PROTEINS: Structure, Function, and Genetics 22:311–321 (1995)

RESEARCH ARTICLES

Automatic Identification of Discrete Substates in Proteins: Singular Value Decomposition Analysis of Time-Averaged Crystallographic Refinements

T. D. Romo,[1,3] J. B. Clarage,[1,3] D. C. Sorensen,[2,3] and G. N. Phillips, Jr.[1,3]
[1]Department of Biochemistry and Cell Biology and [2]Department of Computational and Applied Mathematics, and [3]the W. M. Keck Center for Computational Biology, Rice University, Houston, Texas 77005-1892

ABSTRACT **The singular value decomposition (SVD) provides a method for decomposing a molecular dynamics trajectory into fundamental modes of atomic motion. The right singular vectors are projections of the protein conformations onto these modes showing the protein motion in a generalized low-dimensional basis. Statistical analysis of the right singular vectors can be used to classify discrete configurational substates in the protein. The configuration space portraits formed from the right singular vectors can also be used to visualize complex high-dimensional motion and to examine the extent of configuration space sampling by the simulation.** © 1995 Wiley-Liss, Inc.

Key words: myoglobin, X-ray crystallography, molecular dynamics, conformation analysis, sampling, configuration space

INTRODUCTION

Knowledge of the motions of proteins is important to understanding their function. Proteins are not rigid static structures, but exist as a complex dynamic ensemble of closely related substates.[1–3] A 300 ps dynamics simulation of myoglobin by Elber and Karplus found that approximately 2000 minima were sampled by the simulation.[4] It is the dynamics of the system that permit transitions between these conformational substates and there is evidence that this dynamic behavior is also responsible for the kinetics of ligand entry and ligand binding in systems such as myoglobin.[1, 5, 6] These transitions are believed to involve anharmonic and collective motions within the protein.

X-Ray crystallography is able to provide high-resolution maps of the average structure of a protein, but this structure represents only a static conformation unless multiple sites are manually added via a special refinement scheme.[7] This static model can be augmented with a gaussian approximation of harmonic motion (B-values) for the atoms about their average position, or can it can be extended to a probability ellipsoid (anisotropic B-values). The dynamical information of accessible substates and their transitions that are dependent on anharmonic motion are lost. Molecular dynamics, which solves the Newtonian equations of motion for each atom, is not subject to such approximations and can be used to more closely model the range of motions available to the protein on short (nanosecond) time scales. Time-averaged refinement extends the molecular dynamics concept, restraining the simulation by requiring the time-averaged structure to fit the observed X-ray data.[8, 9] With this methodology, an ensemble of structures that together are consistent with the dynamics force fields and the observed data is generated. This ensemble represents a range of conformations that are accessible to proteins in the crystalline state.

One difficulty with time-averaged refinements is the plethora of generated conformations. No longer is one structure given as the model, rather, thousands of structures collectively describe the model. Although it is possible to analyze this data manually by examining ϕ–ψ plots, or watching movies of each residue in the simulation until some pattern emerges, this is not always a feasible approach. Nor is it feasible to watch an animation of an entire protein and discern the more subtle global conformation states that may exist. If this ensemble of conformations is thought of as a matrix however, then analytical tools from linear algebra can be used.

The singular value decomposition or SVD is an important and popular matrix decomposition.[10–12] It enjoys such popularity because it solves a variety of problems. Among other things, it can determine numerical rank, and can be used to construct a best low rank approximation to a given matrix. The SVD

Received November 9, 1994; revision accepted March 7, 1995.

Address reprint requests to G. N. Phillips, The Department of Biochemistry and Cell Biology, MS140, Rice University, P.O. Box 1892, Houston, TX 77005-1892.

has been used in a wide range of applications ranging from satellite image compression to separating out components of kinetics data.[11] Most of these rely upon the basis set the SVD constructs for the data. In fact, the SVD is central to the technique of principal component analysis in statistics. The size of the matrices involved in analyzing a molecular dynamics trajectory make existing algorithms such as those in LAPACK too expensive to use. A truncated SVD can be computed using an implicit restarted k-step Arnoldi algorithm which involves a relatively inexpensive matrix–vector product operation and is readily parallelizable.

The basis set that the SVD constructs can be thought of as a set of vectors spanning an m-dimensional space, each orthogonal to each other, where m is the number of rows of the matrix being decomposed. For a protein, this basis set describes the modes of motion of the atoms in the protein defined in a least squares sense. The SVD however includes both spatial and temporal information in a convenient mathematical form. It is the temporal information that sets it apart from the more traditional eigendecomposition analysis. The SVD is equivalent to the first stage of a quasinormal mode analysis in that it computes essentially the same quasiharmonic modes from a covariance matrix that is defined from molecular dynamics instead of an analytic harmonic potential.[13] The right singular vectors computed by the SVD however retain all the explicit time dependence and anharmonicity from the original dynamics simulation. Using the SVD, it is possible to extract information about the distribution of motion in the protein, to approximate the protein motion using a lower dimensional basis with less data, and to classify the conformational states of the protein and determine which state the system is in at any point in the simulation. Projections of the entire trajectory onto the SVD basis can also be used for understanding the extent of configuration space sampling of dynamics simulations and the topography of the system's energy hypersurface.[14]

In this paper, we have computed a time-averaged refinement of a mutant myoglobin. Using the SVD, we have characterized the motions within the protein and begun to automate this analysis. We have also used the SVD to distinguish between different conformational states within the dynamics ensemble.

METHODS

The Structure of F46V Met-Myoglobin

The time-averaged refinement was computed for a mutant form of myoglobin, F46V.[15] This is a mutant myoglobin grown in *E. coli* where the normal amino acid at position 46, phenylalanine, has been replaced by a valine.[16] This crystal form has 1277 nonhydrogen atoms with 143 crystallographically identified water molecules with one coordinated to the heme iron. The protein crystallizes in space group $P6$ with cell parameters $a = b = 91.22$ Å, $c = 45.80$ Å and $\alpha = \beta = 90°$, $\gamma = 120°$. A standard crystallographic refinement of this structure was made by Lai et al. who found that the distal histidine (*His*-64) is disordered and can be seen in two different conformations: one with the histidine pointing down toward the heme (closed) and one pointing out toward the solvent (open).[15] The structure is shown in Figure 1 with both distal histidine conformations. This structure with His-64 closed was used as the starting coordinate set for the time-averaged refinements.

