JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Numbers 3–4, 2003

Pp. 617–634

Understanding Protein Flexibility through Dimensionality Reduction

MIGUEL L. TEODORO,[1] GEORGE N. PHILLIPS, JR.,[2] and LYDIA E. KAVRAKI[3]

ABSTRACT

This work shows how to decrease the complexity of modeling flexibility in proteins by reducing the number of dimensions necessary to model important macromolecular motions such as the induced-fit process. Induced fit occurs during the binding of a protein to other proteins, nucleic acids, or small molecules (ligands) and is a critical part of protein function. It is now widely accepted that conformational changes of proteins can affect their ability to bind other molecules and that any progress in modeling protein motion and flexibility will contribute to the understanding of key biological functions. However, modeling protein flexibility has proven a very difficult task. Experimental laboratory methods, such as x-ray crystallography, produce rather limited information, while computational methods such as molecular dynamics are too slow for routine use with large systems. In this work, we show how to use the principal component analysis method, a dimensionality reduction technique, to transform the original high-dimensional representation of protein motion into a lower dimensional representation that captures the dominant modes of motions of proteins. For a medium-sized protein, this corresponds to reducing a problem with a few thousand degrees of freedom to one with less than fifty. Although there is inevitably some loss in accuracy, we show that we can obtain conformations that have been observed in laboratory experiments, starting from different initial conformations and working in a drastically reduced search space.

Key words: dimensionality reduction, principal component analysis, protein motion, protein flexibility.

1. INTRODUCTION

THE FUNCTIONS OF PROTEINS CAN BE AS VARIED as enzymatic catalysis, mechanical support, immune protection, and generation and transmission of nerve impulses, among many others. Today there is a large body of knowledge available on protein structure and function as a result of several decades of

[1]Department of Biochemistry and Cell Biology and Department of Computer Science, Rice University, 6100 Main Street, MS 140, Houston, TX 77005.
[2]Department of Biochemistry and Department of Computer Science, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706-1544.
[3]Department of Computer Science and Department of Bioengineering, Rice University, MS 132, P.O. Box 1892, Houston, TX 77251-1892.

intense research by scientists worldwide. This amount of information is expected to grow at an even faster pace in the coming years due to new efforts in large-scale proteomics and structural genomics projects. In order to make the best use of the exponential increase in the amount of data available, it is imperative that we develop automated methods for extracting relevant information from large amounts of protein structural data. The focus of this paper is on how to obtain a reduced representation of protein flexibility from raw protein structural data.

Protein flexibility is a crucial aspect of the relation between protein structure and function. Proteins change their three-dimensional shapes when binding or unbinding to other molecules. Calmodulin is a representative example. This protein mediates a large number of cellular functions including ion channels, protein synthesis, gene regulation, cell motility, and secretion (Van Eldik and Watterson, 1998). Calmodulin is constituted by two large domains connected by a tether. This protein functions by binding to other proteins, and during this process it undergoes a drastic conformational rearrangement. When the protein binds one of its targets, the tether bends over its length, and the two calcium-binding domains reorient with respect to each other as shown in Fig. 1. Many other proteins undergo conformational rearrangements during the course of their function (Gerstein and Krebs, 1998).

Modeling protein flexibility computationally will be a major benefit to most aspects of biomolecular modeling, and we can envision several applications for our work. Currently used methods in pharmaceutical drug development use information about the 3D structure of a protein in order to find candidate drugs. One of the steps commonly used in the drug design process is to computationally screen large databases of small chemical compounds in search for those that complement the shape of an active site of a target protein. This latter step is known as molecular docking (Martin and Willett, 1998). Candidate drugs bind to the target protein, disrupting its function and leading to a desired pharmaceutical activity. However, during binding, some proteins undergo conformational changes in a process known as induced fit, which allows for higher interaction energy between the two molecules. This experimental fact is ignored by most current docking programs due to the computational complexity of explicitly modeling all the degrees of freedom of a protein (Muegge and Rarey, 2001; Teodoro *et al.*, 2001). Modeling proteins as rigid structures limits the effectiveness of currently used molecular docking methods. Using the approximation described in this paper, it will be possible to include protein flexibility in the drug design process in a computationally efficient way. A second potential application of our work is to model conformational changes that occur during protein–protein and protein–DNA/RNA interactions. Most current methods for studying these interactions are also limited in accuracy and applicability because the proteins involved are modeled as rigid.

