMs are employed as variable axes, supporting studies showing that relatively few, low-frequency NMs can identify the direction of global motions required to achieve state-to-state transitions [18]. Such an analysis can be employed to select $m << d$ NMs as variables if many structures of a protein are available. When this is not the case, there is no general non-parametric rule for an optimal value for m besides the rule of thumb to keep the dimensionality low. In Section 2.3 we analyze in greater detail the relationship

between NMs and PCs, focusing on a well-studied protein, H-Ras, and select m to be the same value whether employing PCs or NMs as variable axes.

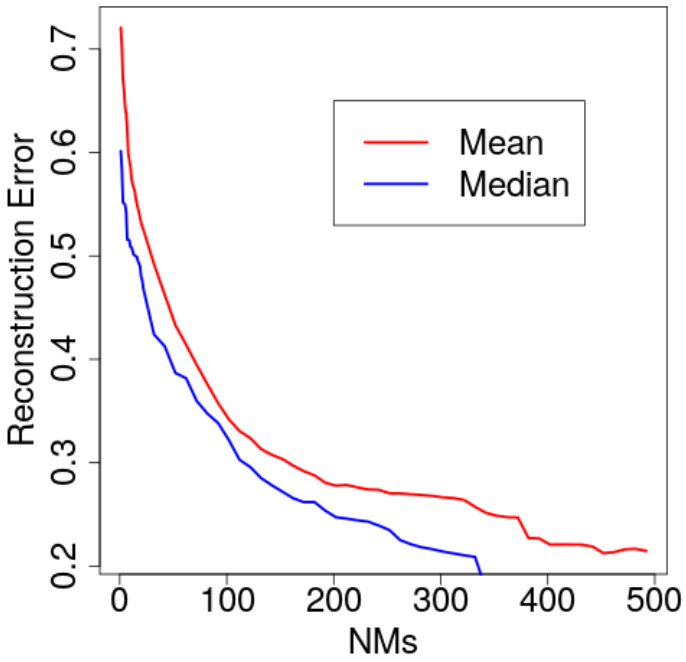


Figure 2.1: Mean and median lrmsds estimate the reconstruction error when using $i \leq d$ frequency-ordered NMs to recover CA traces of experimentally-known structures.

Global Motion Vector

The global motion vector g is adapted from Ref. [18]: $g = \delta \cdot \sqrt{2/m} \cdot \sum_{i=1}^{m} \frac{s_i \mathrm{NM_i}}{f_i}$, where δ is a user-defined parameter, s_i is a sign sampled uniformly at random in $\{-1, +1\}$ for each NM_i (so that displacements can be defined in the positive or negative direction along the principal axis of motion represented by an NM), and the scaling $\frac{1}{f_i}$ is so as to achieve a greater magnitude of displacement along lower-frequency NMs than along the higher-frequency modes under the same fixed energy (with frequencies corresponding to singular values of associated eigenvectors/NMs). This equation is based on the principle that displacements in the direction of each NM must produce a constant-valued energy when averaged over

the resulting path, and the reader is directed to Ref. [18] for the underlying theory and derivation.

2.2.4 Implementation Details and Experimental Setup

A detailed analysis is conducted on a well-studied, 166-amino acid long enzyme, H-Ras, that populates various states. SoPriM-PCA utilizes 87 structures collected from the PDB for H-Ras WT and other variants. Three production runs are used to compute $45,000$ structures ($\delta \in \{1, 2, 3\}$). A detailed analysis in prior work shows these step sizes to balance between exploration and exploitation. SoPriM-NMA utilizes the NMs extracted from a single structure, instead. Two setups are considered, NMs extracted from the Amber ff14SB-minimized structure corresponding to H-Ras PDB entry 1QRA (a representative of the H-Ras GDP-bound/off state) and to H-Ras PDB entry 4Q21 (a representative of the GTP-bound/on state). Three production runs are employed under each setting to compute $45,000$ structures (using $\delta \in \{0.25, 0.5, 0.75\}$; an analysis on optimal values of δ is not shown here in the interest of space).

2.3 Results

2.3.1 Comparison of Intrinsic to Extrinsic Motions

PCs are compared directly to NMs via dot-products $NM_i \cdot PC_j$ with i, j in $[3k]$ (k being the number of CA atoms). Absolute values are used to color-code a heatmap. Fig. 2.2 is limited to the top 100 PCs and top 100 NMs for ease of visualization; the PCs are ordered by their eigenvalues (high to low), and the NMs are ordered by their frequencies (low to high). The highest-similarity pairs are found among the top ten PCs and top ten NMs, as zoomed in on the right of Fig. 2.2. Two setups are considered, on NMs derived from the (Amber ff14SB-minimized) off state representative structure (PDB id 1QRA) and on NMS derived from the (Amber ff14SB-minimized) on state representative structure (PDB id 2Q21). Each of the top ten PCs, which capture more than 80% of the structural variance among known structures of H-Ras, is covered by at least one of the top ten NMs in each setting. In particular, PC1 and PC2 (which cumulatively capture more than 50% of the variance) are best captured by 1QRA-derived NM8 and NM7, respectively, and 4Q21-derived NM4 and NM1, respectively. These results support studies showing that highest-variance PCs correspond better to low-frequency NMs derived from closed (such as 4Q21) than open structures (1QRA).

