Published in final edited form as:

J Phys Chem B. 2009 May 14; 113(19): 6613–6622. doi:10.1021/jp808736c.

Computational Methods for Predicting Sites of Functionally Important Dynamics

Adam D. Schuyler¹, Heather A. Carlson², and Eva L. Feldman¹

¹Department of Neurology, University of Michigan, Ann Arbor, Michigan 48109

²Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan 48109

Abstract

Understanding and controlling biological function of proteins at the atomic level is of great importance; allosteric mechanisms provide such an interface. Experimental and computational methods have been developed to search for residue mutations that produce changes in function by altering sites of correlated motion. These methods are often observational in that altered motions are achieved by random sampling without revealing the underlying mechanism(s). We present two deterministic methods founded on structure-function relationships that predict dynamic control sites (i.e. locations that experience correlated motions as a result of altered dynamics).

The first method (“static”) is based on a single structure conformation (e.g. the wild type (WT)) and utilizes a graph description of atomic connectivity. The local atomic interactions are used to compute the propagation of contact paths. This description of structure connectivity reveals flexible locations that are susceptible to altered dynamics.

The second method (“dynamic”) is a comparative analysis between the normal modes of a WT structure and a mutant structure. A mapping function is defined that quantifies the significance of the motions in one structure projected onto the motions of the other. Each mode is considered up- or down-regulated according to its change in relative significance. This description of altered dynamics is the basis for a motion correlation analysis, from which the dynamic control sites are readily identified.

The methods are theoretically derived and applied using the canonical system dihydrofolate reductase (DHFR). Both methods demonstrate a very high predictive value ($p < 0.005$) in identifying known dynamic control sites. The dynamic method also produces a new hypothesis regarding the mechanism by which the DHFR mutant achieves hyperactivity. These tools are suitable for allosteric investigations and may greatly enhance the speed and effectiveness of other computational and experimental methods.

Keywords

allostery; normal mode analysis (NMA); correlated motion; dihydrofolate reductase (DHFR)

1 Introduction

Macromolecules do not occupy a single rigid conformation, but rather fluctuate over a distribution of possible states [1-3]. Biological function is driven by these motions. Changes

in structure (i.e. by mutation or binding) cause changes in accessible motions, which may alter biological function [4-6]. The correlated motions within the overall ensemble of motions are of particular importance [7,8]; they allow for allosteric activity and possible external control of biological function. It is important to identify the potential control sites and the mechanism by which they interconnect across the structure.

Current experimental methods for discovering dynamic control sites include mutagenesis studies [9,10] and nuclear magnetic resonance (NMR) relaxation [11,12]. The mutagenesis techniques may be guided, but an exhaustive search is often not feasible. The NMR methods are limited in application by protein size and interpretation is often difficult.

Computational alternatives have been developed to more effectively broaden the search space. Simulated mutations are used to map structural perturbations onto a network of cooperative interactions [13]. Ota and Agard [14] developed the ATD (anisotropic thermal diffusion) model for tracking the propagation of kinetic energy emanating from a heated targeted location. This non-equilibrium molecular dynamics method reveals detailed pathways of atomic interactions. However, separate simulations are required for each candidate residue location. In addition, the propagation causes a time shift in the observed correlated dynamics. This time shift is not known in advance and must be identified in the data. The ATD method is quite powerful, especially in testing specific hypothesized mechanisms and testing the effects of mutation and binding events, but it may be limited as a discovery tool.

There has been success in the related study of protein-protein interactions; Russell et al. [15] review experimental and computational methods. Kortemme and Baker [16] review methods for protein-protein interface redesign for the purpose of redirecting biological activity. Computational alanine scanning [17] is useful in systematically predicting the locations of structure hot-spots [18]. However, many of these methods depend on conformation sampling [19] and accurate energy calculations, which are both computationally expensive, at best.

In this paper, we develop two computational models that identify dynamic control sites by analysis of structure geometry and motion. The models provide a foundation for a more complete understanding of the structure-function relationship and provide opportunities for developing external control mechanisms.

The first model, the “static” method, reveals control sites with a geometric analysis of a single conformation (e.g. a wild type (WT)). The second model, the “dynamic” method, characterizes how structure motions change between a reference conformation (e.g. a WT) and a candidate conformation (e.g. a mutant or bound state). The motions are calculated with a normal mode analysis (NMA) variant, further discussed in the NMA Background section. Sites of correlated motion are directly identified within sets of regulated structure motions. The static method is very simple and only requires a single known structure, whereas the dynamic method is more complex and also reveals more information, but requires a pair of available structures.

The methods are derived and applied to the control system, dihydrofolate reductase (DHFR). DHFR is a canonical system for investigating the relationships between function and protein dynamics. DHFR catalyzes the nicotinamide adenine dinucleotide phosphate (NADP) dependent reduction of dihydrofolate to tetrahydrofolate; a folate (FOL) cofactor is involved in the reaction. In the method development and analysis that follows, the WT structure (PDB=1RA2, [20], Figure 1) is compared against a mutant structure (PDB=2D0K, mutations: M1A / M16N / M20L / M42Y / C85A / M92F / C152S, [21]) which demonstrates increased enzymatic activity.

The review article by Hammes-Schiffer and Benkovic [22] addresses the link between protein motion and catalysis, with discussion focussing on mutagenesis, network models and

conserved sequence analysis. Schnell et al. [5] describe the features of the DHFR structure and the motions it experiences in the course of completing the catalytic cycle. The sites correlated with the dynamic driving force of the enzyme are identified and discussed in multiple sources [23-29]; these known residues are collectively referred to as the set $K = \{2, 9 - 10, 17, 20, 38, 42 - 48, 54, 63, 66 - 67, 70, 73 - 75, 88, 98, 106, 113, 117 - 124, 144, 148, 158\}$. The primary goal of the static and dynamic methods is to predict the dynamic control sites of K . Both methods demonstrate very high predictive values ($p < 0.005$) and prove to be powerful tools for investigating allosteric mechanisms.

