An Efficient Method for Quantitative Delineation of How Protein Breathing Motions Open Ligand Migration Channels

Hyuntae Na,* and Guang Song*†§

*Department of Computer Science,
†Program of Bioinformatics and Computational Biology,
§L. H. Baker Center for Bioinformatics and Biological Statistics,
Iowa State University, Ames, IA 50011, USA
Email: htna, gsong@iastate.edu

***Abstract*—Ligand migration and binding are central to the biological functions of many proteins such as myoglobin (Mb). However, there are usually no open channels for a ligand to enter the binding site at the static structure. It is widely accepted that protein breathing motions open these channels dynamically. However, how a protein exerts its control over the opening and closing of these channels through its intrinsic dynamics is not fully understood. Specifically, a quantitative delineation of the breathing motions that are needed to open a given channel is lacking.**

In this work, we present a novel computational method that is able to predict what combinations of normal modes are needed to open a channel. The method provides a complete mapping of ligand migration channels and the actual conformation changes (protein transition pathways) needed to open them. It is able to identify channels that open rarely and require slow dynamics, in which existing methods have more difficulty. Results on myoglobin demonstrate the effectiveness of the method.

I. Introduction

Proteins are one of the fundamental functional units in cells. It is fascinating how proteins exercise precise controls in different functions. Among these, a particular feat is seen in how a protein regulates the ins and outs of small ligands through its matrix. This process is of paramount importance in the proper function of many proteins, such as many enzymes whose efficient catalysis relies directly on the uptake of O_2 or other gaseous molecules. However, the O_2 update mechanism employed by these proteins is poorly understood. As there is often no open channels for ligands to enter into or exit from the interior of a host protein at the static structure, protein dynamics has been thought to open the channels dynamically but it is not fully clear how it does so in many proteins.

Experimentally, flash photolysis and mutagenesis studies are often employed to study the recombination kinetics in heme proteins and to identify ligand migration channels [3, 25, 12, 26, 15]. For example, site-directed mutagenesis of 27 residues was used to map out the ligand pathways [26]. Random mutagenesis studies conducted by Huang and Boxer [15] showed that single mutations of several other clusters of residues far away from the pathways were discovered to profoundly affect the ligand-binding kinetics. Time-resolved X-ray crystallography [28, 27] literally allows one to trace a photo-disassociated ligand as it migrates through the protein, as well as structure relaxation, over a broad range of timescales, from a few nanoseconds to as long as a few milliseconds [24, 7, 6, 23]. It has provided direct insight about how the correlated motions of the backbone and side chains provide a gating mechanism for ligand migration [24].

Computationally, molecular dynamics (MD) has been extensively applied to study ligand migration since late 70's [9, 11, 21, 16, 2, 5, 4]. A recent work by Ruscio et al. [22] obtained a cumulative 7-μs simulations on myoglobin and identified many different trajectories and entry/exit portals on the protein surface. The advantage of using MD is that one can observe actual events of ligand passing in and out of the protein matrix. Its drawback is that it takes extensive time to run the simulations and the process is stochastic. As a result, the less frequently traveled channels are difficult to identify experimentally. Implicit ligand sampling (ILS) [12, 10], an innovative approach developed recently, on the other hand, computes the potential of mean force (PMF) corresponding to the placement of a ligand everywhere inside a protein. ILS provides a complete three-dimension map that identifies the potential cavity sites and the pathways connecting them, some of which are in regions that are difficult to probe experimentally.

Proteins have intrinsic dynamical behaviors that contribute directly to their functions [14]. Most of these dynamical behaviors are captured by its normal modes of motions [13, 8, 17]. For many proteins such as myoglobin, these motions, often called breathing motions, open the channels. However, a quantitative delineation of exactly what breathing motion open a given channel is lacking. In this work we present a novel method that is able to determine exactly what combinations of the intrinsic normal modes may be used to open a channel. The method has two key steps: given a single structure of the protein to be studied, the first step is to apply Voronoi diagram to estimate where the putative ligand migration channels are. This efficient step ($O(n \log(n))$ time) can quickly identify the putative channels [18]. The second step is to gradually stretch open each channel by identifying and applying the best combination of normal modes (see Methods section for

details). The product of this stretching process is a sequence of conformation changes that eventually lead to the full opening of the channel.

The strengths and weaknesses of our method are summarized in Table I in comparison with two commonly used methods for channel mapping and identification, MD and ILS [12, 10]. Compared to ILS, our method is superior in that i) it provides a quantitative description of what combinations of normal modes are needed to open a channel, and ii) it identifies the key residues whose motions contribute the most in opening the channels. Compared to MD, the advantage of our method is that i) it does not require simulations, and ii) the conformation changes needed to open a channel are fully related to the channel opening and are not tangled with other thermal motions of the protein that are irrelevant to the channel opening. This separation may be critical in identifying key motions and key residues that regulate the channels. Lastly, ILS and MD are both simulation-based and suffer an intrinsic limitation that simulation-based approaches share: the narrow range of sampling. Thus, "the effect of slow conformational and allosteric changes will not be observed during the course of the simulation. Therefore, there is no guarantee that all biologically relevant pathways will be discovered through simulation." [10] Our method is normal-mode based and thus can identify channels that open rarely and require slow conformation changes. The weakness of our method is that it does not consider the interaction between the ligand and the protein. It is thus more suitable for studying the migration of small ligands.