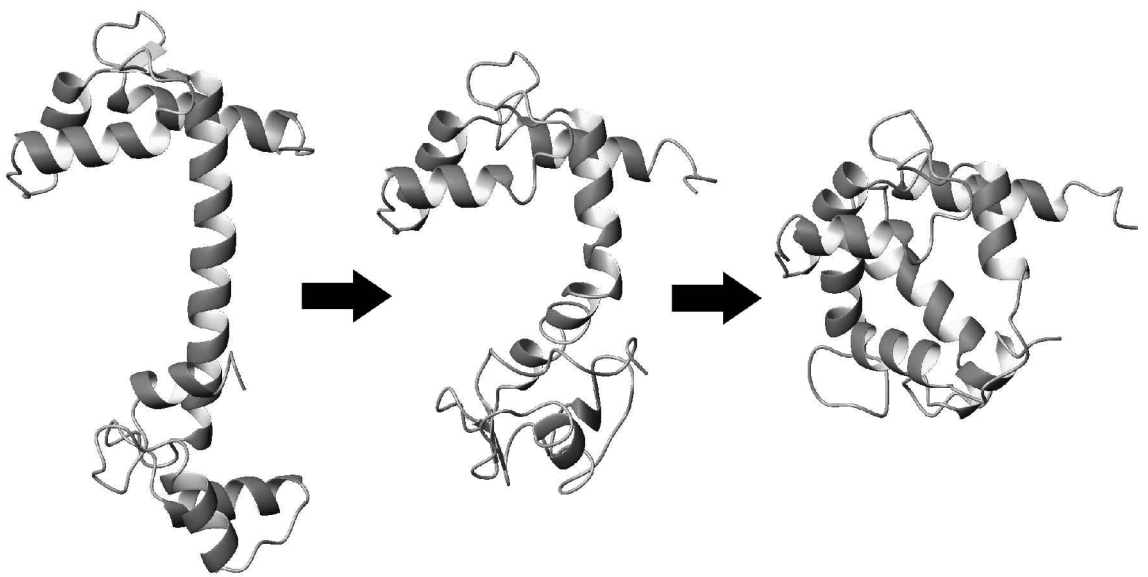

FIG. 1. Conformations of calmodulin. The unbound form shown on the **left** bends the α-helix connecting its two main domains when it binds to a target (not shown). The final bound conformation is on the **right**.

Current structural biology experimental methods are restricted in the amount of information they can provide regarding protein motions because they were designed mainly to determine the three-dimensional static representation of a molecule. The two most common methods in use today are protein x-ray crystallography (Rhodes, 1993) and nuclear magnetic resonance (NMR) (Wüthrich, 1986). The output of these techniques is a set of $\{x, y, z\}$ coordinate values for each atom in a protein. Neither of these methods is able to provide us with a full description, at atomic resolution, of the structural changes that proteins undergo in a timescale relevant to their function. Such information would be ideal to understand and model proteins. The alternative to experimental methods is to use computational methods based on classical (Brooks *et al.*, 1988) or quantum mechanics (Gogonea *et al.*, 2001) to approximate protein flexibility. However, these computations are prohibitively expensive and are not suitable for potential target applications such as the ones described in the previous paragraph. One of the reasons why the above computational methods are expensive is that they try to simulate all possible motions of the protein based on physical laws. For the case of molecular dynamics, the timestep for the numerical integration of such simulations needs to be small (in the order of femtoseconds), while relevant motions occur in a much longer timescale (microseconds to milliseconds). It is unrealistic to expect that one could routinely use molecular dynamics or quantum mechanics methods to simulate large conformational rearrangements of molecules. A medium-sized protein can have as many as several thousand atoms, and each atom can move along three degrees of freedom. Even when considering more restricted versions of protein flexibility that take into account only internal torsional degrees of freedom, or restrict the degrees of freedom to take only a set of discrete values, exploring the conformational space of these proteins is still a formidable combinatorial search problem (Finn and Kavraki, 1999).