2 NMA Background

The structure-function relationship is fundamental to the mechanism question central to this paper. X-ray crystallography reveals atomic structure (protein data bank, PDB [30]). NMA methods produce a set of harmonic motions (the normal modes) around an equilibrium conformation and are particularly effective at establishing a relationship between the crystal coordinates and the corresponding structure dynamics.

Classical NMA methods offer a tremendous improvement in computational performance over molecular dynamics, but are still prohibitive for large structures and for application across large structure sets. Coarse-grained NMA techniques [31,32] reduce the structure by representing groups of atoms as single point masses (e.g. NMA on the alpha-carbon trace). This effectively collapses all atoms of the group and their interactions with other atoms onto a single position. This action reduces the overall number of degrees of freedom (DOFs) in the simulation at the expense of destroying the geometry of the atomic interactions.

The cluster NMA (cNMA) method was developed in response to the inherent inaccuracies of point mass, coarse-grained NMA (see Schuyler and Chirikjian [33,34] for the full derivation and application). cNMA represents groups (or clusters) of atoms as rigid-bodies, thus achieving the desired reduction in DOFs, while maintaining individual atomic positions within each cluster, thereby including the full network of atomic interactions (Figure 2). The result is an efficient and effective way to identify large scale cooperative motions.

Given the coordinates for an n atom structure (including ligands), a point mass is placed at each atom's location. The atoms are grouped into N clusters [34-37]. The authors have previously developed an approach for validating a clustering scheme prior to performing the cNMA computations [38]. In addition, the authors observe that structure rearrangements within a residue are of smaller magnitude than typical crystal structure resolution; this allows clustering by residue. The setup for cNMA is summarized below.

Each cluster's generalized coordinates are given by

$$\delta_c = [\chi_c^T, \gamma_c^T]^T \in \mathbb{R}^6 \quad (1)$$

where $\chi_c \in \mathbb{R}^3$ parameterizes the translational displacement of cluster c and $\gamma_c \in \mathbb{R}^3$ parameterizes the rotational displacement of cluster c . The whole structure's generalized coordinates are the stacked vector

$$\delta = [\delta_1^T, \dots, \delta_N^T]^T \in \mathbb{R}^{6N} \quad (2)$$

The complex set of atomic interactions is well approximated by a harmonic potential, which is implemented as an elastic network model (ENM, [39]). The classic distance-based ENM is

defined as a network of springs connecting all atomic pairs that are within a cutoff distance, r_c . This relationship is expressed as an adjacency matrix, A , where each entry is defined by

$$A_{a_i, a_j} \triangleq A_{i, j} = \begin{cases} 1 & d(a_i, a_j) \leq r_c \\ 0 & \text{else} \end{cases} \quad (3)$$

where a_i is the i^{th} atom and $d(a_i, a_j)$ is the Cartesian distance between atoms a_i and a_j . By convention, self contact is not included ($A_{a_i, a_i} = 0$). In addition to its role in cNMA, the adjacency matrix is the basis for the static analysis.

Varying the cutoff distance has been studied [40] and for structure representations including all atoms, a value of $r_c = 5 \text{ \AA}$ effectively captures the network of atomic linkages. Non-uniform spring constants have also been studied and their use on a C_α trace representation has demonstrated improved alignment with B-factor data [41]. However, this modification is not necessary for the current approach because the rigid clusters of cNMA lock all covalent bond lengths and geometries in their native configuration, with the exception of the peptide bonds. Since the current method utilizes an all-atom representation, springs between nearest and next-nearest neighbors form a dense bundle of interactions along the backbone and stabilize the peptide bonds.

The set of normal modes forms a basis for the space of all possible structure motions. The significance of this basis is that the potential energy associated with displacing the structure along one of its modes is proportional to its frequency (i.e. the lower frequency modes contribute the most to the ensemble dynamics [38]). When ordered by increasing frequency, the first 6 modes are rigid translations and rotations of the structure. The remaining non-rigid modes involve structure deformations and are utilized in the dynamic analysis.

3 Methods

3.1 Static Analysis

The mechanical function of a protein is inherently related to the geometry of its folded structure. The goal of this section is to develop a method for characterizing structural linkages between all pairs of atoms. The commonly used distance based interaction model is a natural starting place and establishes the framework for the new contact path model.

3.1.1 Direct Contact Model (DCM)—The local connectedness of each atom is quantified by its total number of contacts as defined by the adjacency matrix (Equation 3). The degree of an atom (a_i) is the sum of values in the corresponding column of A

$$\Delta(a_i) = \sum_{\alpha=a_1}^{a_n} A_{\alpha, a_i} \quad (4)$$

This function is normalized by scaling the maximum value to 1, thus giving a measure of each atom's relative connectedness with a value on the interval [0, 1].

The DCM is composed of the adjacency matrix and the normalized degree function. This basic model only identifies nearest neighbor atomic interactions, which are fairly uniform across all non-surface atoms (i.e. those atoms that have a complete “shell” of neighbors). However, the distance based description of atomic interactions at the core of the DCM serves as the foundation for the more informative contact path model.

3.1.2 Contact Path Model (CPM)—A contact path is defined as a sequence of serially connected atom pairs, as defined by the adjacency matrix. The i^{th} atom in a path is represented as p_i , where $p_i \in \{a_1, \dots, a_n\}$. For a given path length (s), starting atom (p_0) and ending atom (p_s), the set of all possible paths is represented as

$$\mathcal{P}(p_0, p_s, s) = \{p_0 \leftrightarrow p_1 \leftrightarrow \dots \leftrightarrow p_s\} \quad (5)$$

The double sided arrows indicate pairwise connections and are equivalent to the constraint: $A_{p_{i-1}, p_i} = 1$, for $i \in \{1, \dots, s\}$.

The number of paths in the set, denoted as $|\mathcal{P}(p_0, p_s, s)|$, increases exponentially with path length and direct enumeration is not desirable. However, it is possible to formulate contact path enumeration in terms of the adjacency matrix

$$|\mathcal{P}(p_0, p_s, s)| = [A^s]_{p_0, p_s} \quad (6)$$

This powerful relationship (proved in Appendix A) states that by repeatedly multiplying the adjacency matrix by itself, which is computationally cheap, the entries of the result produce the contact path enumeration values between *all pairs of atoms*.