The PCs-NMs correspondence is further visualized by drawing structures obtained along a selected axis (PC or NM). Instead of adding all the (properly-scaled) PCs or NMs in the global motion vector, only one PC or NM is selected over and over to produce 10 conformations at $\delta \cdot i$ units away along the selected axis, with $i \in [10]$ and using either the Amber ff14SB-minimized structure corresponding to PDB entry 1QRA or that to PDB entry 4Q21 as the selected start structure. The transformation summarized in Section 2.2 is utilized to obtain all-atom structures. The top panel of Fig. 2.3 shows 10 structures obtained by accumulating structural variations captured by PC1 or PC2 starting from 1QRA or 4Q21. The bottom panel shows the structures obtained when the variation is over the two NMs that best agree with PC1 and PC2 (1QRA-derived NM8 and 7, respectively,

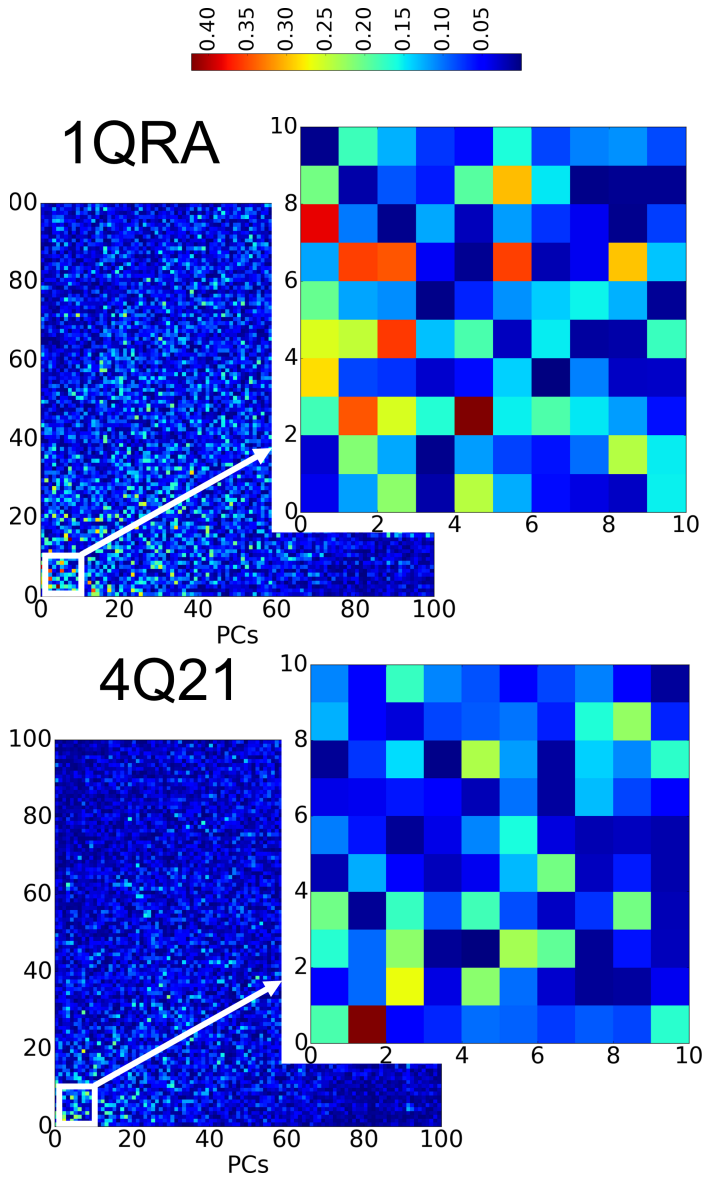


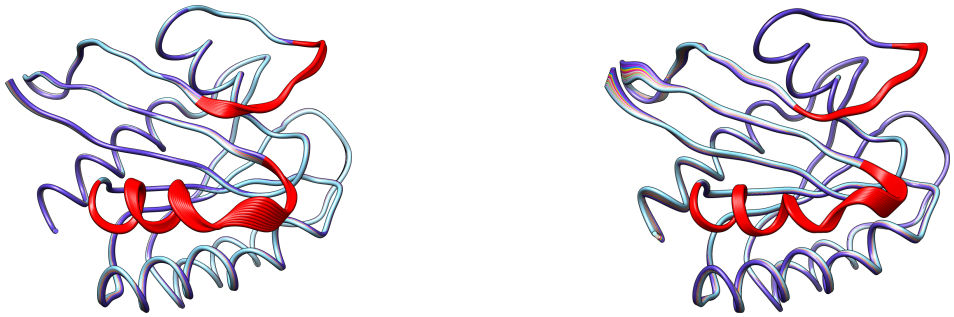
Figure 2.2: Dot products are computed and shown color-coded between each of the top 100 PCs and 100 NMs (derived from 1QRA on the left and from 4Q21 on the right). The heatmap corresponding to the top 10 PCs and NMs is zoomed in.

and 4Q21-derived NM4 and NM1, respectively). Fig. 2.3 visually supports the comparison related in Fig. 2.2 that these NMs encode displacements in the switch I and II functional regions (highlighted in red) of H-Ras.

These results suggest that one structure encodes similar information on the slow dynamics to what can be extracted when one has access to many known structures. While the top ten NMs contain the slow dynamics of interest, the first few (slowest) modes are more likely to capture this dynamics if extracted off a closed structure. In Fig. 2.4 we show that the NMs also encode the organization of the underlying, unknown energy landscape. In the interest of space, we restrict this analysis to comparing projections of PDB-obtained structures of H-Ras on PC1 and PC2 to projections on 1QRA-derived NM8 and NM7 (better results are obtained when using the 4Q21-derived slowest NMs). The annotations in Fig. 2.4 synthesize wet- and dry-laboratory knowledge on H-Ras states and substates. Altogether, the NM-based projections preserve the separation of the On and Off states, together with the co-localization of known structures corresponding to the T (tardy) versus the R+T* (reactive and hydrolyzed tardy) substates. Deformations are present; e.g., the R and T* states are not separable by NM8 and NM7, and smaller substates are also penetrated by projections of structures of other substates. These results support the premise that the NMs can serve as variable axes along which to "fill in" the unknown energy landscape. Based on the constraint to keep the dimensionality low, the rest of the analysis is on structures obtained from SoPriM-NMA with the top ten (m = 10) NMs are variable axes.

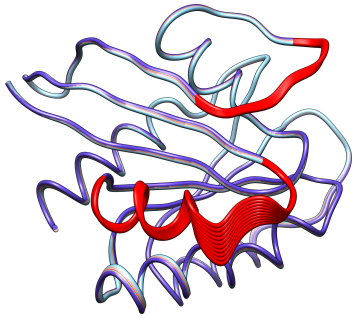
2.3.2 Comparison of Ensembles Generated with SoPriM-PCA and SoPriM-NMA

Below we relate results obtained when using the 1QRA-derived NMs but seeding the initial population of structures with all known PDB structures (threaded onto the WT and Amber ff14SB minimized); many other settings are analyzed but not shown here in the interest of space (such as using only the structure from which NMs are derived in the initial population, using 4Q21-derived NMs, etc.). Fig. 2.5 shows the computed 2D energy landscape by