Time-Averaged Refinement

The refinement was computed using a version of Xplor 2.1 that had been modified for time-averaged refinement. The reader is referred to Gros et al. and Clarage et al. for a more complete description of the algorithms used.[8, 9] Briefly, in a standard or "isotropic" crystallographic refinement, a target function consisting of a combination of stereochemical potentials or restraints, and a crystallographic energy function is minimized. The crystallographic potential is typically defined as

$$\Phi_{\text{X-Ray}} = W_x(d^{\text{T}}Wd) \qquad (1)$$
$$d = |F_{\text{obs}}(\mathbf{h})| - |F_c(\mathbf{h})|$$

In a time-averaged refinement, the instantaneous structure factor F_c is replaced with the time-averaged structure factor,

$$d = |F_{\text{obs}}(\mathbf{h})| - |\langle F_c(\mathbf{h})\rangle_t| \qquad (2)$$

The time-averaged structure factor is computed recursively,

$$\langle F_c(\mathbf{h})\rangle_t = e^{-\Delta t/\tau}\langle F_c(\mathbf{h})\rangle_{t-\Delta t} + (1 - e^{-\Delta t/\tau})F_c^t(\mathbf{h}) \qquad (3)$$

where τ serves as an exponential dampening parameter keeping the relative contribution of each new structure constant.

The statistic of choice among crystallographers is the unweighted R value. This statistic unfortunately cannot discern a model that spuriously fits the data from one that does not. In fact, R can be forced arbitrarily low by simply adding more parameters to the refinement. In a time-averaged refinement, this is precisely what is happening. More and more structures are being averaged together, until the τ_{max} is reached, adding more parameters to the system. Some other statistic must be used to verify that the system is not spuriously fitting noise in the X-ray data.

A standard technique in statistics to try to verify a model is to test the model's predictive ability. Brunger has adapted this technique, called bootstrap or cross-validation, to crystallography defining a new statistic called R_{free}.[17, 18] In this method, the reflection data are partitioned into a *working* set typically consisting of 90% of the reflection data and a *test* set containing the rest of the reflection data. The crystallographic model is refined against the

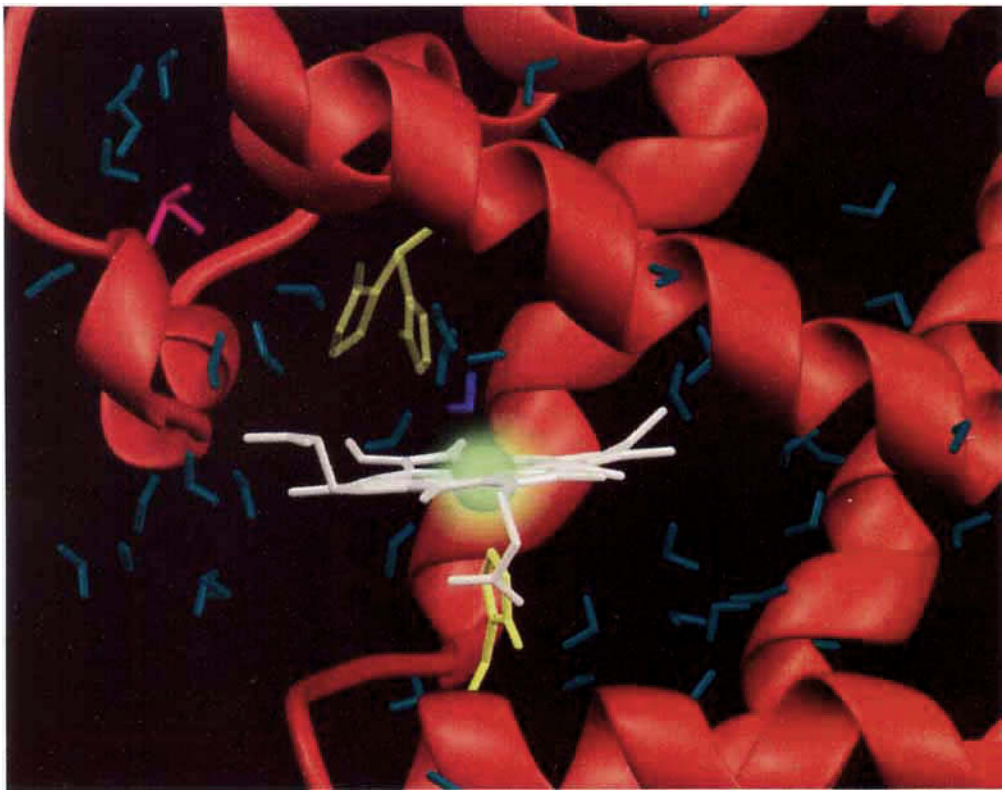

Fig. 1. The isotropically refined structure for F46V met myoglobin solved by Lai et al.[15] showing His-64 in two distinct conformations.

working set. The R_{free} value is derived by computing the R value using the F_c derived from the working set to predict the reflections in the test set. This is defined as

$$R_{free} = \frac{\sum_{\mathbf{h}\in test} \| F_{obs}(\mathbf{h}) - k|F_c(\mathbf{h})| \|}{\sum_{\mathbf{h}\in test} |F_{obs}(\mathbf{h})|} \quad (4)$$

Within the time-averaged refinement scheme then, R_{free} is the statistic of importance. When R_{free} reaches a minimum, this is considered to indicate convergence to at least a local minima for the refinement.

Prior to the time averaged refinement, the diffraction data were partitioned into a 10% test dataset and 90% working dataset. All reflections from 10.0 to 1.7 Å were used. The starting model was relaxed with 100 steps of conjugate gradient minimization in the absence of crystal packing energies followed by 100 steps with the crystal packing energies. Since the standard crystallographic refinement had been computed using all of the reflection data, the structure will be biased toward the test dataset. The starting structure was therefore debiased from the isotropic refinement by running 0.1 ps of unrestrained molecular dynamics.