The solution presented in this work addresses the high-dimensionality problem by transforming the basis of representation of molecular motion. Whereas in the standard representation all degrees of freedom (the $\{x, y, z\}$ values for each atom) of the molecule were of equal importance, in the new representation the new degrees of freedom will be linear combinations of the original variables in such way that some degrees of freedom are significantly more representative of protein flexibility than others. As a result, we can approximate the total molecular flexibility by truncating the new basis of representation and considering only the most significant degrees of freedom. The remaining degrees of freedom can be disregarded, resulting in only a small inaccuracy in the molecular representation. Transformed degrees of freedom will no longer be single atom movements along the Cartesian axes but collective motions affecting the entire configuration of the protein. The main tradeoff of this method is that there is some loss of information due to truncation (of the new basis), but this factor is outweighed by the ability to effectively model protein flexibility in a subspace of largely reduced dimensionality. We also show that there is inevitably some loss of accuracy, but the results are acceptable, consistent with experimental laboratory results, and help shed light on the mechanisms of biomolecular processes.

In this paper, we describe how starting from initial coordinate information from different data sources we apply the principal component analysis method of dimensionality reduction and obtain a new structural representation using collective degrees of freedom. In Section 2, we give some background on the most commonly used techniques of dimensionality reduction and some previously published applications of these in simulating and analyzing protein conformations. In Section 3, we explain how to apply the singular value decomposition method to perform the dimensionality reduction, while in Section 4 we describe the general methodology used to obtain the input data. In Section 5, we present the results we obtained from the dimensionality reduction of data for two protein systems of significant pharmaceutical relevance. Finally, in Section 6, we present our conclusions and discuss directions for future work.

2. BACKGROUND

Dimensionality-reduction techniques aim to determine the underlying true dimensionality of a discrete sampling X of an n-dimensional space. That is, if X is embedded in a subspace of dimensionality m, where $m < n$, then we can find a mapping $F : X \rightarrow Y$ such that $Y \subset B$ and B is an m-dimensional manifold. Dimensionality reduction methods can be divided into two types: linear and nonlinear. The two most commonly used linear methods to find such mappings are multidimensional scaling (MDS) and principal component analysis (PCA).

MDS encompasses a variety of multivariate data analysis techniques that were originally developed in mathematical psychology (Kruskal, 1964; Shepard, 1962) to search for a low-dimensional representation of high-dimensional data. The search is carried out such that the distances between the objects in the lower dimensional space match as well as possible, under some similarity measure between points in the original high-dimensional space.

PCA is a widely used technique for dimensionality reduction. This method, which was first proposed by Pearson (1901) and further developed by Hotelling (1933), involves a mathematical procedure that transforms the original high-dimensional set of (possibly) correlated variables into a reduced set of uncorrelated variables called principal components. These are linear combinations of the original values in which the first principal component accounts for most of the variance in the original data, and each subsequent component accounts for as much of the remaining variance as possible. Note that if the similarity measure of MDS corresponds to the Euclidean distances then the results of MDS are equivalent to PCA. The MDS and PCA dimensionality reduction methods are fast to compute, simple to implement, and since their optimizations do not involve local minima, they are guaranteed to discover the dimensionality of a discrete sample of data on a linear subspace of the original space.

One of the limitations of methods such as MDS and PCA is that their effectiveness is limited by the fact that they are globally linear methods. As a result, if the original data is inherently nonlinear, these methods will represent the true reduced manifold in a subspace of higher dimension than necessary in order to cover nonlinearity. To overcome this limitation, several methods for nonlinear dimensionality reduction have been proposed in recent years. Among these are principal curves (Hastie and Stuetzle, 1989; Tibshirani, 1992), multilayer auto-associative neural networks (Kramer, 1991), local PCA (Kambhatla and Leen, 1997), and generative topographic mapping (Bishop *et al.*, 1998). More recently, Tenenbaum *et al.* (2000) proposed the isomap method and Roweis and Saul (2000) proposed the locally linear embedding method. The main advantage of the last two methods is that the optimization procedure used to find the low-dimensional embedding of the data does not involve local minima. In general, the main disadvantages of nonlinear versus linear dimensionality reduction methods are increased computational cost, difficulty of implementation, and problematic convergence. The development of new methods for dimensionality reduction is an active research area.

