Imaging Protein Statistical Substate Occupancy in a Spectrum-Function Phase Space

W. DeWitt and K. Chu*

Department of Physics, Cook Physical Sciences, University of Vermont, Burlington, Vermont 05405, USA

(Received 7 March 2010; published 24 August 2010)

Hemeprotein ligand rebinding studies reveal varying IR absorbance and rebinding functions across a cryogenic ensemble. Since IR-active vibrations and rebinding barriers couple to structural coordinates, spectral and functional heterogeneity arise from conformational heterogeneity. Modeling rebinding data as a spectrally resolved superposition of first-order rate processes and employing maximum entropy regularization, protein heterogeneity is imaged as an ensemble occupancy of a spectrum-function phase space. Results from myoglobin rebinding carbon monoxide are discussed.

DOI: 10.1103/PhysRevLett.105.098101 PACS numbers: 87.15.kp, 02.30.Zz, 89.70.Cf

Carbon monoxide (CO) undergoes a mid-IR-active stretch vibration when bound to the heme iron of myoglobin, making it the ligand of choice for spectroscopic studies of protein ligand-binding kinetics. Such studies, in conjunction with structural and computational work, have revealed a picture of proteins as dynamically complex molecules [1]. A given primary sequence can adopt any one of a very large number of similar structural conformations. Each conformation is associated with a local minimum in a high-dimensional conformational energy landscape (conformational substate). At physiological temperatures, molecules constantly undergo transitions between substates. At $T < T_g$, such transitions are frozen out so that an ensemble is characterized by static conformational heterogeneity [2].

Ligand binding at low temperature is characterized by two states along a reaction coordinate. The *A* state refers to heme-bound CO. Absorption of a visible photon breaks the covalent Fe-CO bond, leaving the ligand trapped within the frozen protein matrix, or *B* state. Geminate rebinding from the *B* state to the *A* state is nonexponential below 160 K, indicating that different structural conformations rebind at different rates [3]. Also, the CO stretch bands shift and broaden asymmetrically as rebinding progresses at low temperatures, a phenomenon known as kinetic hole burning (KHB) [4,5]. CO stretch frequency differs across conformational substates. An ensemble of carbonmonoxymyoglobin (MbCO) is characterized by a distribution of CO rebinding barriers and a distribution of CO stretch peaks, and these distributions map to each other nonrandomly.

The IR absorbance spectrum of the commonly studied sperm whale MbCO shows three distinct peaks at neutral *p*H; the *A* states A_0, A_1, and A_3. All three display nonexponential kinetics and KHB. This evinces a hierarchy of minima in the energy landscape, with each of the three substates at the first tier (taxonomic substates) separated by higher barriers and composed of many second-tier substates (statistical substates) separated by relatively smaller barriers. Each statistical substate (SS) is characterized by a definite enthalpy barrier to CO rebinding and a definite CO stretch peak frequency. Within each taxonomic substate (TS), the distribution of rebinding barriers and CO stretch peaks is determined by the occupancy of its many SS. In this study we confine our attention to the occupancy of the SS within a single TS. Experimental data were taken using horse myoglobin, which occupies a single TS [6,7].