The atomic interactions of the DCM are represented by a *single* function - the adjacency matrix. In contrast, contact paths are parameterized by path length (s) and form the set of interaction functions

$$\mathcal{P}_s = \{\mathcal{P}(a_i, a_j, 1), \dots, \mathcal{P}(a_i, a_j, s)\} \quad (7)$$

Each function in this family describes atomic interactions over a different length scale. The relationship in Equation 6, allows the interaction family in Equation 7, to be redefined in terms of the adjacency matrix as

$$\mathcal{P}_s = \{A^1, \dots, A^s\} \quad (8)$$

The degree function for the DCM is augmented with a parameter for path length (s) and is applied to each function in $\mathcal{P}_s$. This degree function is normalized and quantifies the relative density of paths of length s that radiate from each atom to the surrounding structure. The CPM is defined as the interaction family $\mathcal{P}_s$ and the corresponding set of normalized degree functions.

The interaction model of the DCM is represented by just the adjacency matrix (i.e. A^1), which is the first interaction model in the CPM series. The CPM is therefore a direct extension of the DCM. As the path length is increased, the longer range, indirect atomic interactions missed by the DCM become accessible to the CPM.

3.1.3 Dual Plot—The connectivity information contained in the CPM is represented by a series of dual plots (Figure 3). Each dual plot is composed of a visualization of an interaction model from $\mathcal{P}_s$ (i.e. a plot showing where the non-zero entries are located) along with a corresponding degree function plot.

Figure 3 shows plots for the CPM at path lengths of $s = \{1, 3, 5, 10, 15, 20\}$, which are categorized into three groups. The first group includes only the $s = 1$ dual plot. This plot

reproduces the DCM and illustrates the common foundation of the CPM and DCM. The second group of plots forms a transition phase in which the features of the DCM evolve into a representation of the CPM. The end of this phase and the beginning of the third phase is marked by the convergence of the dual plot series to a steady state. Upon the completion of this transition, the features of the CPM become dominant. The analysis of the CPM is performed in the third phase on the $s = 20$ dual plot for which the convergence is qualitatively sufficient.

The $s = 20$ degree function exhibits well defined peaks and valleys, as compared to the nearly even “noise” of the $s = 1$ degree function. The peaks in the $s = 20$ degree function indicate highly connected atoms (i.e. rigid locations) and the valleys indicate less connected atoms (i.e. flexible locations). These features are masked by the nearest neighbor DCM, but revealed by the more comprehensive CPM.

3.2 Dynamic Analysis

The dynamic method is primarily concerned with mapping the significance of a mode relative to its own mode set onto the mode set of a modified structure. This requires two components: (1) a way to map the modes from one mode set to another, and (2) a way to assign significance to each mode of a given mode set. The integration of these items allows for the effect that structure modification has on dynamics to be quantified. These regulated modes are the foundation for studying correlated dynamics.

3.2.1 Mode Reduction and Decomposition—The accessible motions of a WT and a mutant structure are determined by performing cNMA (all biologically relevant ligands are included). The structures may be complexed with ligands, even different ligands (e.g. the DHFR WT is complexed with FOL and NADP, while the mutant is only bound with FOL). In order to most directly identify the altered dynamics, the protein's motions within a bound complex are isolated with the following reduction procedure: (1) The ligand motions are excluded by truncating each mode; and (2) the Gram-Schmidt orthonormal process (GSOP) is employed to remove redundant motions and ensure the linear independence within each mode set. This procedure is further discussed below.

The number of DOFs allotted to a given structure for cNMA is represented as

$$d = \underbrace{6N_p}_{d_p} + \underbrace{6N_l}_{d_l} \quad (9)$$

where d_p is the number of DOFs allotted for the N_p clusters defined on the protein (e.g. one for each residue in DHFR) and d_l is the number of DOFs allotted for the N_l clusters defined across the ligands (e.g. Figure 4 for NADP and FOL). For a given set of modes, the first step in the reduction procedure (ligand motion exclusion) is minor; just truncate each mode to remove all ligand motions (it is important to initially compute the modes *with* the ligands, so their contributions are included). This results in a mode set that still has d modes, but each mode only has the d_p DOFs of the protein atoms. Since $d > d_p$, the set of modes is redundant and obscures a direct comparison.

Linear independence is reestablished in the second step of the reduction procedure by trimming the set down to d_p modes. There are many ways to perform such a reduction, but the GSOP gives priority to the lowest modes, while forcing greater modification of the highest modes. This is appropriate, as the lower modes contribute a greater magnitude to the overall structure dynamics.

Upon completion of the mode reduction procedure, the set of mode shape/frequency pairs for the WT is given as

$$W = \{v_i, f_i\} \quad (10)$$

and for the mutant structure as

$$M = \{\widehat{v}_i, \widehat{f}_i\} \quad (11)$$

The motions of W and M are compared with the decomposition matrix D , whose entries are defined as

$$D_{i,j} = |\widehat{v}_i \cdot v_j| \quad (12)$$

The entries of D are on the interval $[0, 1]$ with higher values indicating alignment between the corresponding modes and lower values indicating linear independence. This type of direct comparison between mode sets is made possible by the fact that each mode set forms a basis; any possible structure motion is able to be represented by some combination of the modes in each set. Each column in D corresponds to a mode from W and the entries are the decomposition values over the modes of M . Similarly, each row corresponds to a mode from M and the entries are the decomposition values over the modes of W . This bidirectional relationship means that the decomposition matrix is a conversion function that relates the basis motions between a pair of structures.

The decomposition matrix for the DHFR mutant compared to the WT (Figure 5) is strongly diagonal, as expected, indicating similarity of the motion spaces. However, there is more “spread” from the diagonal than expected, especially given that the mutant only differs from the WT by 1.13Å root mean square displacement (RMSD) over their common atoms. The relative significance values defined in the next section allow specific information about structure dynamics to be extracted from this “spread”.