The application of dimensionality reduction methods, namely PCA, to macromolecular structural data was first described by Garcia (1992) in order to identify high-amplitude modes of fluctuations in macromolecular dynamics simulations. It as also been used to identify and study protein conformational substates (Caves *et al.*, 1998; Kitao and Go, 1999; Romo *et al.*, 1995) as a possible method to extend the timescale of molecular dynamics simulations (Amadei *et al.*, 1993; Amadei *et al.*, 1996) and as a method to perform conformational sampling (de Groot *et al.*, 1996a, 1996b). The validity of the method has also been established by comparison with laboratory experimentally derived data (de Groot *et al.*, 1998; van Aalten *et al.*, 1997). An alternative approach to determine collective modes for proteins uses normal mode analysis (Go *et al.*, 1983; Levitt *et al.*, 1985; Levy and Karplus, 1979) and can also serve as a basis for modeling the flexibility of large molecules (Zacharias and Sklenar, 1999). Normal modes analysis is a direct way to analyze vibrational motions. To determine the vibrational motions of a molecular system, the eigenvalues and the eigenvectors of a mass-weighted matrix of the second derivatives of the potential function are computed. The eigenvectors correspond to collective motions of the molecule, and the eigenvalues are proportional to the squares of the vibrational frequencies. The PCA approach described in this article avoids some of the limitations of normal modes, such as lack of solvent modeling, assumption that the potential energy varies quadratically, and existence of multiple energy minima during large conformational transitions. In contrast to previously published work, we focus on the interpretation of the principal components as biologically relevant motions and on how combinations of a reduced number of these motions can approximate alternative conformations of the protein.

3. PCA OF CONFORMATIONAL DATA

Method

In this paper, we focus our analysis on the application of PCA to protein structural data. For our study, we chose PCA as the dimensionality reduction technique because it is very well established and

efficient algorithms with guaranteed convergence for its computation are readily available. PCA has the advantage over other available methods that the principal components have a direct physical interpretation. As explained later, PCA expresses a new basis for protein motion in terms of the left singular vectors of the matrix of conformational data. The left singular vectors with largest singular values correspond to the principal components. When the principal components are mapped back to the protein structure under investigation, they relate to actual protein movements, also known as modes of motion. It is now possible to define a lower dimensional subspace of protein motion spanned by the principal components and use these to project the initial high-dimensional data onto this subspace. The inverse operation can also be carried out, and it is possible to recover the high-dimensional space with minimal reconstruction error. By contrast, recovering the high-dimensional representation is not readily achievable when using MDS because the definition of the low-dimensional subspace is implicit in the projection and is not defined directly by the left singular vectors as is the case for PCA. The quality of the dimensionality reduction obtained using PCA can be seen as an upper bound on how much we can reduce the representation of conformational flexibility in proteins. The reason for this is that PCA is a linear dimensionality reduction technique and protein motion is in general nonlinear (Garcia, 1992). Hence, it should be possible to obtain an even lower dimensional representation using nonlinear methods. However, we wanted to test the overall approach before proceeding to more expensive methods. For nonlinear methods, the inverse mapping needs to be obtained using, for example, a neural network approach, but the feasibility and efficiency of these mappings has not been tested so far. There is active research in this area and our work will benefit from any progress.

In PCA, principal components are determined so that the first principal component $\mathrm{PC}_{(1)}$ is a linear combination of the initial variables A_j, with $j = 1, 2, \ldots, n$. That is,

$$\mathrm{PC}_{(1)} = w_{(1)1}A_1 + w_{(1)2}A_2 + \cdots + w_{(1)n}A_n,$$

where the weights $w_{(1)1}, w_{(1)2}, \ldots, w_{(1)n}$ have been chosen to maximize the ratio of variance of $\mathrm{PC}_{(1)}$ to the total variation, under the constraint

$$\sum_{j=1}^{n} (w_{(1)j})^2 = 1.$$

Other principal components $\mathrm{PC}_{(p)}$ are similarly linear combinations of the observed variables which are uncorrelated with $\mathrm{PC}_{(1)}, \ldots, \mathrm{PC}_{(p-1)}$ and account for most of the remaining total variation. Although it is possible to determine as many principal components as the number of original variables, this method is typically used to determine the smallest number of uncorrelated principal components that explain a large percentage of the total variation in the data. The exact number of principal components chosen is application dependent and constitutes a truncated basis of representation.