The PARAM19X forcefields were used for the dynamics with the TIP3P potential for the crystallographic waters.[19] Only the crystal waters were used as a solvent. Explicit polar hydrogens were used with extended atom models used for all other atoms. No explicit hydrogen bonding energy was used but was included implicitly in the van der Waals and partial charge parameterization. The nonbonded pair cutoff was set at 7.5 Å. The B-values from the isotropic refinement were retained throughout the time-averaging since for the number of reflections in this dataset, the R_{free} diverged for $B = 0$.[9, 20, 21] A time-step of 0.001 ps was used with structures saved every 0.01 ps. The system was coupled to a 150K heat bath using Langevin dynamics with a frictional coefficient of 84.0 ps^{-1}.[22]

The averaging window variable, or τ, cannot be fully "on" at the start of the simulation. Here, τ was increased gradually to 4 ps when $t = 16$ ps, then ramped to 7.8 ps at $t = 20$ ps, and finally ramped to 16 ps when $t = 26$ ps after which τ was kept constant.

Singular Value Decomposition

The SVD of a matrix, A, is defined as

$$A = U\Sigma V^T \quad (5)$$

where U and V are orthonormal matrices. The columns of U, or u_i, are called the left singular vectors

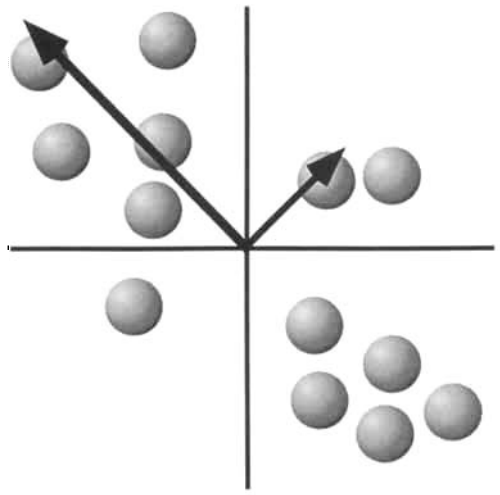

Fig. 2. This figure illustrates the construction of two left singular vectors for an atom constrained to move only in a plane. Each sphere represents a position of the atom at a different time-step. The large arrow represents the dominant first left singular vector and points along the major axis of motion. The smaller arrow represents the second left singular vector, orthogonal to the first. The magnitude of the arrows represent the relative magnitudes of the corresponding singular values.

of A, the columns of V, or v_i, are called the right singular vectors of A, and Σ is a nonnegative diagonal matrix whose diagonal elements, σ_i, are the singular values of A. For a more detailed treatment of the SVD, the reader is referred to Golub and Van Loan.[23]

If the atomic displacement in the $3N$ cartesian coordinates at time t about the geometric mean is defined as

$$a_t = [\Delta x_1, \Delta y_1, \Delta z_1, \ldots, \Delta x_n, \Delta y_n, \Delta z_n]^{\mathrm{T}}$$

then the atomic displacement matrix or conformation matrix is the column-wise concatenation of these displacement vectors,

$$A = [a_0|a_1|\ldots|a_L], \qquad A \in \mathcal{R}^{3n\times L} \tag{6}$$

where n is the number of atoms in the system and L is the number of conformations.

Every atom in the protein has a set of u_i vectors pointing in the direction of their "modes" of motion. These modes are defined as the least squares fit of a line to the space occupied by the atom over time. This is illustrated in Figure 2 for a single atom constrained to move in a plane. Each sphere represents the position of that atom at a different time point. The left singular vectors span the space formed by these positions, pointing in the directions of the axis of motion for the atom. This gives a first left singular vector, or dominant left singular vector, along the direction of largest motion. A second orthogonal vector points in the direction of the secondary mode. Any position of the atom in this plane can then be described by a linear combination of these two vectors. This will have important consequences for approximating the matrix A.

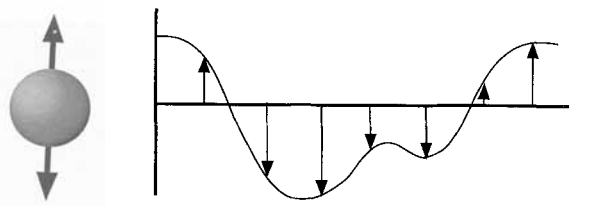

Fig. 3. This figure illustrates the construction of a right singular vector for an atom constrained to move in one dimension along a left singular vector. The curve shows the motion of the atom along this vector over time. The arrows represent the elements of the right singular vector which scale the left singular vector giving the atom's position at each time-step.

The right singular vectors, v_i, are projections of A along the u_is. This means that the v_is characterize the column space of A. For a protein conformation matrix, the v_i tell where each atom is along its u_i at each time point. Another interpretation is that the entire protein trajectory has been projected onto the corresponding left singular vector and this projection then becomes the elements of v_i. This is shown in Figure 3 for a single atom. The curve shows the continuous motion of the atom over time projected onto the left singular vector. The arrows on the curve indicate the scaling of the left singular vector, which are the elements of its right singular vector. If a small number of singular triplets $u_i\sigma_i v_i^{\mathrm{T}}$ dominate the atomic displacements, then the ensemble can be viewed in a lower dimensional configuration space defined by the dominant left singular vectors, as opposed to the entire $3N$ coordinate system. The corresponding right singular vectors then give the projection of the trajectory onto this lower dimensional basis.

It is from the analysis of this reduced configuration space that the notion of conformational substates and their extraction rests on. If a residue, for example, were to exist in a finite number of substates with transitions between them, then the right singular vectors computed from the SVD of that residue should show a similar modality in their probability distribution. This may not always be clear since projecting a higher dimensional topology onto a lower dimensional basis may obscure the substate geometry. If there is sufficient dominance of the singular vector, i.e., a sufficiently large singular value, then these projections will remain interpretable. So by examining the probability distribution for a right singular vector, i.e., its histogram, multimodality is indicative of the existence of conformational substates with respect to the U basis set. Similarly, deviations from a gaussian distribution indicate that the atom's motion derived from the dynamics does not fit the assumptions of the isotropic temperature factors. This substate analysis can be generalized to larger units of the protein such as domains or the whole protein. It is particularly interesting when

there is concerted large-scale motion as this motion will dominate the SVD.