3.2.2 Shift Analysis—Consider a column (j) in the decomposition matrix. By the definition of a basis set, the corresponding mode from W is completely captured by the modes of M . This is represented by the column norm expression

$$\sum_{i=1}^{d_p} [D_{i,j}]^2 = 1 \quad (13)$$

Equation 13 implicitly gives equal significance across all modes in M ; that is, if $D_{a,j} = D_{b,j}$, then mutant modes $\widehat{v}_a$ and $\widehat{v}_b$ contribute equally to WT mode u_j . However, the pair of mutant modes do not necessarily contribute equally to the ensemble of structure dynamics. Normal modes produce motions of equal valued energy when they are scaled inversely to their frequency [42, 43]. The equipartition theorem indicates that the lower frequency modes contribute relatively more to the overall ensemble of structure oscillations [38]. The relative significance of mode i within M is given by

$$\widehat{r}_i = \frac{1/\widehat{f}_i}{\sum_{j=7}^{d_p} 1/\widehat{f}_j} \quad (14)$$

The relative significance values for each mode of $\mathcal{M}$ are incorporated into Equation 13, as weighting factors and yield the modified expression

$$\rho_j = \sum_{i=7}^{d_p} \widehat{r}_i [D_{i,j}]^2 \quad (15)$$

This function uses the decomposition matrix as a conversion function for quantifying the *significance* of each mode from $\mathcal{W}$ with respect to the mode set $\mathcal{M}$.

The effect that mutations have on the ensemble of structure motions is quantified by computing how the significance of each mode from $\mathcal{W}$ shifts when it is projected onto the mode set $\mathcal{M}$. The shift function is defined as the ratio

$$S_i = \frac{\rho_i}{r_i} \quad (16)$$

and produces values greater than 1 for motions that are more significant to the structure dynamics of the mutant than the WT (i.e. up-regulated in mutant) and it produces values less than 1 for motions that are less significant to the mutant than the WT (i.e. down-regulated in mutant). This methodology is derived with modes of $\mathcal{W}$ mapped onto $\mathcal{M}$; a similar derivation holds for mapping the modes of $\mathcal{M}$ onto $\mathcal{W}$. It is critical to note that modes up(down)-regulated from $\mathcal{W}$ into $\mathcal{M}$ are not equivalent to the modes down(up)-regulated from $\mathcal{M}$ into $\mathcal{W}$ (i.e. the significance mapping is not invertible). The shift function is shown in Figure 6, for both mapping directions.

3.2.3 Correlated Motions Expressed by Altered Dynamics—It is common practice to compute the correlated motions across a structure for a given normal mode, but Van Wynsberghe and Cui [8] have demonstrated it is more appropriate to look at correlations over a mode ensemble. The up-and down-regulated mode sets, determined by shift analysis, are the basis for defining such mode ensembles.

The GSOP alters each of the truncated cNMA modes by removing all portions of the mode that align with any of the lower frequency modes. As each mode is altered, its direction becomes unavailable to all subsequent modes. This accumulation of constraints forces the higher modes to undergo a larger adjustment than the lower modes. In addition, this effect is enhanced as mode frequencies typically become less distributed at the higher end of the frequency spectrum [44]. This inherent numerical error is excluded from our analysis by discarding modes that do not retain at least 90% alignment with their original form. A plot of these residual values is shown in Figure 7.

The GSOP is defined on just the protein atoms so that the shift analysis targets the core dynamics. However, once the GSOP transformation is defined, it is applied to *all* atoms so that the bound ligands are observed in the motion correlation analysis.

Figure 8 shows motion correlation plots for the up- and down-regulated modes of WT and mutant DHFR. Each correlation plot (shown as a color map) has an associated plot attached below, that shows each alpha carbon's and each ligand atom's average contributions to positive (red line) and negative (blue line) correlated motions. The data displayed in these plots are utilized to predict the control sites.

3.3 Testing Model Predictions

The predictive quality of the static and dynamic models is tested on DHFR by comparing the models' predicted locations of control sites with those identified in the literature [23-29] and previously stated as set K .

3.3.1 Static Method Prediction—The static method is based on identifying dense networks of contact paths (i.e. rigid locations), whereas atoms involved in altered dynamics require flexible connections to the surrounding structure [5]. Consequently, the locations of valleys in the CPM degree function (i.e. bottom panel of $s = 20$ plot in Figure 3) are proposed as a predictive measure for identifying control sites. The “data vector” ($\mathbf{x}$) is defined as the inverse values of the CPM degree function at $s = 20$. This vector is normalized over the DHFR atoms (ligands are excluded) and has values approaching 1 for the most flexible locations and values approaching 0 for the more rigid locations.

The set of predicted control sites, referred to as $P_{\text{stat}}(\mathcal{W})$, is obtained by applying the following procedure:

1. Compute the following ratio for each alpha-carbon: $\mu_i = \frac{x_i - \langle \mathbf{x} \rangle}{\langle \mathbf{x} \rangle}$, where $\langle \mathbf{x} \rangle$ is the mean value of $\mathbf{x}$. Positive (negative) values of μ_i indicate residues that are more (less) susceptible to involvement in correlated motions.
2. Starting with the residues with the largest μ values, create a set whose size is κ times the size of K .
3. Group these residues into runs of consecutive indices; each group is the top of a peak.
4. Remove any of the groups that do not contain a residue with a μ value at least as large as the threshold cutoff, τ .

There are two parameters which define this procedure. The scaling factor, $\kappa > 1$, allows for more complete identification of control sites, especially since the peaks in the CPM inverse data typically span several residues on either side of the peak. Increasing the size of the prediction set results in more false positives, but many more true positives; both factors are accounted for in the statistical testing that follows. This approach only works for a structure where the number of control sites is known in advance. In general, it will be more practical to set the desired size of the prediction set relative to the size of the structure. In the current DHFR analysis, a value of $\kappa = 1.5$ is used. The threshold cutoff, τ , safeguards against over-interpreting a noisy data set in which most μ values are near the mean. The DHFR analysis is performed with $\tau = 1$.