To illustrate the effectiveness of our method, we apply it to myoglobin. We will show how the method allows one to determine the breathing motions needed to open each of the channels, to identify the residues that play a key role in regulating some of the channels, and to determine the contributions of global backbone motions and local side chain motions to a channel's opening.

II. Methods

To quantitatively determine the most plausible breathing motions that gradually open up a ligand channel, we first define the constraints on the breathing motions that guarantee the radius of the channel continually increases by a small amount until it is fully opened. We then define the criterion for selecting the most plausible breathing motion among those that satisfy the constraints.

A. Constraints on the Breathing Motions that Gradually Open a Channel

1) Definitions: First let us define channels and the radius of a channel in mathematical terms. As in [18], given an input protein structure, we first compute its Voronoi diagram. Protein cavities are denoted by Voronoi vertices that are inside the protein and have large enough clearance. If we think a channel as a sequence of cylindrical pipes connecting together, the axis of the channel is represented by consecutive Voronoi edges, or line segments, that connect the internal cavities to solvent. A channel may have different clearances at different segments of the channel. According to Voronoi diagram computation, each line segment, or Voronoi edge, matches to three atoms

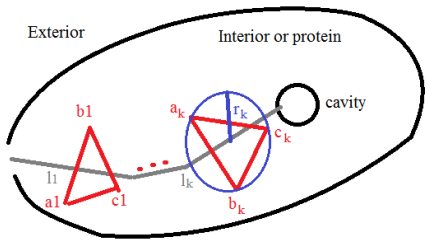


Fig. 1. Illustration of a channel. The channel from solvent to an internal cavity, is represented as a series of line segments, $l_1, ..., l_k$, or Voronoi edges. Each channel/line segment (or Voronoi edge) l_i matches to three atoms, a_i, b_i, c_i, where $1 \leq i \leq k$. The atoms represent lining atoms of the channel and define the channel clearance, which is the circumradius of the triangle formed by the three atoms.

that have equidistance to the edge. This distance represents the minimum clearance at this segment of the channel. As a channel is composed of a sequence of channel segments, the channel radius or channel clearance is defined as the smallest clearance of all channel segments. The location where the clearance is the smallest is the bottleneck of the channel. Let $l_1, ..., l_k$ be the axes of the consecutive segments of a channel, a_i, b_i, c_i be the three atoms corresponding to the Voronoi edge l_i, and r_i be the circumradius of triangle $\triangle a_i, b_i, c_i$, $1 \leq i \leq k$, see Fig. 1. The radius of the channel, denoted by a capital R, is the smallest radius of all the circumradii, i.e., $R = \min(r_1, ..., r_k)$.

2) The Derivative of Channel Clearance with Respect to Normal Modes.: Denote by a, b, and c the three atoms of the channel segment l, and by $\boldsymbol{p}_i$ the Cartesian coordinate of atom i. The circumradius $r_{a,b,c}$ of the triangle $\triangle abc$ is determined using the trigonometry:

$$r_{a,b,c} = \frac{\|\boldsymbol{p}_{a,b}\| \cdot \|\boldsymbol{p}_{b,c}\| \cdot \|\boldsymbol{p}_{c,a}\|}{2\|\boldsymbol{p}_{a,b} \times \boldsymbol{p}_{b,c}\|}, \tag{1}$$

where $\boldsymbol{p}_{a,b} = \boldsymbol{p}_a - \boldsymbol{p}_b$, $\|\boldsymbol{p}\|$ is the norm of vector $\boldsymbol{p}$, and $\boldsymbol{a} \times \boldsymbol{b}$ is the cross product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

Let $\boldsymbol{w}_j$ be j^{th} normal mode of the protein, and $\boldsymbol{w}_{ja}$ its component for atom a. For an instantaneous movement of $t_j \boldsymbol{w}_j$, where t_j is a small scalar value, the new position of the atom a can be written as $\boldsymbol{p}_a + t_j \boldsymbol{w}_{ja}$. In a similar manner, the new positions of atom b (and c) can be obtained. Now, let $\mathrm{r}(t_j)$ be the circumradius (see Eq. (1)) of triangle $\triangle abc$, as a function of t_j. The first derivative of radius r with respect to mode $\boldsymbol{w}_j$ can be written as:

$$\begin{aligned} \dot{r}_{\boldsymbol{w}_j} &= \frac{\partial \mathrm{r}(t_j)}{\partial t_j} = \frac{\partial \mathrm{r}(t_j)}{\partial \mathrm{r}(t_j)^2} \cdot \frac{\partial \mathrm{r}(t_j)^2}{\partial t_j} \\ &= \frac{1}{4r_{a,b,c}} \cdot \frac{f(a,b,c) + f(c,a,b) + f(b,c,a)}{\|\boldsymbol{p}_{ab} \times \boldsymbol{p}_{bc}\|^2} \\ &- r_{a,b,c} \cdot \frac{(\boldsymbol{p}_{ab} \times \boldsymbol{p}_{bc})^{\mathsf{T}} (\boldsymbol{p}_{ab} \times \boldsymbol{w}_{jb,jc} + \boldsymbol{w}_{ja,jb} \times \boldsymbol{p}_{b,c})}{\|\boldsymbol{p}_{ab} \times \boldsymbol{p}_{bc}\|^2}, \end{aligned} \tag{2}$$