The singular values, σ_i define the scale of the motion in the SVD space. It is enticing to state that the σ_is are the displacement in Angstroms along u_i, however, it is easy to see that as more columns are added to A and hence more elements to u_i and v_i, that σ_i must necessarily get larger because of the unitary constraints on u_i and v_i. The singular values, however, give the relative contribution of the corresponding mode. It is this separation of modes by the σ_i that determines the "essential" and "irrelevant" subspaces for the trajectory.[24]

Another formulation of the SVD is using a sum, $\Sigma u_i \sigma_i v_i^T$. It is possible to approximate the original matrix with a subset of singular triplets. An approximation to A can be defined using the first k terms of the SVD as

$$\hat{A} = \sum_{i<k} u_i \sigma_i v_i^T$$

with error,

$$\|A - \hat{A}\|_F = \sum_{i>k} \sigma_j^2$$

where $\|\cdot\|_F$ is the Frobenius norm (see Golub and Van Loan[23]). The error committed in $\hat{A}$ is then dependent on k and the contribution of each of the first k terms of the singular triplet.

Given a conformation matrix, A, the covariance matrix is then given by

$$S = \frac{1}{L} AA^T \tag{7}$$

Since the columns of U are also the eigenvectors for AA^T and the singular values, σ_i, are the square roots of the eigenvalues of AA^T, given an eigendecomposition for (7),

$$AA^T = U \Lambda U^T$$

the right singular vectors of A can be calculated,

$$\Sigma = \Lambda^{1/2}$$

$$V^T = \Sigma^{-1} U^T A \tag{8}$$

In finite precision arithmetic, this formula can be sensitive to roundoff error, but is not in the case of relatively large well-separated singular values. The SVD terms which are not well separated correspond to the high-frequency motions of the protein and are part of the "irrelevant" subspace and hence not interesting here. A truncated SVD was therefore used to compute the singular triplets corresponding to the "essential" subspace.[24]

Truncated SVDs were computed for the algebraically largest singular values. This was done in three steps: (1) a conformation matrix was built for the protein or subsets of the protein and a covariance matrix computed, (2) the algebraically largest eigenpairs were found using the ARPACK software, and (3) the conformation matrix was then projected onto these eigenvectors using (8) to find the right singular vectors.[25] A truncated SVD was calculated for all nonhydrogen and nonsolvent atoms for global motion. The first 5 terms of the SVD were also computed for each residue individually for local substate analysis. The coefficient of Kurtosis, m_4/m_2^2 where m_i is the ith moment of the data, was calculated for the first right singular vector. Those residues with the lowest coefficient of Kurtosis values were then examined in more detail.[26]

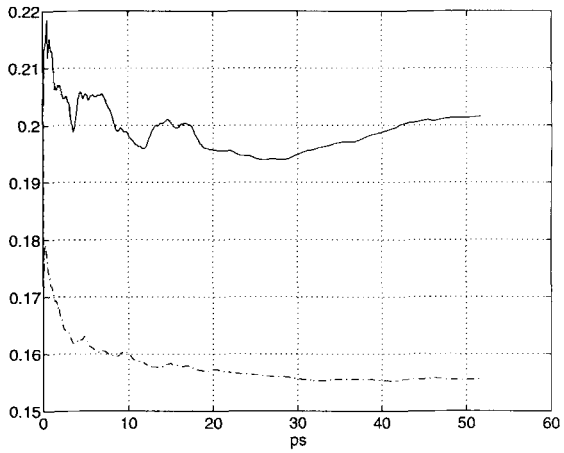

Fig. 4. The crystallographic R (dash-dot line), showing the progress of the model and the R_{free} (solid line) statistic, showing the validity of the model are graphed for the F46V-met time-averaged refinement. Though R remains low, R_{free} begins to increase at t = 30 nominal ps implying that noise is being fit rather than new information is being added to the structure.

RESULTS

Time-Averaged Refinement

The time-averaged refinement was run for a total of 51.8 nominal ps. Graphs of R and R_{free} versus time are shown in Figure 4. During the refinement, a minimum R of 15.52% was found at t = 41.0 nominal ps. The minimum R_{free} was 19.39% at t = 25.9 nominal ps which corresponds to an R of 15.62%. Only the first 30 nominal ps of the simulation was considered since after 30 nominal ps, R_{free} began to strictly increase. It is somewhat unsatisfying that the time-averaged refinement did not compute a significantly lower R value than the standard refinement even though it was not significantly worse. The previous isotropic refinement, however, did not use R_{free} and, in this case, it is the added dynamic information in the resulting time-averaged ensemble that is important.

Left Singular Vector and Singular Value Analysis

A partial singular value decomposition was calculated for all nonhydrogen nonsolvent atoms. The all-

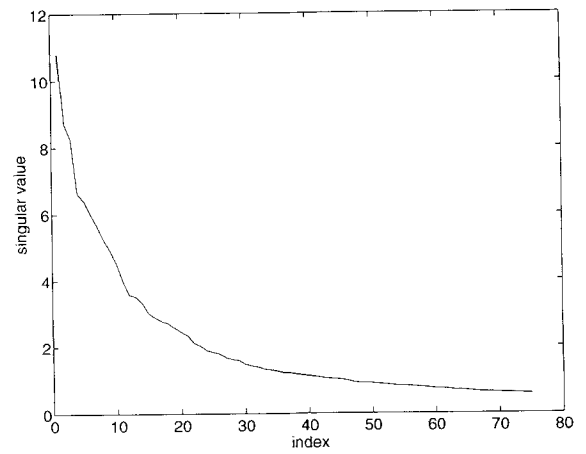

Fig. 5. The singular values from the SVD for all nonhydrogen atoms showing the relative magnitude of motion for each term of the SVD.

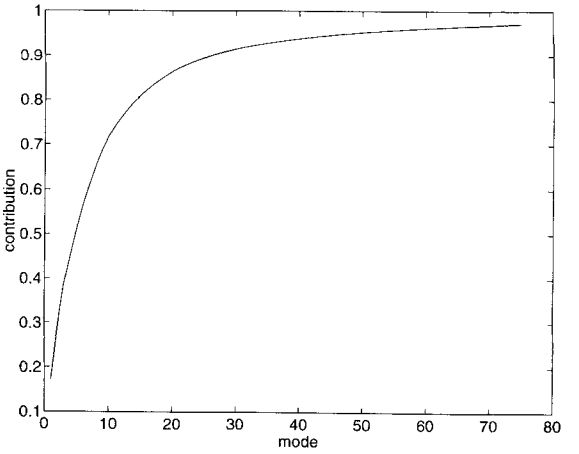

Fig. 6. Cumulative contribution of the ith term to the total motion in the simulation for the all nonhydrogen atom SVD.

atom SVD gave a 3831 × 3831 covariance matrix with a trace of 674.67 Å^2. This corresponds to a mean squared displacement, tr(S)/n, of 0.528 Å^2 for these atoms about their centroid position. The first 75 singular triplets were calculated from this covariance matrix using ARPACK and are shown in Figure 5. The relative contribution of each mode to the overall motion can be calculated by $c_i = \sigma_i^2/\mathrm{tr}(S)$. The cumulative contribution for all 75 modes is shown in Figure 6. The first 12 modes contribute 75% of the motion with 27 modes needed to describe 90% and 49 modes needed to described 95%. The results of the partial SVD for the first 4 residues selected by low coefficient of Kurtosis values are shown in Table I.