3.3.2 Dynamic Method Prediction—Figure 8, shows the results of the dynamic method applied to all 4 combinations of the following parameters: structure $\in \{\mathcal{W}/\mathcal{M}\}$ and regulation $\in \{+/-\}$. For each of the parameter combinations, plots below the motion correlation show each atom's average contribution to positively (red) and negatively (blue) correlated motions. The peaks in the red plots indicate atoms contributing the most to positively correlated motions and the valleys in the blue plots indicate atoms that contribute the most to negatively correlated motions. Each of these 8 data sets serves as the “data vector” ($\mathbf{x}$) and prediction sets are formed according to the procedure in the previous section. For example, $P_{\text{dyn}}(\mathcal{W}, +, -)$ is the set of

predicted control sites resulting from the WT structure's, up-regulated modes showing negatively correlated motions.

The scaling factor parameter has the same meaning as before and the same value is used ($\kappa = 1.5$). The cutoff threshold is specific to the measurement type, which is different between methods. The static method measures connectivity and the dynamic method measures magnitude of motion. The cutoff threshold for the dynamic prediction is set at $\tau = 0.45$.

3.3.3 Testing the Predictions—The prediction sets are tested in two ways:

T1 Each set of consecutive atoms in a prediction set that contains a K residue is counted as a hit, otherwise a miss.

T2 Each residue in a prediction set that is a K residue is counted as a hit, otherwise a miss.

The T1 test ensures that the predicted peaks and valleys hit K residues and the T2 test makes sure each K residue is hit by the prediction set. The true positives and false positives observed in the above tests are analyzed with a chi-squared test to produce an associated p-value. The results are shown for the static model in Figure 9 and for the dynamic model in Figure 10. The yellow highlighting in each Figure marks significant findings.

4 Discussion and Summary

The static and dynamic methods are developed with the purpose of predicting sites of altered dynamics in a system with an unknown mechanism. In the following sections: (1) the static and dynamic methods are validated against the known, experimentally derived control sites of DHFR, (2) the parameters which define the dynamic method are correlated against the findings to reveal a mechanistic driving force behind the hyperactive DHFR mutant, and (3) a workflow for the study of new systems with unknown control sites is presented.

4.1 Method Validation Against DHFR

The static method utilizes a basic structure description and demonstrates a high predictive value for the analysis of both the WT and mutant structures (Figure 9). In practice, the static method only needs to be applied to a single structure, but this application demonstrates that any form of the candidate structure may be suitable. In addition, the CPM is a robust model which captures the general distribution of contacts throughout a structure. This topological description is largely invariant to local structure rearrangements and is even suitable for structures of low resolution.

The majority of the eight parameter combinations in the dynamic method (Figure 10) yield statistically significant predictions. There are two prediction sets which are clear outliers: the $P_{\text{dyn}}(\text{M}, +, +)$ set is empty and the $P_{\text{dyn}}(\text{W}, -, +)$ set contains only a single control location, which is not observed in any of the other sets. The consensus prediction set is defined to include residues that are predicted by at least half of the remaining prediction sets. The consensus predictions are: {15 - 24, 48 - 52, 63 - 71, 86 - 89, 119 - 123, 141 - 150} and result in a T1 p-value $< 1e - 6$ and a T2 p-value of 0.08. The residues define 4 distinct spatial locations in the structure: (1) the Met20 loop and supporting structure, (2) the αC helix, which is across the binding site from Met20, (3) the CD loop of the adenosine binding domain, and (4) part of the DE loop, which is on the opposite side of the structure as the binding site.

4.2 DHFR Hyperactivity Mechanism

In the previous section, the dynamic method was validated by extracting a consensus set of predicted residues across all parameter sets. In this section, the consensus parameters are

extracted by partitioning the prediction sets according to statistical significance. The known control sites allow the identification of parameters which most strongly produce statistically significant prediction sets. These identified parameters are the basis for hypothesizing the hyperactivity mechanism of the mutant form.

The first group is derived from the up-regulated WT motions (i.e. mutant shows “more” of what WT has) and the down-regulated mutant motions (i.e. WT shows “less” of what the mutant has). The corresponding prediction sets are

$$G^+ = \{P_{\text{dyn}}(W, +, +), P_{\text{dyn}}(W, +, -), P_{\text{dyn}}(M, -, +), P_{\text{dyn}}(M, -, -)\} \quad (17)$$

The second group is derived from the down-regulated WT motions (i.e. mutant shows “less” of what WT has) and the up-regulated mutant motions (i.e. WT shows “more” of what the mutant has). The corresponding prediction sets are

$$G^- = \{P_{\text{dyn}}(W, -, +), P_{\text{dyn}}(W, -, -), P_{\text{dyn}}(M, +, +), P_{\text{dyn}}(M, +, -)\} \quad (18)$$

The G^+ group captures dynamics of the mutant that are enhanced relative to the dynamics of the WT and the G^- group captures the dynamics of the WT that have been damped in the mutant. Given this physical interpretation of the partition groups, the statistical differentiation of these groups is now demonstrated. The G^+ group has T1 p-values < 0.005 for all of its prediction sets, while the G^- group only has one T1 p-value < 0.005 and a second p-value < 0.05 . The G^+ group has two T2 p-values that are significant and the G^- set has none. In fact, one of the G^- data sets fails to predict a single control site and another set predicts only a single location, which is a false positive. The G^+ set only has 2 false positive locations and both are adjacent to known control sites. The G^- set contains 8 false positive locations, including 3 locations that are not adjacent to control sites.

The high predictive value of the G^+ group relative to the G^- group prompts the hypothesis that the hyperactivity of the DHFR mutant is caused by an enhanced representation of the WT motions that most effectively trigger the control sites. Iwakura et al. [21] classify the hyperactive mutant as being more flexible and having more surface cavities with a larger total volume, which “increase the room for dynamic motions of side chains and even the backbone, making the molecule flexible”. Our hypothesis of enhanced motion is consistent with these findings. Iwakura et al. [21] state that the THF release rate of the enzyme·NADPH·THF complex accounts for most of the hyperactivity, but the remaining increase may be regulated by the G^+ motions.

4.3 Discovering Unknown Allosteric Mechanisms

Given a single WT structure the goal is to discover locations that are susceptible to altered dynamics and develop a methodology for triggering these locations, thereby activating an inherent allosteric mechanism. The characterization of these interactions will lead to a better understanding of the biological mechanism and potentially allow for its control.