where $\boldsymbol{a}^{\mathsf{T}}$ is the transpose of a vector $\boldsymbol{a}$, $f(a,b,c) = ((\boldsymbol{w}_{ja} - \boldsymbol{w}_{jb})^{\mathsf{T}} \boldsymbol{p}_{a,b}) \|\boldsymbol{p}_{a,c}\|^2 \|\boldsymbol{p}_{b,c}\|^2$, and $\boldsymbol{w}_{ja,jb} = \boldsymbol{w}_{ja} - \boldsymbol{w}_{jb}$. $\dot{r}_{\boldsymbol{w}_j}$ represents the rate of change in radius r with respect to the movement along mode $\boldsymbol{w}_j$. A large (small) $\dot{r}_{\boldsymbol{w}_j}$ for a particular mode means the channel radius quickly (slowly) increases or decreases as the protein makes a breathing motion along that

	our method	**ILS**	**MD**
Method basis (require simulations?)	normal modes (no)	MD (yes)	MD (yes)
Completeness in mapping ligand channels	close to complete	close to complete	not complete
Transition pathways that open the channels	Yes	No	Yes
Identify normal modes contribute the most	Yes	No	No
Running time	short to medium	short to medium	long
Ligand size	small	small	small or large
Channel prediction quality	estimate	estimate	more realistic
Ligand-protein interactions	not considered	not considered	considered

TABLE I
A COMPARISON BETWEEN OUR METHOD AND THE TWO OTHER WELL-KNOWN COMPUTATIONAL METHODS FOR LIGAND CHANNEL MAPPING AND IDENTIFICATION, NAMELY IMPLICIT LIGAND SAMPLING (ILS) [12, 10] AND MOLECULAR DYNAMICS (MD).

mode. As the protein fluctuates, the circumradius $\mathrm{r}(t_j)$ at $\triangle abc$ is approximated as:

$$\mathrm{r}(t_j; a, b, c) = r_{a,b,c} + \dot{r}_{\boldsymbol{w}_j} \cdot t_j. \tag{3}$$

Now consider the effect of a movement that involves a combination of m modes, i.e., $\sum_{i=1}^{m} t_j \boldsymbol{w}_j$. The circumradius r should become:

$$\mathrm{r}(t_1, ..., t_m; a_i, b_i, c_i) = r_{a_i,b_i,c_i} + \sum_{i=1}^{m} \dot{r}_{\boldsymbol{w}_j|i} \cdot t_j, \tag{4}$$

where r_{a_i,b_i,c_i} is the circumradius of i^{th} triangle $\triangle a_i b_i c_i$, and $\dot{r}_{\boldsymbol{w}_j|i}$ represents $\dot{r}_{\boldsymbol{w}_j}$ for $\triangle a_i b_i c_i$.

3) The Needed Constraints on Normal Modes that Gradually Open a Channel: Now we determine what combination of normal modes are needed to gradually open an initially closed channel. Eq. (4) specifies how channel radii change as a result of mode motions. Our goal is to find the right combination of t_js such that the channel radius is guaranteed to increase by a certain small amount at each iteration step.

Assume that a channel is composed of k consecutive channel segments, and that each channel segment i is lined by three atoms a_i, b_i, and c_i, where $1 \leq i \leq k$. Recall that the channel radius R is the smallest radius of all channel segments:

$$R = \min_{1 \leq i \leq k} r_{a_i,b_i,c_i}. \tag{5}$$

After a small movement of $\sum_{i=1}^{m} t_i \boldsymbol{w}_i$, the channel radius becomes:

$$R_{\text{after}} = \min_{1 \leq i \leq k} \mathrm{r}(t_1, ..., t_m; a_i, b_i, c_i). \tag{6}$$

Now we require that at each iteration, the channel radius increases by a small amount of s, which is a model parameters, i.e.,

$$R_{\text{after}} - R \geq s. \tag{7}$$

Let $\boldsymbol{t} = (t_1, ..., t_m)^{\mathsf{T}}$ be the vector form of the mode motion, and $\dot{\boldsymbol{r}}_i = (\dot{r}_{\boldsymbol{w}_1|i}, \dot{r}_{\boldsymbol{w}_2|i}, ..., \dot{r}_{\boldsymbol{w}_m|i})^{\mathsf{T}}$. The set of $\boldsymbol{t}$ that increases the channel radius by s satisfies the following constraints:

$$\dot{\boldsymbol{r}}_i^{\mathsf{T}} \boldsymbol{t} \geq u_i, \tag{8}$$

where $1 \leq i \leq k$, and

$$u_i = \min(r_{a_1,b_1,c_1}, ..., r_{a_k,b_k,c_k}) + s - r_{a_i,b_i,c_i}. \tag{9}$$

B. Selecting the Most Plausible Breathing Motion

The constraint in Eq. (8) specifies what breathing motions can gradually open a channel. There exist many combinations of modes satisfying this constraint. Which one should we choose? Which criterion should we follow to identify the most plausible breathing motion? One apparent choice is to minimize the amount of work required to open a channel. The optimal breathing motion that can open a channel should be the one that takes the least effort. Another consideration, which is less obvious but very necessary, is to realize that the derivation of Eq. (8) requires that the channel opening process should take place via many small, ideally infinitesimal, steps. Therefore, in the process of searching for the optimal breathing motion, we have also to require that the magnitude of the motion at each iteration step, which is $\|\boldsymbol{t}\|$, be small.