TABLE I. Statistics for the First 4 Residues Selected on the Coefficient of Kurtosis*

Residue	Trace	% Contribution	Kurtosis
Glu-148	3.37	51.1	1.09
His-64	18.50	91.5	1.22
Asp-61	2.83	64.9	1.29
Leu-89	4.62	55.2	1.31

*Trace is trace of AA^T calculated for the residue. Contribution is the relative contribution of the first left singular vector to the residue's motion. Kurtosis is the coefficient of Kurtosis calculated from the first right singular vector.

Right Singular Vector Analysis

Right singular vectors (RSVs) were computed for each corresponding left singular vector using Eq. (8). Histograms of the elements of a right singular vector show the probability distribution of conformational states on the SVD basis. Histograms for the first several RSVs for the all-atom SVD are shown in Figure 7. A gaussian approximation to each distribution is also shown. The first 10 modes bear little resemblance to a gaussian distribution. Only around the 30th mode does the distribution approach a gaussian.

Right singular vectors were also computed for each side chain and ranked in order of increasing Kurtosis. Histograms for the first RSV for Glu-148, Asp-60, and Leu-89 are shown in Figure 8. The first RSV, which is the projection of the side chain along its first LSV, is also shown. These plots clearly show the existence of two discrete configurational states. In each case, the state change involves flipping about the χ_2 bond.

The first RSV for the distal histidine is shown in Figure 9a. This plot shows the transitions of the histidine between the "closed" and "open" conformation states. Figure 9b shows the bimodal distribution of states along this RSV and the tight clustering of conformations. Plotting the first 3 RSVs as a single curve in 3-space gives a projection of the His-64 trajectory from its canonical 30-dimensional configuration space down to a readily visualized 3-D portrait, and is shown in Figure 10. This figure shows the clustering of this distal histidine conformational states in this new 3-D basis set and the narrow swing-pathway between the two clusters. The first left singular vector computed for each atom from the initial histidine position is shown mapped back onto their respective atoms as vectors in Figure 11. Direct inspection of His-64 under the influence of the time-averaged dynamics showed the histidine relaxing quickly to a conformation relatively close to the initial X-ray structure. Throughout the simulation, the histidine swung from closed to open, displacing waters, and then back again to interact with the coordinated water in the heme pocket. In addition to the major swing motion along the χ_1 and χ_2 bonds, there is a slight "wobble" of the imidazole during the rotation. Representative conformations of the histidine during the swing are shown in Figure 12.

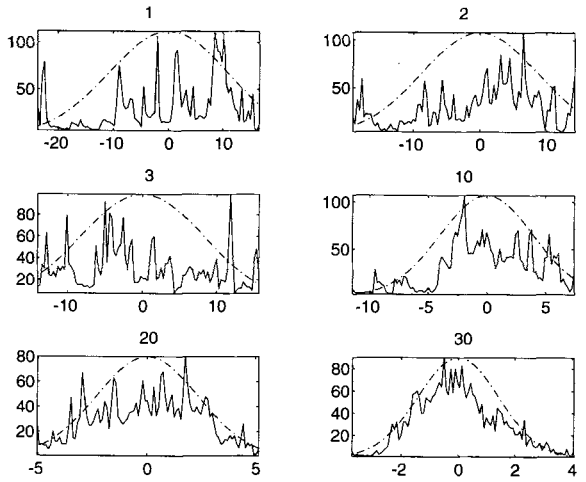

Fig. 7. Histograms of the right singular vectors for all nonhydrogen atom positions for the indicated mode showing the distribution of conformational states along the corresponding left singular vector. The dash–dot line represents a gaussian approximation to the distribution.

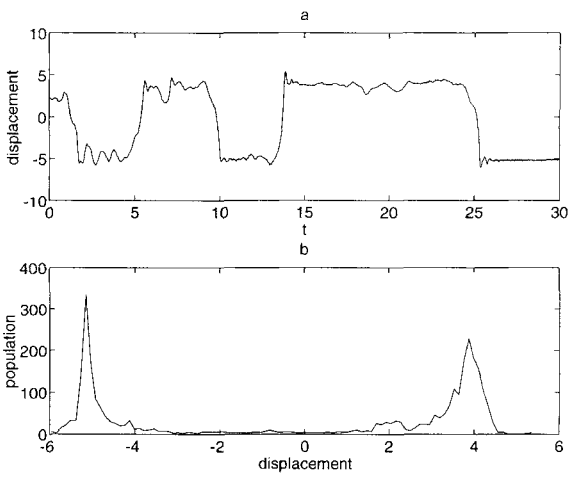

Fig. 9. **a:** The dominant right singular vector for His-64 plotted over the first 30 nominal ps of the simulation. **b:** The histogram for the first right singular vector for His-64 showing its bimodal distribution corresponding to the "up" and "down" configurations.

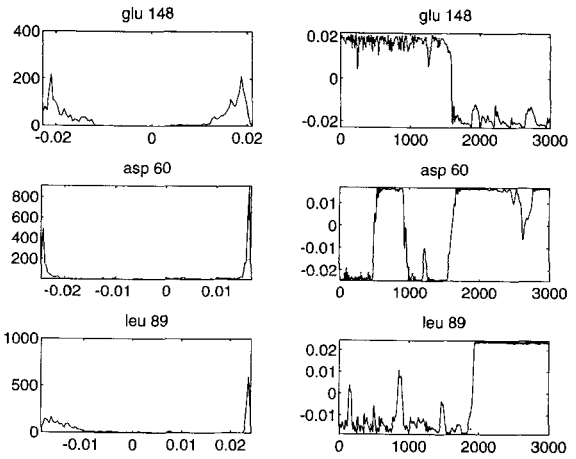

Fig. 8. The left column shows the histogram for the first right singular vector for each residue selected by the coefficient of Kurtosis. The right column shows the first right singular vector in nominal femtosecond steps for each residue indicating the number of transitions between states and where during the simulation they occurred.