The control sites for a given system are predicted by applying the static method to the WT structure. The next task is to determine how to trigger these control sites; there are two obvious avenues of pursuit. The most direct option is to use docking methods to design small molecules that bind to the surface exposed control sites. Additional computational and experimental studies of such derived complexes may reveal altered function.

A second option is to find structure mutations which activate the dynamic control sites. The RMSD distance from each of the DHFR mutation site alpha-carbons to each of the known control site alpha-carbons is $> 21\text{\AA}$. This indicates that mutations directed at the control sites may not provide the optimal approach. Computational mutagenesis may serve as a screening tool. Any computationally derived, candidate mutant structure that shows increased mobility at the predicted control sites (as simply determined by cNMA) is further analyzed against the WT with the dynamic method. The extent of mode regulation is an indication of the mutation's effect on correlated dynamics.

The DHFR mutant is structurally similar to the WT (RMSD $< 1.2\text{\AA}$); this gives the physical justification for mapping the cNMA motions of one structure onto the other. If a large structural change is introduced by a given mutation the physical meaning of such a mapping becomes less apparent. This limitation may be a factor in studying extreme structure modifications, but it is extremely unlikely that structures with only a few mutations will substantially differ in conformation from the WT. However, if such a change is observed, then the static method is suitable for directly analyzing the mutant. If the dynamic method is also desired, then it may be necessary to redefine the mapping function to account for the altered geometry (e.g. the optimal alignment of a subregion in one structure with the corresponding subregion in the other may be used to define a shift in position and orientation).

Studies of additional systems may reveal correlations between types of biological functional and the parameters that define the dynamic model. For example, the DHFR analysis identified the G^+ motions as being more significant than the G^- motions. In general, it is conceivable that damped motions (i.e. those of G^-) play a role in modulating enzymatic activity, perhaps this will be observed for hypo-active structures.

4.4 Summary

The static and dynamic methods employ a flexible framework that can be readily utilized by other techniques. From the static method: the dual plot is an effective way to evaluate the impact of a local contact property as it propagates across a structure. From the dynamic method: the significance weighted decomposition mapping employed by the shift function (Equation 16) is suitable for mapping any motion-related property of one structure onto another. The static and dynamic analysis techniques are suitable as stand alone methods, but complement each other in a very powerful combined approach to allosteric investigation.

A Contact Path Enumeration

The relationship in Equation 6, is proven by induction in two parts. First, the result for a base case (i.e. for path length $s = 2$) is established. Second, it is shown that given the property for path length $s - 1$, the property holds for path length s .

The definition of matrix multiplication applied to the self multiplication of the adjacency matrix gives the expression

$$[A^2]_{a_i, a_j} = \sum_{\alpha=a_1}^{a_n} [A]_{a_i, \alpha} [A]_{\alpha, a_j} \quad (19)$$

The first term in the product within the summation is only non-zero if atom a_i is directly linked to atom α . Similarly, the second term in the product is only non-zero if atom is directly linked to atom a_j . Therefore the product in the summation is only non-zero when atom a_i connects to some intermediate atom (α) which also connects to atom a_j . This is precisely the definition of

a contact path of length 2. Since the summation varies the intermediate atom over all atoms of the structure, all possible contact paths connecting a_i to a_j of length 2 are enumerated.

For the inductive step, we again start with the application of the matrix multiplication definition

$$[A^s]_{a_i, a_j} = \sum_{\alpha=a_1}^{a_n} [A^{s-1}]_{a_i, \alpha} [A]_{\alpha, a_j} \quad (20)$$

Since it is given that the property holds for paths of length $s - 1$ we can substitute into Equation 20, to

$$[A^s]_{a_i, a_j} = \sum_{\alpha=a_1}^{a_n} |\mathcal{P}(a_i, \alpha, s - 1)| \cdot [A]_{\alpha, a_j} \quad (21)$$

The first term of the product within the summation gives the number of paths of length $s - 1$ that connect atom a_i to atom α . The second term in this product indicates if atom α directly connects to atom a_j . Since this summation varies over all intermediate atoms (α), the result is the total number of paths of length s connecting atom a_i to atom a_j . Therefore, the relationship stated in Equation 6, is verified.

5 Acknowledgements

A.D.S. is supported by the Sinai Medical Sta Foundation, E.L.F. is supported by a grant from the National Institutes of Health (U01 DK076160), the Juvenile Diabetes Research Foundation Center for the Study of Complications in Diabetes, the Program for Neurology Research and Discovery and the A. Alfred Taubman Medical Research Institute, and H.A.C. is supported by grants from the National Institutes of Health (GM65372) and the National Science Foundation (MCB 0546073).

References

1. Goh, Chern-Sing; Milburn, Duncan; Gerstein, Mark. Conformational changes associated with protein-protein interactions. *Current Opinion in Structural Biology* Feb;2004 14(1):104–109. [PubMed: 15102456]
2. Miyashita O, Onuchic JN, Wolynes PG. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences of the United States of America* Oct;2003 100(22):12570–12575. [PubMed: 14566052]
3. Kovacs, Julio A.; Cavasotto, Claudio N.; Abagyan, Ruben. Conformational sampling of protein flexibility in generalized coordinates: Application to ligand docking. *Journal of Computational and Theoretical Nanoscience* Sep;2005 2(3):354–361.
4. Arora, Karunesh; Brooks, Charles L. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proceedings of the National Academy of Sciences of the United States of America* Nov;2007 104(47):18496–18501. [PubMed: 18000050]
5. Schnell, Jason R.; Jane Dyson, H.; Wright, Peter E. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annual Review of Biophysics and Biomolecular Structure* 2004;33:119–140.
6. Henzler-Wildman, Katherine A.; Lei, Ming; Thai, Vu; Kerns, S Jordan; Karplus, Martin; Kern, Dorothee. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* Dec; 2007 450(7171):913–916. [PubMed: 18026087]
7. Ma, Jianpeng; Karplus, Martin. The allosteric mechanism of the chaperonin GroEL: A dynamic analysis. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95 (15):8502–8507. [PubMed: 9671707]
8. Van Wynsberghe, Adam W.; Cui, Qiang. Interpreting correlated motions using normal mode analysis. *Structure* Nov;2006 14(11):1647–1653. [PubMed: 17098190]