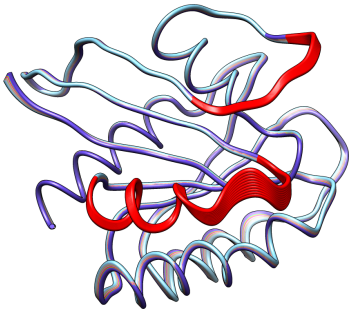


from 1QRA along PC1

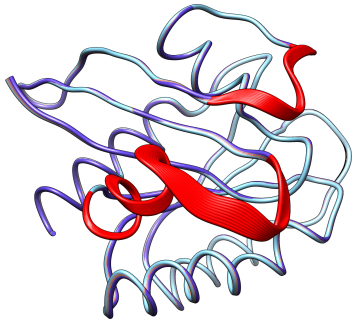
from 1QRA along NM8



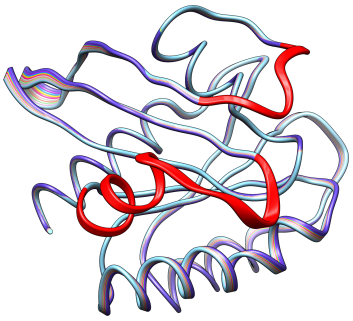
from 1QRA along PC2



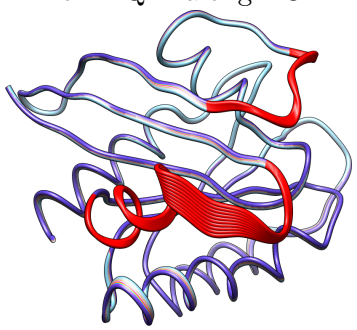
from 1QRA along NM7



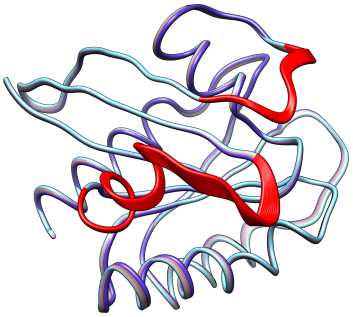
from 4Q21 along PC1



from 4Q21 along NM4



from 4Q21 along PC2



from 4Q21 along NM1