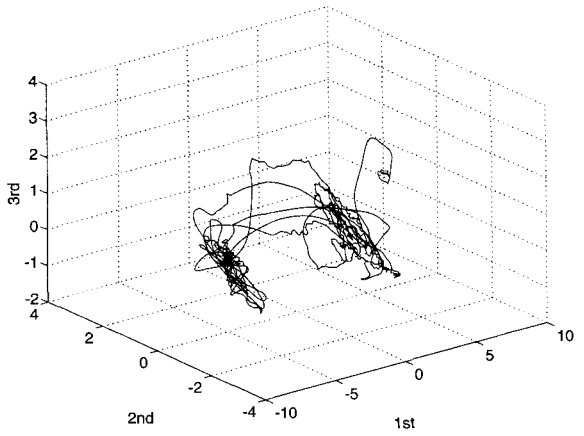

Fig. 10. The three-dimensional configuration space formed by the first, second, and third right singular vectors showing the tight clustering of the two configurations, up and down, and the narrow imidazole swing pathway.

DISCUSSION

The time-averaged refinement results have provided new insight into both the structure and dynamics of F46V met-myoglobin. In particular, the swinging of the distal histidine shows a biologically important motion that would not have been as readily found with an isotropic X-ray refinement. Additionally, the acceleration of the dynamics by the X-ray energy enables a larger region of the accessible conformation space of the protein to be searched in a smaller amount of "wall time" than with conventional molecular dynamics.

In X-ray crystallography, atoms are typically refined using a "single-site" model where each atom occupies its average position in the structure. A first approximation to the motion about this average is to use the isotropic B-factors which impose a spherical gaussian probability distribution function on the electron density of the atom. This effectively smears out the atom's density into a larger sphere. If there are sufficient experimental observations, this can be extended to an ellipsoidal function using anisotropic B-values. This model, however, still relies on a fundamental assumption of harmonicity or, equivalently, gaussian displacements in the motions of the atoms. Such assumptions hold for small molecule systems from whence X-ray crystallography sprung, but it is increasingly evident that this is not the case

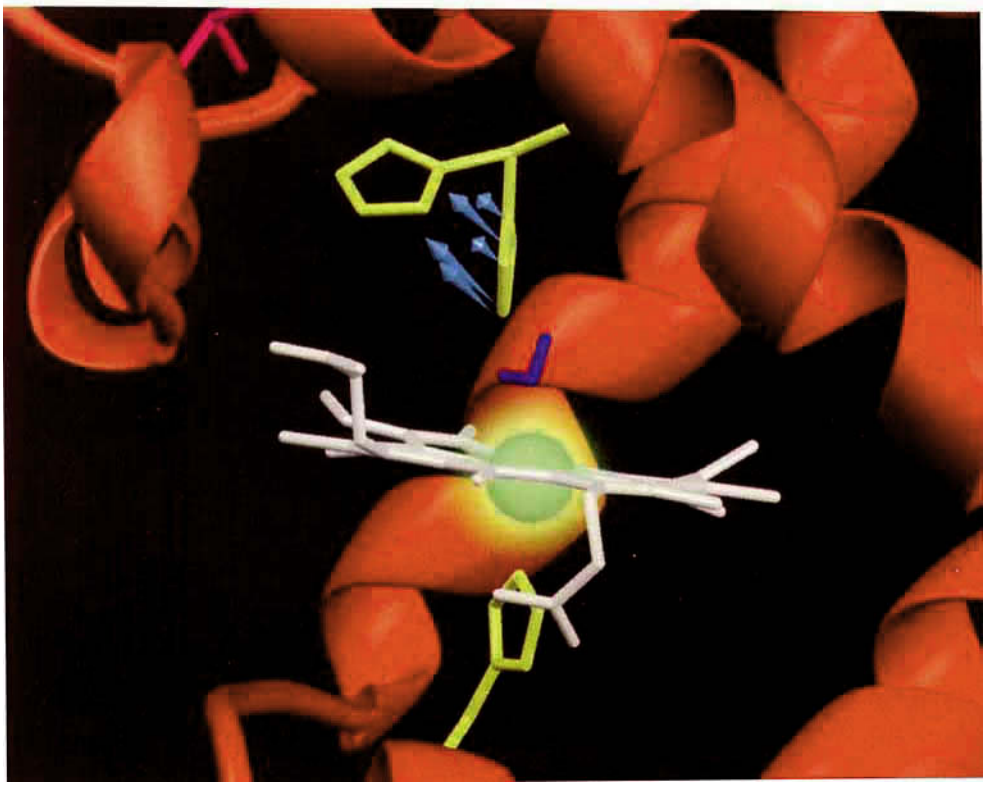

Fig. 11. The first dominant left singular vector for His-64 is shown based on the starting structure. These vectors point in the direction of the swing-path for the distal histidine and comprises 91.5% of the total motion in the simulation.

for macromolecular systems.[1]. This is particularly true for larger concerted motions within the protein. These deviations from anharmonicity are then modeled with larger and larger B-values which improperly characterize the motion and give a poor fit for the electron density.

Time-averaged refinement attempts to circumvent this problem by combining molecular dynamics with experimental structure factor restraints to model atomic motion in the protein. Since molecular dynamics solves the Newtonian equations of motion for each atom, it is not subject to the same harmonic constraints. The protein model is now free to explore a much larger region of its conformation space subject to these looser molecular dynamics restraints.

The use of molecular dynamics in X-ray crystallography to explore a protein's configuration space is not new. Crystallographic simulated annealing uses molecular dynamics with an added energy potential, such as in Eq. (1), to explore large regions of the protein's accessible configuration space in the context of an optimization problem.[27] In time-averaged refinement, this dynamic solution is cast in terms of finding an ensemble of structures which collectively fit the observed data. In other words, the computed electron density for the average conformation is restrained to match the observed X-ray data. In terms of energy potentials, the instantaneous structure factor is replaced with a time-averaged structure factor.