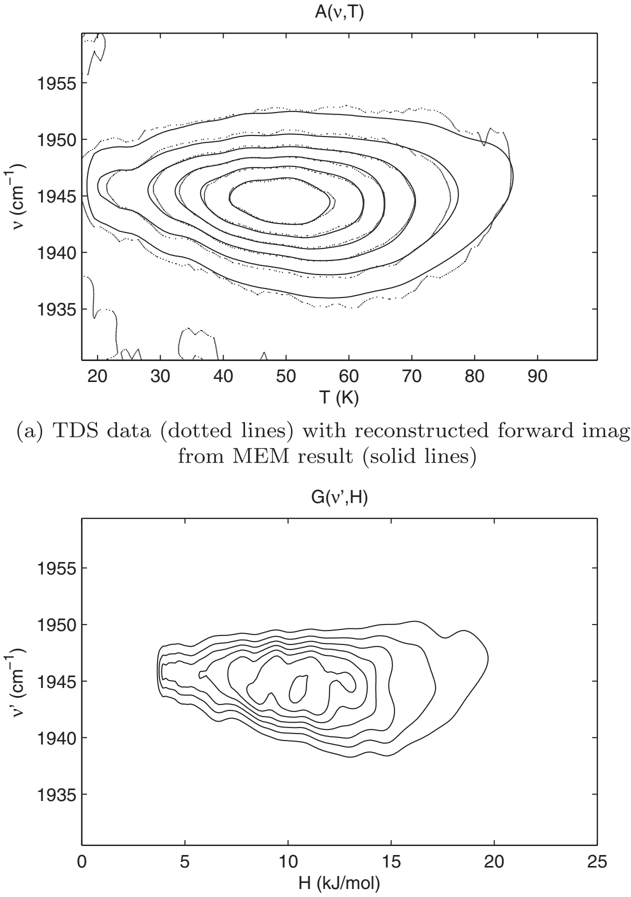
Temperature derivative spectroscopy (TDS) is an experimental protocol that samples the distribution of rebinding enthalpy barriers by measuring rebinding across a range of temperatures [8]. Following photolysis at low temperature, spectra are taken continuously as the temperature is ramped linearly ($T = T_i + \beta t$). The differences between consecutive spectra are calculated, producing the TDS surface $\Delta\mathcal{A}(\nu, T)$ which is approximately proportional to the temperature (or time) differentiated CO absorbance spectrum. Spectrally integrating the TDS surface gives a curve proportional to the temperature differentiated *B*-state population. At lower temperature, rebinding is slow because only proteins with small enthalpy barriers are involved. As the temperature increases, more proteins with higher barriers can participate in the rebinding, until the rate drops again due to depletion of the *B* state. Figure 1(a) depicts the TDS surface $\Delta\mathcal{A}(\nu, T)$ for horse MbCO ($\beta =$ 0.5 K/ min, $T_i = 15$ K). The sample was prepared in a 75% glycerol solution according to standard techniques [8].

The rebinding of a SS with enthalpy barrier *H* can be modeled as a first-order rate processes with an Arrhenius-like temperature dependent rate function,

$$k(H, T) = A\left(\frac{T}{T_0}\right)e^{-(H/RT)}.$$

A is an experimentally determined frequency factor, and T_0 is an arbitrary reference temperature (taken to be 100 K). Solving the first-order equation with the TDS temperature constraint results in the intrinsic TDS line shape,

$$\mathcal{I}(H, T) = \frac{k(H, T)}{\beta} e^{-\Theta(H,T)},$$



(a) TDS data (dotted lines) with reconstructed forward image from MEM result (solid lines)

(b) MEM inverse image

FIG. 1. MEM inversion of horse MbCO data. $p = 0.1$, $\gamma = 2.0\ \mathrm{cm}^{-1}$, $\sigma = 0.02\max_{\nu,T}\Delta\mathcal{A}(\nu, T)$, TEST threshold $= 10^{-4}$, $A = 10^{9.2}\ \mathrm{s}^{-1}$. Initial $\mathbf{G}$ and $\mathbf{G}'$ were uniform and very small.

where

$$\Theta(H, T) = \int_{T_i}^{T} dT' \frac{k(H, T')}{\beta}$$

is a function related to the exponential integral of order 3 that can be evaluated numerically. In addition to being assigned a specific enthalpy barrier H, each SS is also associated with a specific CO stretch spectrum peak wave number. This is modeled as a Lorentzian spectrum, $\mathcal{L}(\nu, \nu')$, with HWHM γ whose peak ν' may vary across SS. The unity normalized contribution to the TDS surface $\Delta\mathcal{A}(\nu, T)$ from a SS with enthalpy barrier H and CO stretch peak ν' is given by

$$\mathcal{K}(\nu, T, H, \nu') = \mathcal{L}(\nu, \nu')\mathcal{I}(H, T). \tag{1}$$