9. Lockless, Steve W.; Ranganathan, Rama. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* Oct;1999 286(5438):295–299. [PubMed: 10514373]
10. Goodey, Nina M.; Benkovic, Stephen J. Allosteric regulation and catalysis emerge via a common route. *Nature Chemical Biology* Aug;2008 4(8):474–482.
11. Wand, A Joshua. Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nature Structural Biology* Nov;2001 8(11):926–931.
12. Popovych, Nataliya; Sun, Shangjin; Ebright, Richard H.; Kalodimos, Charalampos G. Dynamically driven protein allostery. *Nature Structural & Molecular Biology* Sep;2006 13(9):831–838.
13. Hilser, Vincent J.; Dowdy, David; Oas, Terrence G.; Freire, Ernesto. The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. *Proceedings of the National Academy of Sciences of the United States of America* Aug;1998 95(17):9903–9908. [PubMed: 9707573]
14. Ota, Nobuyuki; Agard, David A. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *Journal of Molecular Biology* Aug;2005 351(2):345–354. [PubMed: 16005893]
15. Russell, Robert B.; Alber, Frank; Aloy, Patrick; Davis, Fred P.; Korkin, Dmitry; Pichaud, Matthieu; Topf, Maya; Sali, Andrej. A structural perspective on protein-protein interactions. *Current Opinion in Structural Biology* Jun;2004 14(3):313–324. [PubMed: 15193311]
16. Kortemme, Tanja; Baker, David. Computational design of protein-protein interactions. *Current Opinion in Chemical Biology* Feb;2004 8(1):91–97. [PubMed: 15036162]
17. Massova, Irina; Kollman, Peter A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society* 1999;121(36):8133–8143.
18. Kortemme, Tanja; Baker, David. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America* Oct;2002 99(22):14116–14121. [PubMed: 12381794]
19. Hilser VJ, Freire E. Structure-based calculation of the equilibrium folding pathway of proteins. correlation with hydrogen exchange protection factors. *Journal of Molecular Biology* Oct;1996 262(5):756–772. [PubMed: 8876652]
20. Sawaya MR, Kraut J. Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence. *Biochemistry* Jan;1997 36(3):586–603. [PubMed: 9012674]
21. Iwakura, Masahiro; Maki, Kosuke; Takahashi, Hisashi; Takenawa, Tatsuyuki; Yokota, Akiko; Katayanagi, Katsuo; Kamiyama, Tadashi; Gekko, Kunihiro. Evolutional design of a hyperactive cysteineand methionine-free mutant of escherichia coli dihydrofolate reductase. *Journal of Biological Chemistry* May;2006 281(19):13234–13246. [PubMed: 16510443]
22. Hammes-Schiffer, Sharon; Benkovic, Stephen J. Relating protein motion to catalysis. *Annual Review of Biochemistry* 2006;75:519–541.
23. Radkiewicz, Jennifer L.; Brooks, Charles L. Protein dynamics in enzymatic catalysis: Exploration of dihydrofolate reductase. *Journal of the American Chemical Society* 2000;122(2):225–231.
24. Agarwal, Pratul K.; Billeter, Salomon R.; Rajagopalan, PT Ravi; Benkovic, Stephen J.; Hammes-Schiffer, Sharon. Network of coupled promoting motions in enzyme catalysis. *Proceedings of the National Academy of Sciences of the United States of America* Mar;2002 99(5):2794–2799. [PubMed: 11867722]
25. Rajagopalan, P. T Ravi; Lutz, Stefan; Benkovic, Stephen J. Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates. *Biochemistry* Oct;2002 41(42):12618–12628. [PubMed: 12379104]
26. Rod, Thomas H.; Radkiewicz, Jennifer L.; Brooks, Charles L. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the United States of America* Jun;2003 100(12):6980–6985. [PubMed: 12756296]
27. Wong, Kim F.; Watney, James B.; Hammes-Schiffer, Sharon. Analysis of electrostatics and correlated motions for hydride transfer in dihydrofolate reductase. *Journal of Physical Chemistry B* 2004;108(32):12231–12241.
28. Wong, Kim F.; Selzer, Tzvia; Benkovic, Stephen J.; Hammes-Schiffer, Sharon. Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate

- reductase. Proceedings of the National Academy of Sciences of the United States of America May; 2005 102(19):6807–6812. [PubMed: 15811945]
29. Chen, Jie; Dima, Ruxandra I.; Thirumalai, D. Allosteric communication in dihydrofolate reductase: signaling network and pathways for closed to occluded transition and back. *Journal of Molecular Biology* Nov;2007 374(1):250–266. [PubMed: 17916364]
 30. Berman, Helen M.; Westbrook, John; Feng, Zukang; Gilliland, Gary; Bhat, TN.; Weissig, Helge; Shindyalov, Ilya N.; Bourne, Philip E. The protein data bank. *Nucleic Acids Research* 2000;28(1): 235–242. [PubMed: 10592235] URL <http://www.pdb.org/>
 31. Tozzini, Valentina. Coarse-grained models for proteins. *Current Opinion in Structural Biology* Apr; 2005 15(2):144–150. [PubMed: 15837171]
 32. Bahar, Ivett; Rader, AJ. Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology* Oct;2005 15(5):586–592. [PubMed: 16143512]
 33. Schuyler, Adam D.; Chirikjian, Gregory S. Normal mode analysis of proteins: a comparison of rigid cluster modes with C_α coarse graining. *Journal of Molecular Graphics and Modelling* 2004;22(3): 183–193. [PubMed: 14629977]
 34. Schuyler, Adam D.; Chirikjian, Gregory S. Efficient determination of low-frequency normal modes of large protein structures by cluster-NMA. *Journal of Molecular Graphics and Modelling* 2005;24 (1):46–58. [PubMed: 15990344]
 35. Thorpe MF, Lei Ming, Rader AJ, Jacobs Donald J, Kuhn Leslie A. Protein flexibility and dynamics using constraint theory. *Journal of Molecular Graphics and Modelling* 2001;19(1):60–69. [PubMed: 11381531]
 36. Rader AJ, Hespenheide Brandon M, Kuhn Leslie A, Thorpe MF. Protein unfolding: Rigidity lost. *Proceedings of the National Academy of Sciences of the United States of America* 2002;99(6):3540–3545. [PubMed: 11891336]
 37. Gohlke, Holger; Thorpe, MF. A natural coarse graining for simulating large biomolecular motion. *Biophysical Journal* Sep;2006 91(6):2115–2120. [PubMed: 16815893]
 38. Schuyler, Adam D.; Jernigan, Robert L.; Qasba, Pradman K.; Ramakrishnan, Boopathy; Chirikjian, Gregory S. Iterative cluster-NMA: A tool for generating conformational transitions in proteins. *Proteins: Structure, Function, and Bioinformatics* Feb;2009 74(3):760–776.
 39. Tirion, Monique M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters* 1996;77(9):1905–1908. [PubMed: 10063201]
 40. Ahmed, Aqeel; Gohlke, Holger. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins: Structure, Function, and Genetics* Jun;2006 63(4):1038–1051.
 41. Kondrashov, Dmitry A.; Cui, Qiang; Phillips, George N. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophysical Journal* Oct;2006 91(8):2760–2767. [PubMed: 16891367]
 42. Haliloglu, Turkan; Bahar, Ivett; Erman, Burak. Gaussian dynamics of folded proteins. *Physical Review Letters* 1997;79(16):3090–3093.
 43. Bahar, Ivett; Atilgan, Ali Rana; Demirel, Melik C.; Erman, Burak. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Physical Review Letters* 1998;80(12):2733–2736.
 44. Hinsen, Konrad. Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. Chapman & Hall/CRC; 2006. p. 1-16. *Mathematical and Computational Biology Series* chapter 1
 45. DeLano, WL. The PyMOL molecular graphics system. 2007. URL <http://www.pymol.org>