The phenylalanine at position 46 in myoglobin is a highly conserved residue and part of the "second shell" of residues around the heme binding pocket. Site-directed mutagenesis studies of this residue indicate that it plays an important role in determining the functional properties of myoglobin.[15] This residue helps to maintain the structural integrity of the CD corner and the entrance to the binding pocket. It also sterically limits the motion of His-64 maintaining an orientation optimal for hydrogen bonding with ligands and limits solvent access to the pocket both through the orientation of the distal histidine and by blocking the pocket entrance.

Changing Phe-46 to a Val removes the steric constraints on His-64 imposed by Phe-46. This provides a space for His-64 to rotate about its χ_1 and χ_2 bonds. In the X-ray crystal structure for F46V metmyoglobin solved by Lai et al. the distal histidine becomes disordered and can be modeled in either an "up" conformation pointing toward where the Phe-46 was or in a "down" conformation pointing toward the heme iron. There is not sufficient electron density present

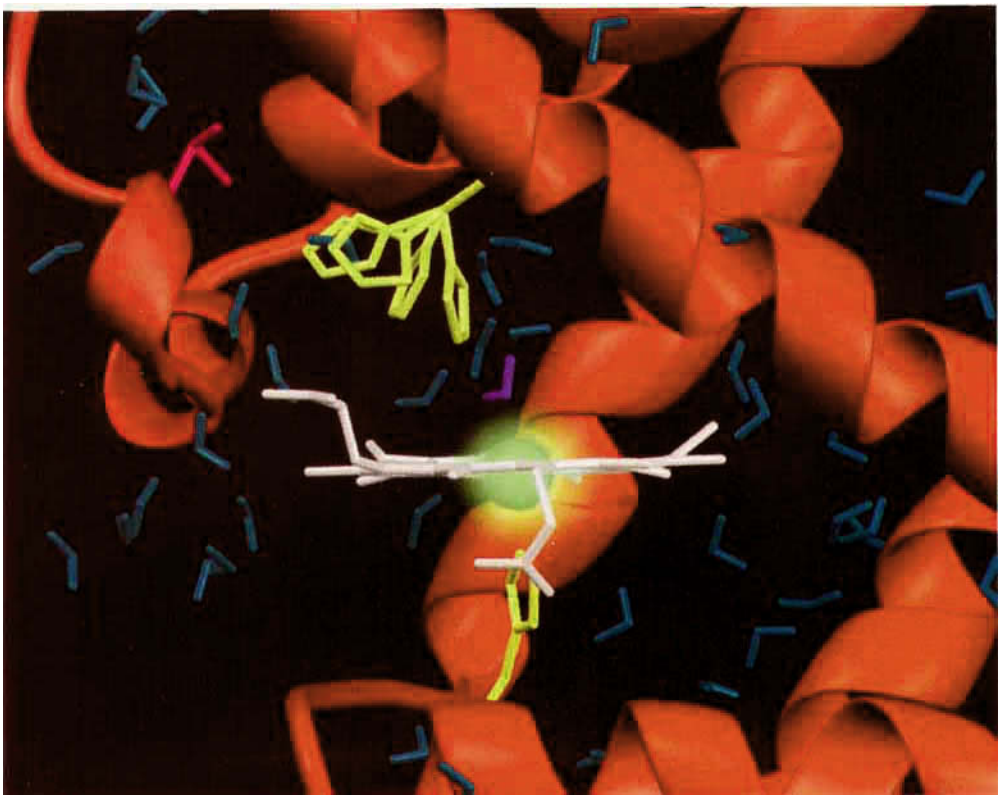

Fig. 12. Representative conformations of the His-64 during the first 30 nominal ps of the simulation showing the principal swinging motion and a secondary rotation of the imidazole ring.

to fully model a complete imidazole in either conformation however.[15] Presumably, His-64 does not exist as two distinct and rigid conformers in the crystal, but as a continuum of states ranging from the "up" and "down" conformations including the states comprising the swing pathway between those two extremum states.

In this refinement then, the density for each His-64 conformation provides an additional energy into the dynamics that forces the histidine to find both states. This energy input greatly accelerates the transitions of the histidine between its two configurational states. The normal time scale for such transitions is estimated to be on the order of nanoseconds, yet here the first swing is in 2 nomimal ps and there are 5 transitions in 30 nominal ps. The entire protein is driven by this added X-ray energy, in effect heated by the extra energy. This is why the system was coupled to a 150 K heat-bath. However, the extent to which the dynamics was accelerated as a whole is not clear.

The SVD is a powerful tool for analyzing not just proteins, but the configuration and motions of any n-particle system. One drawback to using the SVD for biological systems however is the lack of constraints. The SVD does not understand stereochemistry—there is no guarantee of physicality for the vectors and vector spaces the SVD constructs. The left singular vectors do point along modes of motion though these motions may not necessarily make good stereochemical sense. Similarly, conformations reconstructed using a partial SVD may not necessarily be stereochemically accurate. However, including sufficient numbers of terms from the SVD can approximate each atom's position to within a small enough error bounds that the effect of stereochemical constraints can be achieved. Such accuracy comes at the expense of compression performance. Even though these vectors are not strictly physical, these results clearly identify and quantify the distal histidine's motion in myoglobin.

It is intriguing to note that none of the low-frequency mode histograms bear much resemblance to a gaussian distribution. For crystallographic refinement, had the dynamics been consistent with a harmonic gaussian approximation, i.e., the isotropic B-values, then nearly all of these histograms would have been expected to resemble a gaussian. Amadei et al. have demonstrated this non-gaussian behavior for an MD simulation of solvated lysozyme over one nanosecond as well and used this to describe an "essential" and an "irrelevant" subspace in the modes of motion for their simulation.[24] It is possible however that this is an effect of inadequate sampling of

the accessible configuration space of the protein; were the simulation run longer, the lower frequency modes would become more and more gaussian. The SVD defined configuration portraits or projections of the system from the $3N$ configuration space to a few visualizable dimensions defined by the SVD may also assist in understanding the extent of sampling of this configuration space. This configuration space is defined in terms of the conformations found during the trajectory rather than a potential energy function. As the system fills in its configuration space and conformations are revisited, this should be apparent in the right singular vectors, such as with the distal histidine. This has been suggested by Clarage et al. on both experimental and theoretical grounds.[14] All of the modes tending to become gaussian cannot be universally true, the distal histidine clearly shows this. If there is discrete modality in the conformation space, then there will never be a gaussian distribution regardless of how long the simulation is run. It seems reasonable then to suspect that the partitioning of this configuration space is contextual, that is it depends on the system being simulated as well as the conditions of the simulation.