Conformational data

The data used as input for PCA is in the form of several atomic displacement vectors corresponding to different structural conformations which together constitute a vector set. We will call this set the conformational vector set. Each vector in the conformational vector set has dimension $3N$, where N is the number of atoms in the protein being studied and is of the form $[x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N]$, where $[x_i, y_i, z_i]$ corresponds to Cartesian coordinate information for the i^{th} atom. The first step in the generation of the atomic displacement vectors is to determine the average protein vector for each conformational vector set. This is achieved by first removing the translational and rotational degrees of freedom from the considered molecule by doing a rigid least squares fit (Kabsch, 1976) of all the structures to one of the structures in the vector set and then averaging the values for each of the $3N$ degrees of freedom. The resulting average structure vector is then subtracted from all other structures in the conformational vector set to compute the final atomic displacement vectors.

Singular valve decomposition (SVD)

In this work, we use the singular value decomposition (SVD) as an efficient computational method to calculate the principal components (Romo, 1998). The SVD of a matrix, A, is defined as

$$A = U\Sigma V^T,$$

where U and V are orthonormal matrices and Σ is a nonnegative diagonal matrix whose diagonal elements are the singular values of A. The columns of matrices U and V are called the left and right singular vectors, respectively. The square of each singular value corresponds to the variance of the data in A along its corresponding left singular vector, and the trace of Σ is the total variance in A. For our purposes, matrix A is constructed by the column-wise concatenation of the elements of a conformational vector set. If there are m conformations of size $3N$ in the vector set, this results in a matrix of size $3N \times m$. The left singular vectors of the SVD of A are equivalent to the principal components (Romo, 1998) and will span the space sampled by the original data. The right singular vectors are projections of the original data along the principal components. The right singular vectors also provide useful molecular information by helping to identify preferred protein conformations (Romo *et al.*, 1995; Teodoro *et al.*, 2000). For this paper, we construct A using the conformational vector sets of Section 4. The SVD of matrix A was computed using the ARPACK library (Lehoucq *et al.*, 1998). ARPACK is a collection of Fortran77 subroutines designed to solve large-scale eigenvalue problems. It is based upon an algorithmic variant of the Arnoldi process called the Implicitly Restarted Arnoldi Method (Lehoucq and Sorensen, 1996).

4. OBTAINING CONFORMATIONAL DATA

Ideally, the input data for dimensionality reduction would come from an accurate experimental technique that would permit the determination of the 3D structure at atomic resolution as it changes as a function of time. Using such a technique, we could collect a large number of samples at short time intervals (picosecond to nanosecond intervals). Unfortunately, such an experimental technique is not currently available. In order to perform the dimensionality reduction, we need to obtain as much data as possible about the protein system being studied from all available sources. The most common data sources are the experimental laboratory methods of x-ray crystallography and NMR, and force field based computational sampling methods such as molecular dynamics. Of these, the laboratory methods generate less data but do it with a greater accuracy.

X-ray crystallography

The most established and accurate method of determining the structure of a protein is protein x-ray crystallography (Rhodes, 1993). This technique is based on the collection of diffraction data generated by exposing a protein crystal to an x-ray beam. The experimental diffraction data is then computationally processed to yield a three-dimensional representation of the electron cloud of the molecule under study. Using these electron maps, it is then possible to accurately determine the spatial position of protein atoms in a process called fitting. The final outcome of the structural determination process is a single set of coordinates of all atoms in the molecule. The main limitation of this experimental technique is that it is necessary to obtain protein crystals in order to collect experimental data. Unfortunately, generating protein crystals is a very lengthy and laborious process which is not always successful. For example, membrane proteins, which are extremely important biologically, are notoriously difficult to crystallize, and very few structures of this class of proteins have been determined to date (Byrne and Iwata, 2002).

X-ray crystallography is sensitive to the experimental conditions under which it is performed, and changes of these conditions change the result. However, this a priori disadvantage can be sometimes beneficial. For example, by determining the structure of a protein in the presence of different inhibitors, we can determine different binding modes for potential drug candidates and observe how the protein is able to adapt to different molecules. For a number of interesting molecules, there are several structures available, and our analysis can be performed directly on the conformational vector set defined collectively by the experimental structures. This constitutes an advantage since no errors are introduced by the approximations

used in current computational methods. One example of such a protein is HIV-1 protease. For this system, there are several publicly available structures of the protein bound to widely different inhibitors. This is possible because the binding site of this protease is very flexible and is able to change its shape in order to complement the three-dimensional shape of the smaller molecules that bind to it (ligands). By analyzing the conformational data using a dimensionality reduction method such as PCA, we are able to reduce the overwhelming amount of information obtained from dozens of different structures to a small set of motions that summarize how the protein is able to adjust the shape of its binding site.