Denote E, $\boldsymbol{f}$, and H as the potential energy, the force, and the Hessian matrix of the current conformation, respectively. Denote W as the matrix form of m modes: $W = (\boldsymbol{w}_1, ..., \boldsymbol{w}_m)$. A protein movement can be written as $W\boldsymbol{t}$, where $\boldsymbol{t}$ is the vector form of the m modes' contributions and is $\boldsymbol{t} = (t_1, ..., t_m)^{\mathsf{T}}$. The potential energy change δE due to movement $W\boldsymbol{t}$ can be approximated to the second order as:

$$\delta E = \boldsymbol{f}^{\mathsf{T}} W \boldsymbol{t} + \frac{1}{2} \boldsymbol{t}^{\mathsf{T}} W^{\mathsf{T}} H W \boldsymbol{t}. \tag{10}$$

Denote $\widetilde{H} = W^{\mathsf{T}} H W$. In this work, we use all the torsional modes solved by TNM [20] for W. H is the full Hessian matrix written in the Cartesian space. In general $\boldsymbol{w}_i$ is not an eigenvector of H. Let Λ and V be the matrix forms of the eigenvalues and eigenvectors of $\widetilde{H}$, respectively, i.e., $\widetilde{H} = V \Lambda V^{\mathsf{T}}$.

Now define,

$$\boldsymbol{t}^* = |\Lambda|^{1/2} V^{\mathsf{T}} \boldsymbol{t}; \tag{11}$$

$$\boldsymbol{f}^* = |\Lambda|^{-1/2} V^{\mathsf{T}} W^{\mathsf{T}} \boldsymbol{f}, \tag{12}$$

where $|\Lambda|$ is a matrix whose elements are the absolute values of the elements in Λ, and $|\Lambda|^{1/2}$ is the square root of matrix $|\Lambda|$. Note Λ is a diagonal matrix. Let $\mathrm{sign}(\Lambda)$ be a matrix whose elements take the signs of the elements of Λ. We have, $\Lambda = |\Lambda|^{1/2}\ \mathrm{sign}(\Lambda)\ |\Lambda|^{1/2}$.

Using transformed variables $\boldsymbol{t}^*$ and $\boldsymbol{f}^*$, Eq. (10) becomes:

$$\delta E = \boldsymbol{f}^{*\mathsf{T}} \boldsymbol{t}^* + \frac{1}{2} \boldsymbol{t}^{*\mathsf{T}}\ \mathrm{sign}(\Lambda)\ \boldsymbol{t}^*. \tag{13}$$

Note that $\boldsymbol{t}^*$ (see Eq. (11)) is scaled by the square roots of the eigenvalues (i.e., $|\Lambda|^{1/2}$), and $\boldsymbol{f}^*$ is the generalized force

and is the gradient of the potential δE with respect to $\boldsymbol{t}^*$. The advantage of using $\boldsymbol{t}^*$ over $\boldsymbol{t}$ is that the preference for lower frequency modes is naturally taken into account and the components of $\boldsymbol{t}^*$ can now be treated equally when finding the optimal $\boldsymbol{t}^*$. Λ as a diagonal matrix contains the eigenvalues of the Hessian matrix H in the mode space defined by W. sign(Λ) represents the signs of these eigenvalues. When the conformation is at a local energy minimum, all the eigenvalues are positive and sign(Λ) becomes an identity matrix. At other places, such as at a saddle point, some eigenvalues may be negative and sign(Λ) may contain some -1's along the diagonal.

Recall our discussion in the beginning of this section that in finding the most plausible breathing motion that can open a channel, we have two considerations. One is to find a motion that takes the least work, and the other is to have a small $\|\boldsymbol{t}\|$ or $\|\boldsymbol{t}^*\|$ at each iteration. Only minimizing the work without requiring the magnitude of $\boldsymbol{t}$ being small may result in large, unrealistic moves. To allow searching in all directions in the mode space and yet to bias the search towards the direction of the potential energy's gradient descent, we set the search space to be of the shape of an ellipsoid, whose one focus is the origin of the mode space for $\boldsymbol{t}^*$ and whose other focus is in the direction of potential energy gradient descent. The eccentricity of the ellipsoid controls the degree of the biasedness towards potential energy's gradient descent. An eccentricity of 0 means no bias, while an eccentricity of 1 means the search is fully along the direction of potential energy's gradient descent. Eccentricity is a model parameter in this work. Experiments show that an eccentricity of 0.7 works well. The function of such an ellipsoid is:

$$\|\boldsymbol{t}^*\|(1 - e \cdot \cos(\theta)) = const, \tag{14}$$

where e is the eccentricity, θ is angle between the direction $\boldsymbol{t}^*$ and the direction of potential energy's gradient descent, i.e.,

$$\cos(\theta) = -\frac{\boldsymbol{f}^{*\mathsf{T}}\boldsymbol{t}^*}{\|\boldsymbol{f}^*\| \cdot \|\boldsymbol{t}^*\|}. \tag{15}$$