Our results for the SVD of the individual residues show one method for automating the extraction of salient information from a dynamics trajectory. In this case, the coefficient of Kurtosis for each residue's first right singular vector is a simple statistic that can be used to select "interesting" residues though it is not the only such statistic. It is possible that there are other conformational substates within the SVD of the residues beyond the first right singular vector. Since the contribution of this vector was large relative to the other modes, the substates defined by the higher frequency modes are not significant. The swinging distal histidine and its existence in two states as defined by X-ray crystallography provides a convenient "control" for the SVD analysis. The distal histidine was the second out of the top four residues found with this technique.

Recalling that the right singular vectors are the projection of the set of conformations, A, onto the corresponding left singular vector, then the right singular vector shows where in this new SVD basis the residue is at each time point in the simulation. The distribution of the right singular vector indicates what substates, if any, the residue is in. These states can also be found by visual inspection, however, the right singular vector is an easily interpretable 1-D curve. The state of the residue can be classified by this curve and its classification tracked with respect to time. This provides a natural way to quantify transitions based on conformations in cartesian space.

CONCLUSION

Our results show that the singular value decomposition can be used to identify discrete substates in large complex molecular dynamics trajectories. Using the SVD, any unit of the protein—residue, helix, domain—can be abstracted to a single point wandering around in a lower dimensional configuration space. This is a convenient method for visualizing complex motions and collective motions in a protein during a simulation both at the local level of residues and at the global level of the entire protein dynamics trajectory. The SVD basis provides a "natural" basis set since it represents the least-squares fit to the modes of motion for the unit being analyzed. We also show that time-averaged refinement can simulate biologically relevant motion in proteins beyond searching for lower R values.

ACKNOWLEDGMENTS

We thank John Olson and Henry Lai for providing the myoglobin mutant and starting X-ray structures as well as their insightful discussions. We also thank Luis Soltero for his guidance in coding. This work was supported by NIH grants AR40252 and GM-13945, the W. M. Keck Foundation, and NLM Grant LM07093, NSF Grant MCB-9315840, and NSF cooperative agreement CCR-9120008, and the Robert A. Welch Foundation.

REFERENCES

1. Frauenfelder, H., Sligar, S. G., Wolynes, P. G. The energy landscapes and motions of proteins. Science 254:1598–1603, 1991.
2. Garcia, A. E. Large-amplitude nonlinear motions in proteins. Phys. Rev. Lett. 68:2696–2699, 1992.
3. Straub, J. E., Rashkin, A. B., Thirumalai, D. Dynamics in rugged energy landscapes with applications to the S-peptide and ribonuclease A. J. Am. Chem. Soc. 116:2049–2063, 1994.
4. Elber, R., Karplus, M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. Science 235:318–321, 1987.
5. Hong, M. K. Conformational substates and motions in myoglobin: External influences on structure and dynamics. Biophys. J. 58:429–436, 1990.
6. Steinbach, P. J. et al. Ligand binding to heme proteins: Connection between dynamics and function. Biochemistry 30:3988–4001, 1991.
7. Kuriyan, J., Osapay, K., Burley, S .K. Exploration of disorder in protein structures by X-ray restrained molecular dynamics. Proteins 10:340–358, 1991.
8. Gros, P., van Gunsteren, W. F., Hol, W. G. J. Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. Science 249:1149–1152, 1990.
9. Clarage, J. B., Phillips, G. N., Jr. Cross-validation tests of time-averaged molecular dynamics refinements for determination of protein structures by X-ray crystallography. Acta Cryst D50:24–36, 1994.
10. Stewart, G. W. "Introduction to Matrix Computations." New York: Academic Press, 1973.
11. Strang, G. "Linear Algebra and Its Applications." New York: Harcourt, Brace Jovanovich, 1988.
12. Stewart, G. W. On the early history of the singular value decomposition. SIAM Rev. 35:551–566, 1993.
13. Levy, R. M., Karplus, M., Kushinck, J., Perahia, D. Evaluation of the configurational entropy for proteins: Application to molecular dynamics simulations of an α-helix. Macromolecules 17:1370–1374, 1984.
14. Clarage, J. B., Romo, T. D., Pettit, B. M., Phillips, G. N., Jr. A sampling problem in molecular dynamics simulations of macromolecules. Proc. Natl. Acad. Sci. U.S.A. 92: 3288–3292, 1995.

15. Lai, H., Li, T., Lyons, D. S., and Phillips, G. N., Jr. Phe46(CD4) orients the distal histidine for hydrogen bonding to bound ligands in sperm whale myoglobin. Proteins 22:322–339, 1995.
16. Springer, B. A., Sligar, S. G. High-level expression of sperm whale myoglobin in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 84:8961–8965, 1987.
17. Brunger, A. T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. Nature (London) 355:472–475, 1992.
18. Brunger, A. T. Assessment of phase accuracy by cross validation—The free R-value—methods and applications. Acta Cryst. D49:24–36, 1993.
19. Brunger, A. T. Xplor 2.1. Yale University, 1989.
20. Shiffer, C. A., Gros, P., van Gunsteren, W. F. The flexibility of time-averaged crystallographic refinement: α-Cyclodextrin as a test system. Aota Cryst. D. 51:85–92, 1995.
21. Gros, P., van Gunsteren, W. F. Crystallographic refinement and structure-factor time averaging by molecular dynamics in the absence of a physical force field. Mol. Simmulation 10:377–395, 1993.
22. Berendsen, H. J. C. et al. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81:3684–3690, 1984.
23. Golub, G. H., Van Loan, C. F. "Matrix Computations." Baltimore: Johns Hopkins University Press, 1989.
24. Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. Essential dynamics of proteins. Proteins 17:412–425, 1993.
25. Lehoucq, R., Sorensen, D. C., Vu, P. A. ARPACK: Fortran subroutines for solving large scale eigenvalue problems, Release 2.1, 1994. (Release 2.1 of ARPACK available from sorensen@rice.edu.)
26. Kuriyan, J., Petsko, G. A., Levy, R. M., Karplus, M. Effect of anisotropy and anharmonicity on protein crystallographic refinement. J. Mol. Biol. 190:227–254, 1986.
27. Brunger, A. T. Crystallographic refinement by simulated annealing: Application to crambin. Acta Cryst. A45:50–61, 1989.