Rebinding of a TS can be modeled as a superposition of such contributions, with weights given according to the occupancy of the SS. If there are many closely spaced SS, then this occupancy can be described by a continuous distribution $\mathcal{G}(\nu', H)$ of enthalpy barriers and CO stretch peaks.

$$\begin{aligned}\Delta\mathcal{A}(\nu, T) &= \int_0^\infty d\nu' \int_0^\infty dH \mathcal{K}\mathcal{G} \\ &= \int_0^\infty d\nu' \int_0^\infty dH \mathcal{L}\mathcal{I}\mathcal{G}.\end{aligned} \tag{2}$$

Consider an inverse problem of the form of Eq. (2). The problem is to determine the image $\mathcal{G}$ given a noisy sampling of $\Delta\mathcal{A}$ and a known kernel function $\mathcal{K}$, which is partially separable according to Eq. (1). A discrete analogue of the problem takes the matrix form

$$\mathbf{\Delta A} = \mathbf{LGI}, \tag{3}$$

where the transformation matrices $\mathbf{L}$ and $\mathbf{I}$ are sampled from the integral operators as $L_{ik} = \mathcal{L}(\nu_i, \nu'_k)\Delta\nu'$, and $I_{lj} = \mathcal{I}(H_l, T_j)\Delta H$. The sampling is such that the linear system (3) is highly underdetermined.

Such problems are generally ill posed—for a given goodness of fit there are many different images that are consistent with the data. To form a well-posed problem with a unique solution we introduce a regularization using the maximum entropy method (MEM) [9]. The MEM seeks to maximize the entropy in the image with respect to a prior image $\mathbf{G}'$, so that, among all the potential images that fit the data to within the noise, the one with minimum spurious structure is selected.

To obtain the MEM solution we maximize

$$F(\mathbf{G}) = S(\mathbf{G}) - \lambda\chi^2(\mathbf{G}),$$

where

$$S(\mathbf{G}) = \sum_{kl}\left(G_{kl} - G'_{kl} - G_{kl}\ln\frac{G_{kl}}{G'_{kl}}\right)$$

is the entropy of a candidate inverse image $\mathbf{G}$ with respect to the prior image $\mathbf{G}'$ as defined by Skilling [9,10], and

$$\chi^2(\mathbf{G}) = \frac{\|\mathbf{\Delta A} - \mathbf{LGI}\|_2{}^2}{|\mathbf{\Delta A}|}$$

is the goodness of fit, computed as the mean squared residuals. If σ is the standard deviation of Gaussian error in each measurement, then we wish to maximize S subject to the constraint $\chi^2 \simeq \sigma^2$. This can be achieved by tuning the Lagrange multiplier λ and maximizing F, which requires $\nabla F = \mathbf{0}$. This gives

$$\mathbf{G} = \mathbf{G}' \cdot e^{\lambda\mathbf{L}^T(\mathbf{\Delta A} - \mathbf{LGI})\mathbf{I}^T},$$

where $\cdot$ denotes the Hadamard product, the exponential is taken elementwise, and we have rescaled the Lagrange multiplier $2\lambda/|\mathbf{\Delta A}| \rightarrow \lambda$.

A relation of this form can be used iteratively to generate a solution, although some care must be taken to ensure convergence. Iteration is prone to instability due to the exponential. Following Gull and Daniel [11], we introduce the parameter p to smooth successive iterates.

$$\mathbf{G}_{n+1} = (1 - p)\mathbf{G}_n + p\mathbf{G}' \cdot e^{\lambda\mathbf{L}^T(\mathbf{\Delta A} - \mathbf{LG}_n\mathbf{I})\mathbf{I}^T}. \tag{4}$$

For sufficiently small p the iteration converges.