Now the search for the most plausible breathing motion $\boldsymbol{t}^*$ that opens a channel becomes an optimization problem:

$$\underset{\boldsymbol{t}^*}{\operatorname{argmin}} \ \|\boldsymbol{t}^*\|^2(1 - e \cdot \cos(\theta))^2. \tag{16}$$

The above optimization is subject to the constraints given in Eq. (8), which can be rewritten, using the new variable $\boldsymbol{t}^*$, as follows:

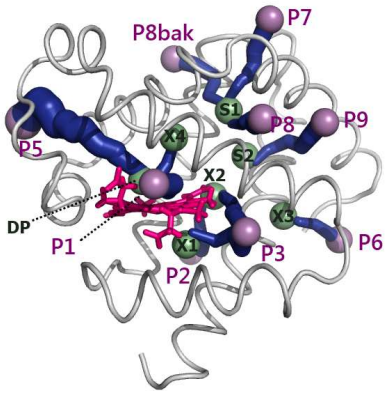
$$(\boldsymbol{r}_i^*)^{\mathsf{T}}\boldsymbol{t}^* \geq u_i, \tag{17}$$

where

$$\boldsymbol{r}_i^* = |\Lambda|^{-1/2}V^{\mathsf{T}}\dot{\boldsymbol{r}}_i. \tag{18}$$

C. The Iterative Procedure for Opening Up A Channel

In this section, we summarize the iterative steps for opening up a channel. Given a protein structure, it is first energetically minimized using the CHARMM22 force field [19]. From the minimized structure, Hessian matrix H is computed. Voronoi diagram are then computed and cavities and putative ligand migration channels are identified.



(a)

label of channel	initial clearance
DP-P1	0.60
DP-P5	0.53
X1-P3	0.74
X2-P2	0.88
X2-P3	0.90
X3-P6	0.60
X4-P1	0.60
X4-P5	0.53
S1-P7	0.49
S1-P8	0.55
S1-P8bak	0.49
S2-P9	1.04

(b)

Fig. 2. The channels in myoglobin. X1 to X4 stands for the four xeon binding sites. S1 and S2 are two cavities identified in MD simulations [4]. DP stands for distal pocket. The surface portals are labeled as P1, P2, etc.. (a) A cartoon image of myoglobin (pdbid: 1A6G) overlaid with cavities (green spheres), portals (red labels), and channels (rugged blue tubes). (b) The channels between cavities and solvent and their initial radii (in parentheses).

Algorithm 1 lists the rest of the steps that follow. The algorithm receives as input the initial protein conformation $\boldsymbol{p}_0$ (i.e., the minimized structure), a putative channel path L_0, the targeting channel radius h required for an open channel, and the radius increase s at each iteration. In the algorithm, channel L_0 is gradually opened by applying the most plausible breathing motion until the channel radius reaches h, and the final conformation will be returned. In each iteration, the mode matrix W are determined by using the torsional network model (TNM) [20]. The determined $\boldsymbol{t}^*$, and therefore $\boldsymbol{t}$, is then used to update the conformation. There are two major advantages in using torsional modes: i) it reduces the number of degrees of freedom; ii) it avoids sharp potential energy increases caused by distorted protein geometry.

Algorithm 1 ExpandChannel($\boldsymbol{p}_0, L_0, h, s$)

$\boldsymbol{p} \leftarrow \boldsymbol{p}_0, \quad L \leftarrow L_0$
$M \leftarrow \{\langle a_i, b_i, c_i\rangle \mid$ triangle atoms a_i, b_i, c_i of channel segment $l_i \in L\}$
while $\min(r_{a_1,b_1,c_1}, r_{a_2,b_2,c_2}, ...) < h$ **do**
  Compute force $\boldsymbol{f}$ and Hessian matrix H of conformation $\boldsymbol{p}$
  Compute modes $W = (\boldsymbol{w}_1, ..., \boldsymbol{w}_m)$ using TNM [20].
  Determine constraints $\boldsymbol{r}_i^{\mathsf{T}}\boldsymbol{t} \geq u_i$ according to (8), for all $\langle a_i, b_i, c_i\rangle \in M$
  Find the optimal $\boldsymbol{t}^*$ according to (12), (15) - (18)
  $\boldsymbol{t} \leftarrow V|\Lambda|^{-1/2}\boldsymbol{t}^*$
  Update conformation: $\boldsymbol{p} \leftarrow \boldsymbol{p} + W\boldsymbol{t}$
  Update channel segments of L using the new $\boldsymbol{p}$
  Update M.
end while

III. Results

In this section, we apply the proposed method to myoglobin to determine what and how breathing motions open each of its channels.