Figure 2.3: A few structures along selected principal components and normal modes are drawn, starting from either the GDP-bound (off) representative structure (PDB id 1QRA) or the GTP-bound (on) representative structure (PDB id 4Q21), and superimposed over the start structures. The switch I and II functional regions are in red.

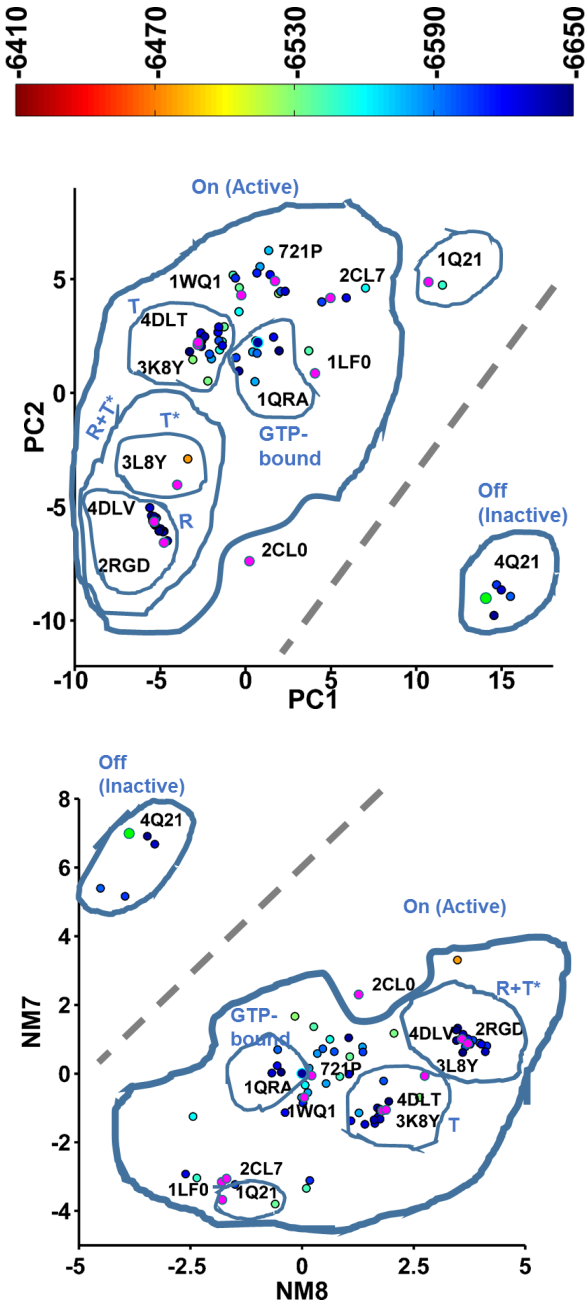
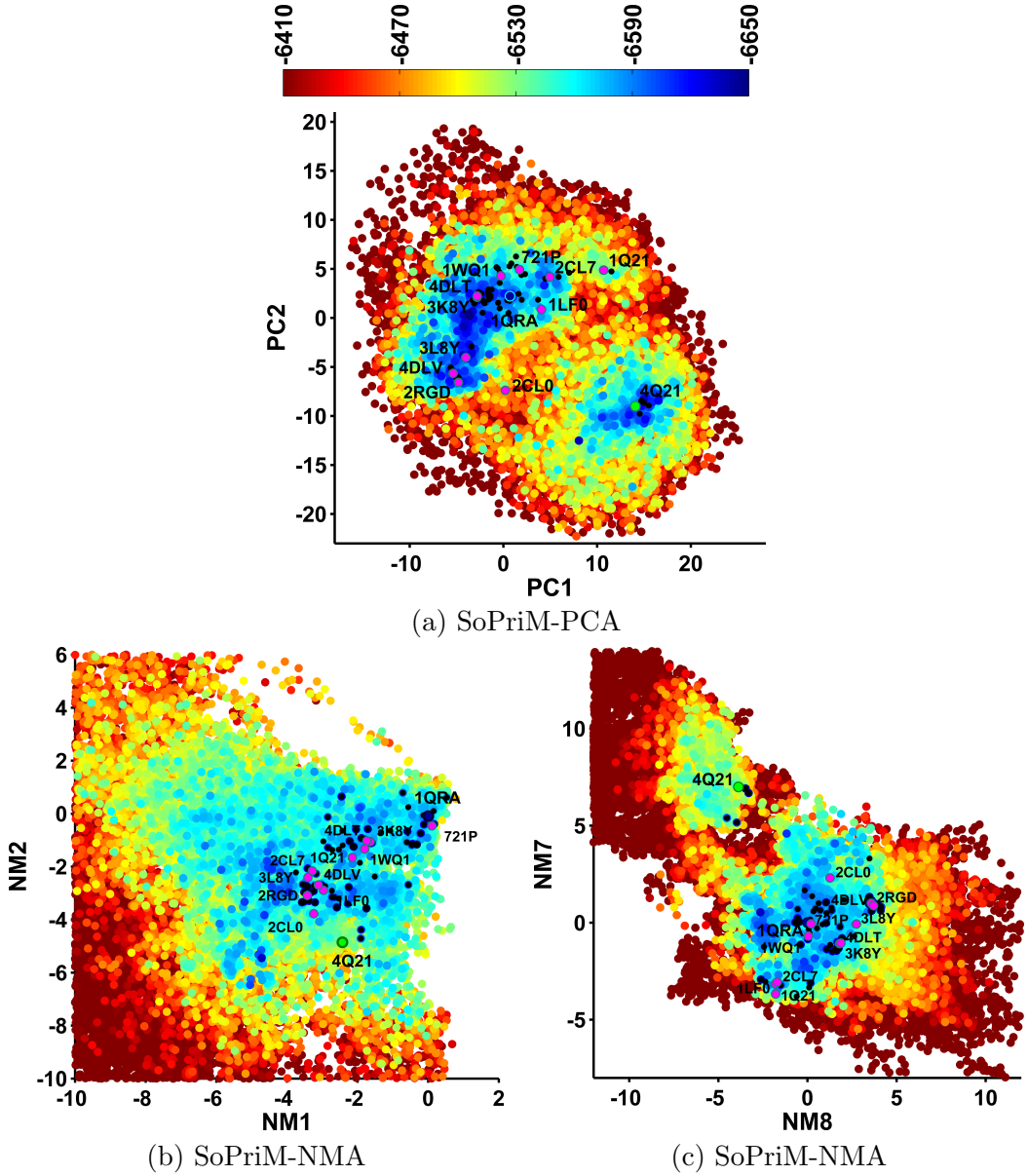