Skilling and Bryan [12] note that for a true maximum entropy solution, the gradients of S and χ^2 are parallel. Thus, the quantity

$$\text{TEST} = \frac{1}{2}\left\| \frac{\nabla S}{\|\nabla S\|_2} - \frac{\nabla \chi^2}{\|\nabla \chi^2\|_2} \right\|_2^2$$

can be used as a test of convergence—the iteration proceeds according to Eq. (4) until TEST is below a specified tolerance. After convergence is achieved, λ is incremented and the prior distribution $\mathbf{G}'$ is updated with the resulting $\mathbf{G}$. The iteration procedure is started again, and continues in this fashion until λ is sufficiently high so that the reconstructed forward image fits the experimental data to within the noise ($\chi^2 \simeq \sigma^2$). We note the significant computational cost reduction associated with partial kernel separability as expressed in Eq. (1), which permits formulation of Eq. (4) in terms of standard matrix operations.

The MEM described above has been used to successfully invert simulated TDS data with Gaussian noise. A variety of artificial images $\mathbf{G}$ can be recovered accurately from their noisy forward images, with particularly good results when $\mathbf{G}$ is sampled from a smooth function $\mathcal{G}$. The particular initial image and initial prior image chosen do not significantly effect the final image, only the number of iterations required, so the MEM image is not biased by the prior image. This success with inversion of simulated data gives some confidence in the robustness of the numerical scheme.

Figure 1 shows results for MEM inversion of MbCO TDS data. Isothermal rebinding experiments show the preexponential A to be about $10^{9.2}$ s^{-1} for horse MbCO [7]. γ is not known precisely; however, it must be less than the HWHM of the A_1 band (which is a superposition of single molecule absorbance bands) at about 4.5 cm^{-1} [13]. A lesser upper bound can be determined; it will not be possible to attain adequate forward image reconstruction if the underlying spectra are too broad. This upper bound is found to be about 2 cm^{-1}. Furthermore, not knowing γ, we take the view that using the largest value that permits adequate reconstruction is most parsimonious, as it minimizes spectral heterogeneity manifested in the inverse image. This can be seen as being in the same spirit of the MEM itself, that of Occam's razor. We are as conservative as the data allow in introducing structure in the image, or equivalently, we attempt to keep the protein ensemble maximally homogeneous. In any case, certain interesting features of $\mathcal{G}$ turn out not to depend strongly on γ.

KHB arises from a nonrandom mapping between CO stretch frequencies and rebinding barriers across the ensemble, and suggests that spectral and functional heterogeneity may share a common structural origin—both ν' and H are coupled to some conformational coordinate which varies across the ensemble. With the results of the previous section, namely $\mathcal{G}$, we are equipped to describe this mapping quantitatively. We may compute the conditional expectation of ν' at any H as

$$\mu_{\nu'}(H) = \int_0^\infty d\nu' \nu' \frac{\mathcal{G}(\nu', H)}{g(H)}.$$

Figure 2 shows $\mu_{\nu'}(H)$ computed from $\mathcal{G}$ using various γ. Evidently this measure of spectrum-function association is not very sensitive to γ. What is interesting about the curve is its nonlinear character, which suggests a possible structural interpretation.

Coupling of CO stretch frequency to structural coordinates is mediated by the vibrational Stark effect [14,15]. Modulation of stretch frequency is proportional to the projection of the local electric field along the CO transition dipole (approximately along the CO bond) $\Delta\nu' \propto \mathbf{E} \cdot \Delta\mu$. Structural and simulation studies have shown that the distal residue His64 is most responsible for modulating the CO stretch due to its proximity to the active site and protonation of N_ϵ (Fig. 3). The taxonomic state A_0 observed in the IR at low pH is associated with His64 swung outside the distal cavity, away from the active site [16]. The A_1 and A_3 bands have been identified with a rotation about the C_β-C_δ bond, such that N_ϵ-H is oriented toward or away from the active site [17].