1A6G.pdb is used as the initial structure and is energetically minimized in vacuum using the Tinker [1] software and CHARMM22 force field [19]. The minimized structure is then

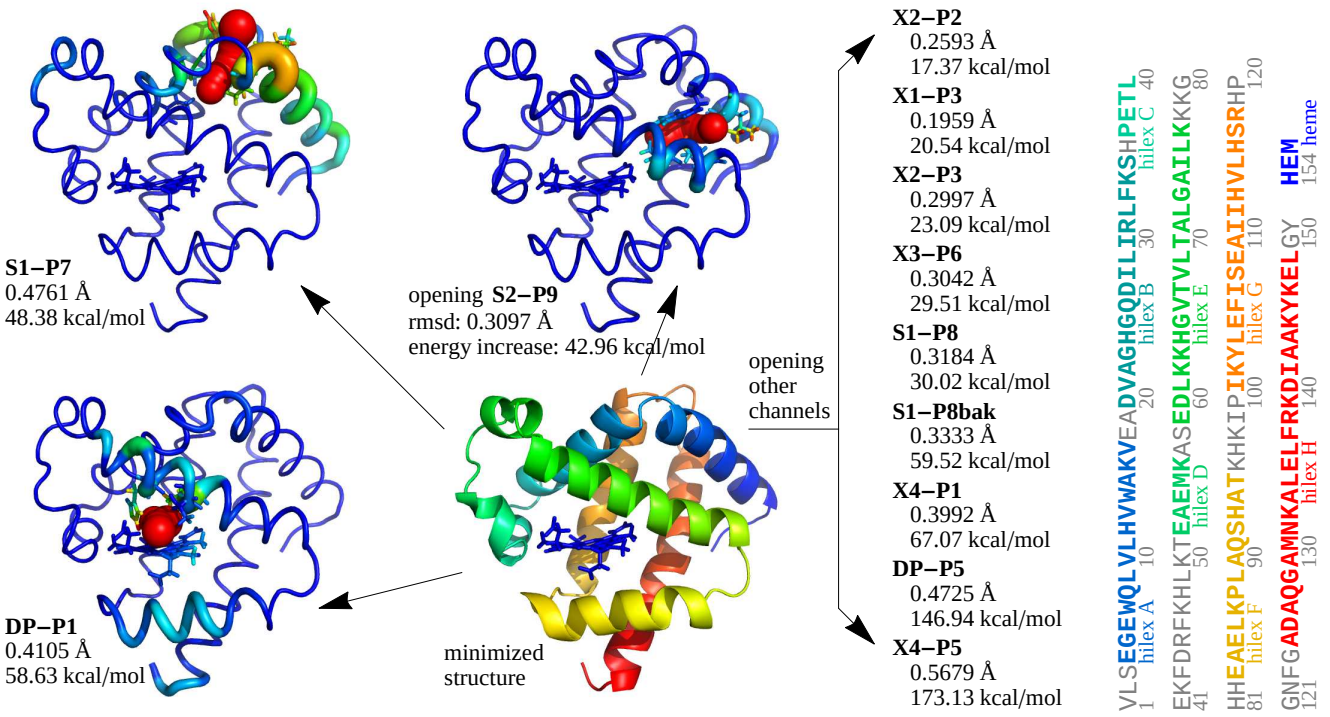


Fig. 3. By applying our method to the minimized structure (center), different conformation changes (shown by the arrows) needed to open the channels are obtained. For three of the channels, the final conformations at which the channels are fully open are given. The thickness along the backbone trace shows the magnitude of the motions of the residues as the channels open up. The RMSD deviations between the final conformations and the minimized structure and their potential energy differences are also given. The primary sequence of myoglobin and its secondary structures (helices A to H) are given on the right.

used as the input conformation $\boldsymbol{p}_0$ in Algorithm 1. Then, putative ligand channels L_0 are determined using the Voronoi diagram [18]. Specifically, any channel that is between a cavity and the solvent and whose clearance is greater than a given threshold is considered as a putative channel. For comparison purpose, putative channels that match to the channels identified in [22] are listed in Fig. 2(b), along with their initial clearances at the minimized structure. Fig. 2(a) shows the cartoon image of the cavities and these channels. All the channels on the list are initially closed, but as the protein "breathes", they should open up dynamically. Included in the list is also the S1-P8bak channel that was not identified by MD [22]. It costs more energy to open channel S1-P8bak than the other two that are connected to cavity S1, i.e., S1-P7 or S1-P8 (see Fig. 5). This may be the reason why this channel was not observed in the MD simulation [22].

In our experiment, every putative channel identified through Voronoi diagram is iteratively expanded using the Algorithm 1 until its clearance reaches up to a preset threshold of 1.8Å, which is enough for a gaseous ligand molecule to pass through. The amount of work needed to open each channel is also computed and is used to estimate the relative likelihood that a channel may be used by a ligand to enter into or exit from the protein matrix.

A. Quantitative Delineation of What Breathing Motions Open Ligand Migration Channels

To determine the most plausible breathing motions that open the ligand migration channels of myoglobin that are listed in Figure 2, we apply our channel stretching procedure to each of these channels. The procedure identifies the best combination of normal modes that are able to gradually open up the channel at each step. This sequence of steps from the initial conformation to the conformation where the channel is fully open represents the transition pathway needed to open the channel. The rocking motion along this pathway provides a quantitative description of what breathing motions are needed and how they open and close the channel.

Figure 3 shows how the minimized structure (center) undergoes different conformation changes to open each and every channel. The RMSD deviations between the final conformations and minimized structures are given in the figure, which are on average less than 0.5 Å. Also shown in the figure are the potential energy differences between final conformations and the minimized structure. These values provide an estimate of the relative likelihood that a channel may be used by a ligand to enter into or exit from the protein matrix.

All the final conformations (in PDB format) at which one of the channels are open, the transition pathways (in PDB format) needed to open each and every of these channels as well as movies that display the breathing motions needed to open all these channels are available at http://www.cs.iastate.edu/~gsong/CSB/channels/.