Figure 2.4: WT-minimized, known structures of H-Ras are projected onto PC1 and PC2 (left), and 1QRA-derived NM8 and NM7 (right). The projections are color-coded based on all-atom Amber ff14SB energies. PDB ids are shown alongside projections of selected structures. Annotations indicate known states and substates relevant for function.

drawing 2D projections of computed structures onto the top two axes and color-coding the projections by the Amber ff14SB energies of the corresponding structures. Fig. 2.5(a) shows the PC1-PC2 landscape and serves as the baseline, showing the ability of SoPriM-PCA to reproduce the main On and Off states and even substates under-probed in the wet laboratory (as related in prior work). Fig. 2.5(a) and 2.5(b) show the NM1-NM2 and NM8-NM7 landscapes, respectively, obtained when projecting SoPriM-NMA computed structures. The main On and Off states are captured well, but the smaller substates are not as well populated as when using the top ten PCs as variables. Better results are obtained when using the 4Q21-derived NMs (data not shown here). When the initial population is seeded to contain only one structure, the exploration capability of SoPriM-NMA suffers (data not shown), as more time is needed to expand to other regions of the structure space.



(a) SoPriM-PCA

(b) SoPriM-NMA

(c) SoPriM-NMA

Figure 2.5: Structures obtained with SoPriM-PCA are projected onto PC1 and PC2 in (a), and those obtained with SoPriM-NMA are projected onto NM1 and NM2 in (b) and NM8 and NM7 in (c). The projections are color-coded based on Amber ff14SB energies. Projections of known structures are also shown, and selected ones are annotated with their PDB ids to allow visualization of the main functional states as captured by each algorithm.

2.4 Conclusion

This study shows that much information can be inferred on the slow dynamics and even the energy landscape even when only one structure is available for a protein under investigation. The SoPriM framework allows leveraging the NMs extracted off a single structure to build a sample-based representation of the underlying energy landscape that reveals functional states and substates and separating barriers. While the availability of more wet-laboratory structural data is desired, the study presented here opens further lines of enquiry onto leveraging structures of a protein or members in its superfamily to compute energy landscapes.