Figure 1. WT DHFR and Mutations

Cartoon representation of WT structure: DHFR (blue ribbon and mesh), FOL (red spheres), NADP (orange spheres), each mutation site (yellow sticks). The complex is compact and difficult to view as a stationary image. The main observable features are the spatial distribution of the mutation sites relative to the bound ligands. (Image prepared with PyMOL [45].)

Figure 2. cNMA Schematic

A) External forces suggest a rotation (blue arrows). In all-atom NMA, each of the atoms follows the tangential force (red) and extrapolation of that force pulls the atoms apart. B) Point mass, coarse-grained NMA collapses the representation; all forces (red) cancel and no motion is detected. C) In cNMA, the rotational motion is now captured and the cluster (yellow) maintains its local geometry.

Figure 3. Dual Plots for WT DHFR (interaction models on top, normalized degree functions on bottom)

These plots are part of the CPM interaction family defined in Equation 8. The $s = 1$ plot reproduces the DCM model and the increasing path lengths show the evolution of contact paths. The vertical axes of the interaction models and the horizontal axes of the degree functions are labeled with atomic indices. The horizontal and vertical blue lines partition the plots into DHFR, FOL and NADP. The normalized degree function indicates average atomic connectivity.

Figure 4. Ligand Clustering

The atoms of NADP and FOL are clustered according to the sets defined to the right of each structure illustration. The sets use atom labels as defined in the PDB entries. This clustering is based on a qualitative assessment of structure rigidity.

Figure 5. Decomposition Matrix

This plot is a direct comparison of the DHFR mutant modes and the WT modes. The black indicates mode alignment and fades to white with increasing dissimilarity. The strong diagonal indicates similarity between the mode sets and the spread indicates variations in dynamics.

Figure 6. Mode Regulation

Plots of the shift function for $\mathcal{W}$ projected onto $\mathcal{M}$ (top) and $\mathcal{M}$ projected onto $\mathcal{W}$ (bottom). The horizontal dashed lines mark the transition from down-regulated to up-regulated modes. The mean values and standard deviations are 1.03 and 0.08 for the top plot and 1.01 and 0.08 for the bottom plot. There are 137 $\mathcal{W}$ modes that are up-regulated more than one standard deviation and 162 that are down-regulated more than one standard deviation. There are 133 $\mathcal{M}$ modes that are up-regulated more than one standard deviation and 118 that are down-regulated more than one standard deviation.

Figure 7. Mode Residuals After Reduction Procedure

This plot shows the alignment of each structure's reduced mode with its original cNMA mode. Only modes which retain a $> 90\%$ residual are included in the motion correlation analysis. The GSOP correctly targets the lowest modes: 79% and 66% of the lowest 200 modes are included for the WT and the mutant, respectively.

Figure 8. Motion Correlation Matrices for WT and Mutant DHFR

The up- and down-regulated mode sets are each used to produce a motion correlation plot for each structure. These correlations are computed based on the cNMA motions of all atoms in each structure, but for clarity, the plots show only the alpha-carbon data for DHFR and all atom data for the bound ligands (these structure components are labeled on the x-axes and marked with vertical and horizontal lines across the plots). The average values for each atom's positive (negative) correlations is plot as a red (blue) curve under each correlation matrix. The most positive (red peaks) and negative (blue valleys) indicate the atoms most involved in correlated motion.

Figure 9. Static Method: Predicted Control Sites

The predicted control sites produced by the static method applied to the WT and mutant structures are diagramed (rectangles spanning the corresponding index ranges) in comparison to the known control sites. The blue bands labeling the known control sites are extended up as a background for the other plots. The T1 and T2 tests are applied to each of the data sets and the significance of the prediction sets is shown using the following symbols: * $\rightarrow p < 0.05$, ** $\rightarrow p < 0.005$. The yellow highlighting marks significant findings.

Figure 10. Dynamic Method: Predicted Control Sites

The predicted control sites produced by the dynamic method applied to each of the 8 parameter combinations of Figure 8, are diagramed (rectangles spanning the corresponding index ranges) in comparison to the known control sites. The blue bands labeling the known control sites are extended up as a background for the other plots. The T1 and T2 tests are applied to each of the data sets and the significance of the prediction sets is shown using the following symbols: * $\rightarrow p < 0.05$, ** $\rightarrow p < 0.005$. The yellow highlighting marks significant findings.