B. Identify Key Normal Mode Motions and Key Residues that Regulate the Channels

Since the modes selected at each iteration are torsional modes, the combination of these modes represent a displacement in the torsional space. The residues whose torsional angles (could be either backbone ϕ/ψ angles or side chain χ angles) display the largest change are those that contribute the most to the opening of the channels. Figure 4 shows the

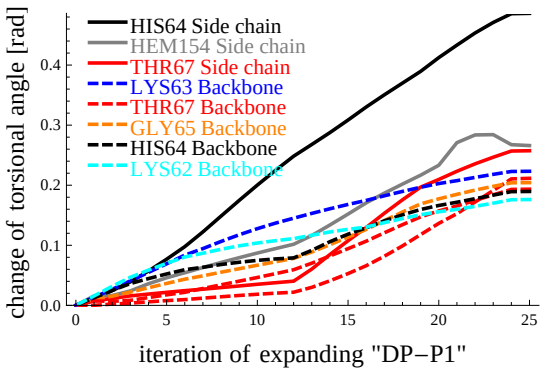


Fig. 4. The largest torsional angle changes in opining up channel DP to P1

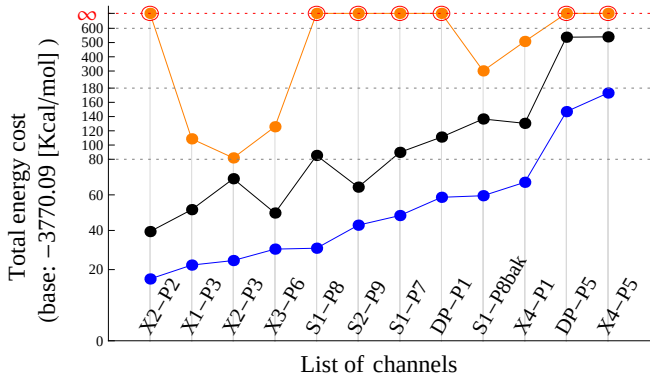


Fig. 5. Comparison of energy costs (y axis) in opening channels (listed in x axis) when allowing only backbone (black), only side chains (orange), or the both (blue). Channels are sorted by their energy costs on the blue line. Infinity ∞ energy cost represents that the channel cannot be opened.

accumulative torsional angle changes over time for the top few torsional angles that display the largest contributions to channel opening.

Channel DP to P1. This channel is the well-known HIS gate of myoglobin. It is clearly seen from Figure 4 that the largest contribution comes from the side chain swing motions of residue His64, the well-known HIS gate. Thr67, another lining residue of this channel, also makes measurable movements to accommodate the opening of this channel. The contributions of backbone motions, shown in the dashed lines, are much smaller.

C. Can the Channels Be Opened by Backbone Motions or Side-Chain Motions Alone?

The opening and closing of each channel is controlled by the interplay of backbone motions and side chain motions. To assess the necessity of side chain motions or backbone motions in opening a channel, we apply our method to investigate if the channels can be opened by either of them alone, and if so, what extra work is needed.

Fig. 5 compares the energy cost to open the 12 channels with only backbone motions (black), only side chain motions (orange), and the both (blue). Since the three approaches require different magnitude of work to open the channels, we divide the work required (shown in the y-axis in Figure 5) into four ranges with different scalings; the four ranges are separated by dotted lines. We denoted by the infinity ∞ the case in which a channel cannot be opened.

Results in Fig. 5 show that many channels cannot be opened by side chain motions alone (orange). The backbone motions (black) are capable to open each channel even though it would require much more work. Interestingly, the well-known HIS channel (DP to P1) cannot be fully opened by side chain motions alone. It is generally accepted that Histidine 64 plays a gating role in the opening of the channel. A plausible explanation is that backbone breathing motions are needed to create enough space for the sidechain of Histidine 64 to swing open.

IV. Conclusion

In this paper, we present a novel normal mode-based method that is able to provide a complete mapping of ligand migration channels of a given protein. The method has two key steps. The first step is to use the highly efficient Voronoi diagram to estimate putative ligand migration channels. The second step, which is the bulk of this work, is a method that determines the best combinations of normal modes that gradually open a channel.

There are three notable contributions in this method development. The first is to compute the derivative of a channel clearance relative to any given mode. This derivative describes how rapidly (or slowly) a breathing motion along a particular normal mode increases or decreases the clearance of a given channel. It is foreseeable that such derivatives relative to the normal modes may be used to study how other structural or functional properties of proteins depend on protein dynamics. To the best of our knowledge, this has not been done before. Secondly, finding the best combination of normal modes that gradually open a channel is rephrased as an optimization problem. Thirdly, the optimization function we used is innovative. When defining what is the best combination of normal modes that slowly open a channel, there are two attributes that we desire to minimize at each iteration: one is the work required and the other is the magnitude of the conformation change. How to minimize them both at the same time? Though it is desired that the work cost to stretch open a channel should be minimized, however, if we focus on minimizing the work cost alone, it may result in large unrealistic movements. Therefore, we need to somehow minimize the work cost and the magnitude of the motions (i.e., $\|\boldsymbol{t}\|$) at the same time. Our optimization function, which is of ellipsoidal form and has an adjustable eccentricity factor, provides a perfect balance between the two requirements.