Chapter 3: Initial Analysis of Calmodulin

A protein other than H-Ras, Calmodulin, is a good candidate for further testing SoPriM-NMA. Calmodulin (CaM) is a protein with three states, which shows more deformation than H-Ras. Preliminary heatmap comparisons of CaM PCs to CaM NMs suggests that SoPriM-NMA will likely recover the CaM energy landscape well.

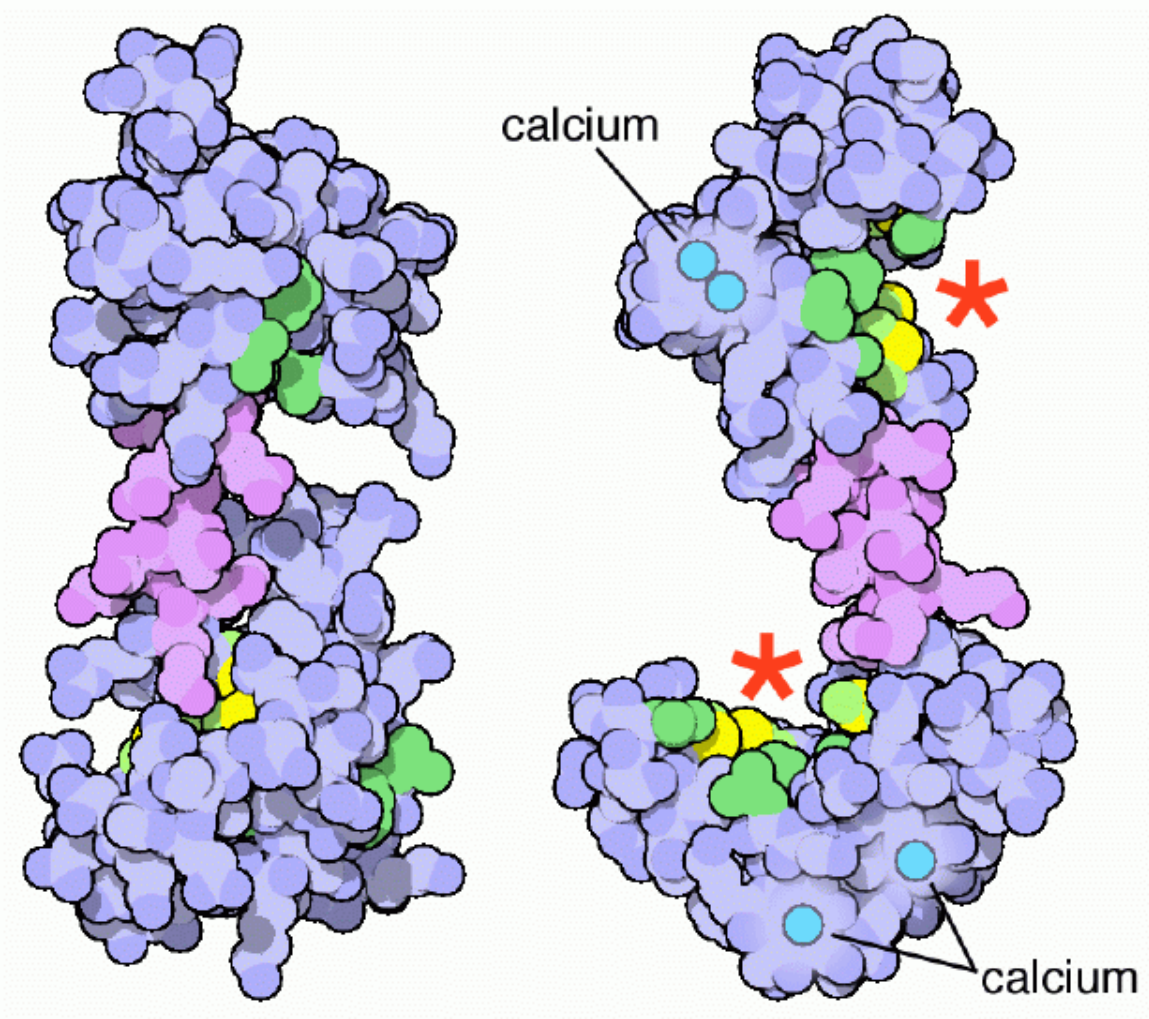


Figure 3.1: Two states of CaM. Courtesy of RCSB.

The heatmaps of cosine similarities of NMs and PCs of H-Ras presented earlier showed greater similarities in the lower left corner. That suggests that PCA and NMA agree on the nature of the more relevant motions of the protein H-Ras.