The presence of His64 also regulates ligand binding through steric hindrance (forcing the CO to bind at an angle to the heme normal) and electrostatic interactions [18]. We suggest that the spectrum-function association described in Fig. 2 arises from varying swing orientation of His64 toward the heme center. As the residue swings toward the binding site, the rebinding barrier increases. However, for this same motion of the residue, the modulation of the CO stretch frequency due to the charged N_ϵ-H reaches an extremum, and then decreases to zero as N_ϵ-H approaches the plane that bisects the CO bond (where the projection of the field along the CO transition dipole vanishes). Rebinding barrier H increases along the structural coordinate while the CO stretch peak ν' reaches an extremum within the same range, producing the observed association between H and ν'.

To test this structural interpretation, varying His64 swing angles were simulated by rotating the C_α-C_β bond

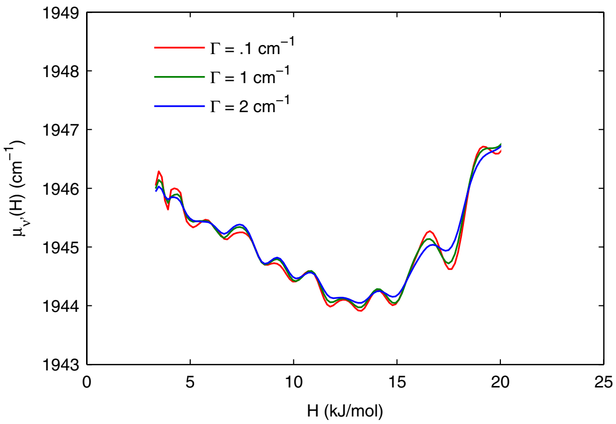


FIG. 2 (color online). Spectrum-function association.

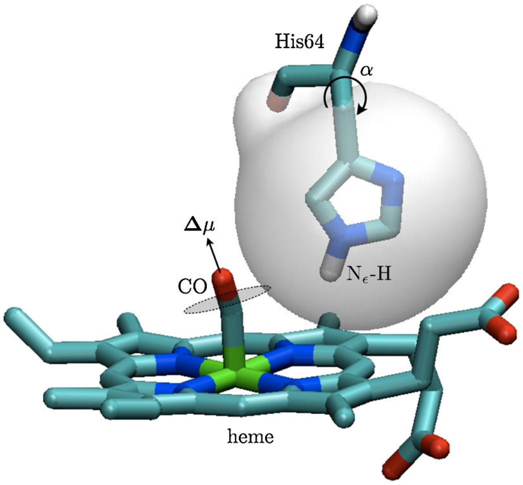


FIG. 3 (color online). Active site with proposed structural coordinate α and electrostatic potential isosurface at $-300(kT_0/e)$. Structure data are from PDB code 1DWR [6].

of a crystal structure (Fig. 3) using Molefacture Plugin in VMD [19], which allows one to freely adjust various conformational coordinates. For each angle α, the electrostatic potential due to the His64 residue was computed via the particle-mesh Ewald method on a dense grid containing the active site. The computation was done using PME Plugin for VMD [20] with an Ewald factor of 1 Å^{-1}. The electric field at the CO midpoint was then computed for each α and projected onto the normalized CO bond vector $\widehat{\Delta\mu}$. The dependence of the projected electric field on the swing angle is depicted in Fig. 4. This structure-spectrum association displays a nonlinear feature analogous to the spectrum-function association of Fig. 2, thereby corroborating the interpretation.

The concept of protein ensemble occupancy of a conformational energy landscape has proven crucial in understanding complex kinetic and spectroscopic behavior. Among TS the connections between spectrum, function, and structure have been investigated thoroughly and the state of knowledge at this level could be described as mature. The goal of the present study has been to extend this investigation to the second tier of the energy landscape. Because the SS constitute an effective continuum of states, the problem of teasing out relationships between spectrum, function, and structure is not as simple as for the TS, where each substate can be isolated and monitored individually. The problem has been formulated naturally as an inverse problem. Using an inversion method (MEM), which simulations indicate is robust, the ensemble occupancy of SS has been imaged in a spectrum-function phase space. Examining the phase space image $\mathcal{G}$ yields the spectrum-function association $\mu_{\nu'}(H)$, of which a structural interpretation has been described and computationally validated.