Results on myoglobin clearly demonstrate the effectiveness of our method. Our method provides all the actual transition pathways (conformation changes of the protein) needed to open each and every channel. Transition pathways (in PDB format) and movies that display how breathing motions open the channels are available at http://www.cs.iastate.edu/~gsong/CSB/channels/. Such actual transition pathways are not available from other channel mapping methods such as ILS [10], and it would take a much longer time to obtain them using MD simulations. In addition, the method identifies some of key residues whose side chain motions or backbone motions contribute the most in regulating the channels.

The limitation of the current method is that it does not consider the potential interactions between the ligand and

the host protein. While this is acceptable to small gaseous ligands [12], care must be taken when applying this method to larger ligands or charged ligands (such as proton or metal ions), as in which case the electrostatic interactions between the ligand and the host protein may strongly affect of the normal modes of the host protein.

ACKNOWLEDGMENTS

Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged.

REFERENCES

[1] J. W. Ponder and F. M. Richards. An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.*, 8(7):1016-1024, 1987.

[2] M. Anselmi, A. Di Nola, and A. Amadei. The kinetics of ligand migration in crystallized myoglobin as revealed by molecular dynamics simulations. *Biophys J*, 94(11): 4277–81, 2008.

[3] R. H. Austin et al. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–73, 1975.

[4] C. Bossa et al. Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys J*, 86(6):3855–62, 2004.

[5] C. Bossa et al. Molecular dynamics simulation of sperm whale myoglobin: effects of mutations and trapped co on the structure and dynamics of cavities. *Biophys J*, 89(1): 465–74, 2005.

[6] D. Bourgeois et al. Complex landscape of protein structural dynamics unveiled by nanosecond laue crystallography. *Proc Natl Acad Sci U S A*, 100(15):8704–9, 2003.

[7] D. Bourgeois et al. Extended subnanosecond structural dynamics of myoglobin revealed by laue crystallography. *Proc Natl Acad Sci U S A*, 103(13):4924–9, 2006.

[8] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA*, 80(21): 6571–6575, November 1983. ISSN 0027-8424. doi: 10.1073/pnas.80.21.6571.

[9] D. A. Case and M. Karplus. Dynamics of ligand binding to heme proteins. *J Mol Biol*, 132(3):343–68, 1979.

[10] J. Cohen, K.W. Olsen, and K. Schulten. Finding gas migration pathways in proteins using implicit ligand sampling. *Methods in Enzymology*, 437:439–457, 2008. doi: 10.1016/S0076-6879(07)37022-5.

[11] R. Elber and M. Karplus. Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J Am Chem Soc*, 112:9161–9175, 1990.

[12] H. Frauenfelder et al. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc Natl Acad Sci U S A*, 98 (5):2370–4, 2001.

[13] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Nut1. Acad. Sci. USA*, 80(12):3696–3700, June 1983. doi: 10.1073/pnas.80.12.3696.

[14] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172): 964–972, December 2007. ISSN 1476-4687. doi: 10.1038/nature06522.

[15] X. Huang and S. G. Boxer. Discovery of new ligand binding pathways in myoglobin by random mutagenesis. *Nat Struct Biol*, 1(4):226–9, 1994.

[16] G. Hummer, F. Schotte, and P. A. Anfinrud. Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations. *Proc Natl Acad Sci U S A*, 101(43):15330–4, 2004.

[17] M. Levitt, C. Sander, and P. S. Stern. The Normal Modes of a protein: Native bovine Pancreatic Trypsin inhibitor. *Int. J. Quant. Chem*, 10:181–199, 1983.

[18] Tu-Liang Lin and Guang Song. Efficient mapping of ligand migration channel networks in dynamic proteins. *Proteins*, 79(8):24752490, August 2011. doi: 10.1002/prot.23071.

[19] MacKerell et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *J. Phys. Chem. B*, 102(18):3586–3616, April 1998. doi: 10.1021/jp973084f.

[20] Raul Mendez and Ugo Bastolla. Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins. *Physical Review Letters*, 104(22):228103+, June 2010. doi: 10.1103/physrevlett.104.228103.

[21] D. R. Nutt and M. Meuwly. Co migration in native and mutant myoglobin: atomistic simulations for the understanding of protein function. *Proc Natl Acad Sci U S A*, 101(16):5998–6002, 2004.

[22] J. Z. Ruscio et al. Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin. *Proc Natl Acad Sci U S A*, 105(27):9204–9, 2008.

[23] M. Schmidt et al. Ligand migration pathway and protein dynamics in myoglobin: a time-resolved crystallographic study on l29w mbco. *Proc Natl Acad Sci U S A*, 102 (33):11704–9, 2005.

[24] F. Schotte et al. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300 (5627):1944–7, 2003.

[25] E. E. Scott and Q. H. Gibson. Ligand migration in sperm whale myoglobin. *Biochemistry*, 36(39):11909–17, 1997.

[26] E. E. Scott, Q. H. Gibson, and J. S. Olson. Mapping the pathways for o2 entry into and exit from myoglobin. *J Biol Chem*, 276(7):5177–88, 2001.

[27] V. Srajer et al. Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. *Science*, 274(5293):1726–9, 1996.

[28] V. Srajer et al. Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved laue x-ray diffraction. *Biochemistry*, 40(46):13802–15, 2001.