The cosine similarities presented now suggest that PCA and NMA agree on the important movements of CaM. This preliminary analysis is encouraging for a follow-up test of SoPriM-NMA on CaM.

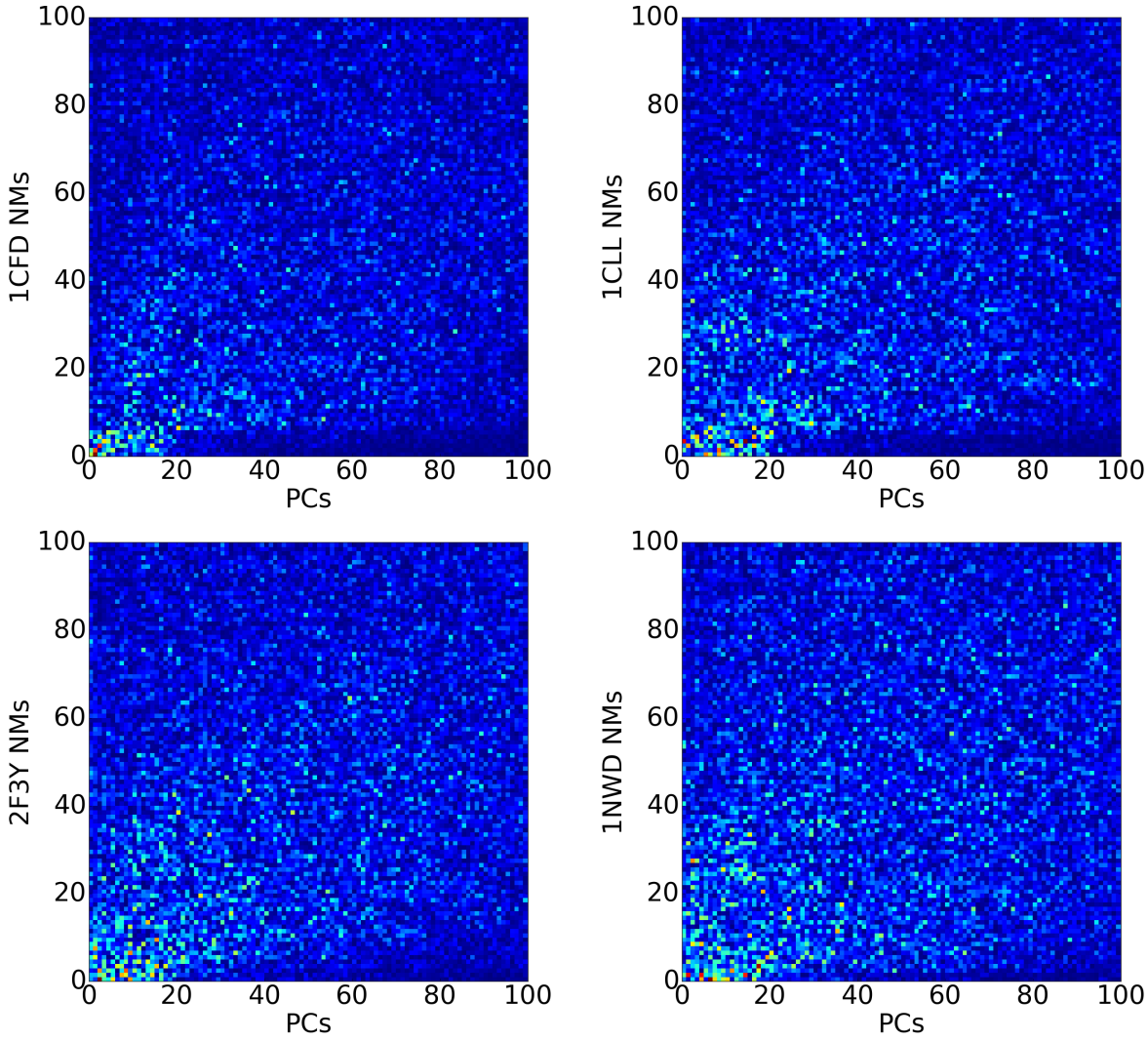


Figure 3.2: Heatmaps comparing CaM Principal Components to Normal Modes extracted from various CaM structures.

Bibliography

Bibliography

[1] D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformational ensembles in biomolecular recognition," *Nature Chem Biol*, vol. 5, no. 11, pp. 789–96, 2009.

[2] R. Nussinov and P. G. Wolynes, "A second molecular biology revolution? the energy landscapes of biomolecular function," *Phys Chem Chem Phys*, vol. 16, no. 14, pp. 6321–6322, 2014.