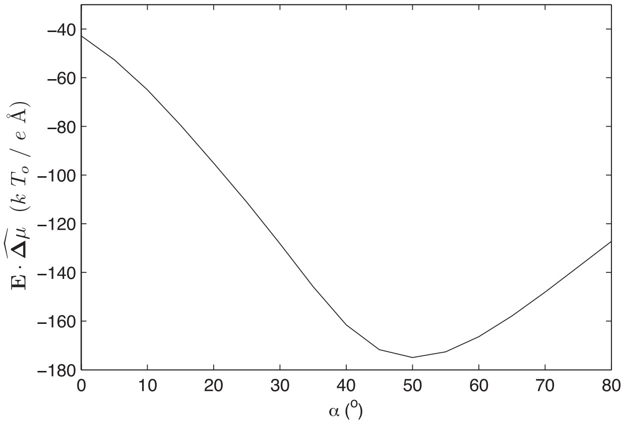


FIG. 4. Computed structure-spectrum association. The crystal structure in Fig. 3 corresponds to $\alpha = 0$.

*kelvin.chu@uvm.edu

[1] H. Frauenfelder, F. Parak, and R. Young, Annu. Rev. Biophys. Biophys. Chem. **17**, 451 (1988).
[2] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
[3] R. Austin, K. Beeson, and L. Eisenstein, Phys. Rev. Lett. **32**, 403 (1974).
[4] B. F. Cambell, M. R. Chance, and J. M. Friedman, Science **238**, 373 (1987).
[5] P. Ormos, A. Ansari, D. Braunstein, B. Cowen, H. Frauenfelder, M. Hong, I. Iben, T. Sauke, P. Steinbach, and R. Young, Biophys. J. **57**, 191 (1990).
[6] K. Chu, J. Vojtchovsky, B. McMahon, R. M. Sweet, J. Berendzen, and I. Schlichting, Nature (London) **403**, 921 (2000).
[7] K. Chu, R. M. Ernst, H. Frauenfelder, J. R. Mourant, G. U. Nienhaus, and R. Philipp, Phys. Rev. Lett. **74**, 2607 (1995).
[8] J. Berendzen and D. Braunstein, Proc. Natl. Acad. Sci. U.S.A. **87**, 1 (1990).
[9] *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer, Dordrecht, The Netherlands, 1989).
[10] P. Steinbach, Biophys. J. **70**, 1521 (1996).
[11] S. Gull and G. Daniell, Nature (London) **272**, 686 (1978).
[12] J. Skilling and R. Bryan, Mon. Not. R. Astron. Soc. **211**, 111 (1984).
[13] J. Moore, P. Hansen, and R. Hochstrasser, Proc. Natl. Acad. Sci. U.S.A. **85**, 5062 (1988).
[14] E. Park and S. Boxer, J. Phys. Chem. B **106**, 5800 (2002).
[15] H. Lehle, J. Kriegl, K. Nienhaus, P. Deng, S. Fengler, and G. Nienhaus, Biophys. J. **88**, 1978 (2005).
[16] F. Yang and G. Phillips, Jr., J. Mol. Biol. **256**, 762 (1996).
[17] K. Merchant, W. Noid, D. Thompson, R. Akiyama, R. Loring, and M. Fayer, J. Phys. Chem. B **107**, 4 (2003).
[18] T. Spiro and P. Kozlowski, Acc. Chem. Res. **34**, 137 (2001).
[19] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[20] A. Aksimentiev and K. Schulten, Biophys. J. **88**, 3745 (2005).