[3] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, "Principles and overview of sampling methods for modeling macromolecular structure and dynamics," *PLoS Comp. Biol.*, vol. 12, no. 4, p. e1004619, 2016.

[4] D. Russel, K. Lasker, J. Phillips, D. Schneidman-Duhovny, J. A. Veláquez-Muriel, and A. Sali, "The structural dynamics of macromolecular processes," *Curr Opin Cell Biol*, vol. 21, no. 1, pp. 97–108, 2009.

[5] A. Shehu and E. Plaku, "A survey of computational treatments of biomolecules by robotics-inspired methods modeling equilibrium structure and dynamics," *J Artif Intel Res*, vol. 597, pp. 509–572, 2016.

[6] T. Maximova, E. Plaku, and A. Shehu, "Structure-guided protein transition modeling with a probabilistic roadmap algorithm," *IEEE/ACM Trans. Bioinf. and Comp. Biol.*, 2017, doi: 10.1109/TCBB.2016.2586044.

[7] E. Sapin, K. A. De Jong, and A. Shehu, "From optimization to mapping: An evolutionary algorithm for protein energy landscapes," *IEEE/ACM Trans. Bioinf. and Comp. Biol.*, 2017, doi: 10.1109/TCBB.2016.2628745.

[8] E. Sapin, D. B. Carr, K. A. De Jong, and A. Shehu, "Computing energy landscape maps and structural excursions of proteins," *BMC Genomics*, vol. 17, no. Suppl 4, p. 456, 2016.

[9] T. Maximova, D. Carr, E. Plaku, and A. Shehu, "Sample-based models of protein structural transitions," in *ACM Conf Bioinf & Comp Biol (BCB)*, Seattle, WA, 2016, pp. 128–137.

[10] T. Maximova, E. Plaku, and A. Shehu, "Computing transition paths in multiple-basin proteins with a probabilistic roadmap algorithm guided by structure data," in *IEEE Intl. Conf. Bioinf. & Biomed.*, Washington, D.C., 2015, pp. 35–42.

[11] R. Clausen and A. Shehu, "A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes," *J Comp Biol*, vol. 22, no. 9, pp. 844–860, 2015.

[12] R. Clausen, B. Ma, R. Nussinov, and A. Shehu, "Mapping the conformation space of wildtype and mutant H-Ras with a memetic, cellular, and multiscale evolutionary algorithm," *PLoS Comput Biol*, vol. 11, no. 9, p. e1004470, 2015.

[13] W. Qiao, T. Maximova, E. Plaku, and A. Shehu, "Statistical analysis of computed energy landscapes to understand dysfunction in pathogenic protein variants," in *ACM Conf on Bioinf and Comput Biol Workshops (BCBW): Comput Struct Biol Workshop (CSBW)*, Boston, MA, 2017, pp. 1–6.

[14] M. M. Tirion, "Large amplitude elastic motions in proteins from a single parameter, atomic analysis," *Phys Rev Lett*, vol. 77, no. 9, pp. 1905–1908, 1996.

[15] I. Bahar, T. R. Lezon, L. W. Yang, and E. Eyal, "Global dynamics of proteins: bridging between structure and function," *Annu Rev Biophys*, vol. 39, pp. 23–42, 2010.

[16] A. Das, M. Gur, M. H. Cheng, S. Jo, I. Bahar, and B. Roux, "Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003521, 2014.

[17] I. Al-Bluwi, M. Vaisset, T. Siméon, and J. Cortés, "Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods," *BMC Struct Biol*, vol. 13, no. S2, p. Suppl 1, 2013.

[18] A. D. Schuyler, R. L. Jernigan, P. K. Wasba, B. Ramakrishnan, and G. S. Chirikjian, "Iterative cluster-NMA: a tool for generating conformational transitions in proteins," *Proteins: Struct Funct Bioinf*, vol. 74, no. 3, pp. 760–776, 2009.

[19] N. Kantarci-Carsibasi, T. Haliloglu, and P. Doruker, "Conformational transition pathways explored by Monte Carlo simulation integrated with collective modes," *Biophys J*, vol. 95, no. 12, pp. 5862–5873, 2008.

[20] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves, "Bio3D: an R package for the comparative analysis of protein structures," *Bioinformatics*, vol. 22, no. 21, pp. 2695–2696, 2006.

[21] A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallogr A*, vol. 26, no. 6, pp. 656–657, 1972.

Biography

David R.S. Morris grew up in Virginia. He attended Northern Virginia Community college, where he received his Associate of Science degree in Mathematics in 2013. He went on to receive his Bachelor of Science degree in Computer Science from George Mason University in 2015. He then received his Master of Science in Computer Science from George Mason University in 2017. He intends to study towards a PhD in Computer Science in Germany starting in Winter 2017.