Nuclear magnetic resonance (NMR)

The second most common method of determining the structure of a protein is NMR (Wüthrich, 1986). This method uses a spectroscopy approach to collect the experimental data necessary for structure determination. The most important data obtained is a set of NOE (nuclear Overhauser effect) intensities produced by dipolar relaxation from neighboring spin systems. This intensity is inversely proportional to the distance between the spin systems, and as such it is possible to collect a series of distance constraints between atoms which enable the determination of the three-dimensional structure of a protein by solving a distance geometry problem (Guntert *et al.*, 1991).

This method is in general not as accurate as x-ray crystallography, and its use is limited to small and medium-sized proteins. However, it provides useful information about protein dynamics directly and avoids some of the problems of x-ray crystallography, such as protein crystallization. Another advantage of using NMR structures is that the final solution is not a single structure but a family of structures as required for input for dimensionality reduction. Although this family is usually composed of 10 to 50 structures, this number can be made as large as necessary by deriving more structures that satisfy the NMR experimental constraints to the same level as the original number. It is not clear, however, if new information is always useful for the dimensionality reduction technique since all the structures in the family are derived from the same set of experimental observations. Structures derived using either x-ray crystallography or NMR are stored in major databanks, such as the Protein Data Bank (Berman *et al.*, 2000).

Molecular dynamics (MD)

An alternative to using experimental methods to derive structural data is using computational methods such as MD (Brooks *et al.*, 1988). In fact, computational methods are used to augment existing experimental data since MD simulations typically start from a three-dimensional protein structure determined by x-ray crystallography or NMR. MD uses a force field (Cornell *et al.*, 1995; MacKerell *et al.*, 1998) to approximate the potential energy surface of a protein. The force field measures energy through a combination of bonded terms (bond distances, bond angles, torsional angles, etc.) and nonbonded terms (van der Waals and electrostatics). The relative contributions of these terms are different for the different types of atoms in the simulated molecule. They are determined by adjusting a series of parameters so that the molecule displays characteristics that have been observed experimentally or have been calculated from first principles. Once the force field has been specified, the time evolution of the system at an atomic scale is determined by solving Newton's equations of motion. One of the main disadvantages of MD is that it is very computationally expensive, which makes it impossible to run simulations for a timescale relevant to a majority of biological processes. Given current computer hardware, the timescales that are practically feasible to simulate even for a medium sized protein are usually less than 10–50 ns. The longest simulation published so far is 1 μs (Duan and Kollman, 1998). These simulations can take from days to months of computer time, even on parallel machines, depending on the timescale simulated and the complexity of the molecular model. Nonetheless, shorter simulations can provide us with invaluable data since they are the only method of observing proteins in "real time" and with atomic detail. For the purposes of determining a set of collective modes of motion representing the flexibility of the protein, it has been verified that it is only necessary to run short MD simulations ($<$ 1 ns) (de Groot *et al.*, 1996c) because the subspace spanned by the calculated left singular vectors derived from MD converge quickly.

MD is a good data source for our purposes because it can provide a large number of conformations of a molecule. However, MD is not as accurate as experimental methods due to approximations introduced in the computational process in order to make the MD simulation computationally practical. Among these approximations is the over simplified treatment of solvation effects and the lack of polarizability

representation in commonly used force fields. Addressing some of these limitations is currently the focus of many researchers, and there have been significant advances in the area of continuum electrostatic models and polarizable force fields. Improvements to the data generated by computational methods will reflect positively on the quality of the information that is obtained using dimensionality reduction.

When carrying out the dimensionality reduction described in this work, we must chose among the data sources which are available for the molecular system being studied. It is unlikely that data from all sources described in this section will be available simultaneously for any particular system. Furthermore, data obtained using exclusively experimental data sources is especially difficult to obtain. However, the availability of experimental data is very likely to increase in the future due to methodological improvements resulting from increased automation advances resulting in part from structural genomics projects. In the next section, we apply our method to two model systems using available sources of data.

5. APPLICATION TO SPECIFIC SYSTEMS

HIV-1 protease

The first model system used in this study is HIV-1 protease. HIV-1 protease is a homodimeric aspartyl protease with each subunit containing 99 residues. The active site of HIV-1 protease is formed by the homodimer interface and is capped by two identical β-hairpin loops from each monomer, which are referred usually as flaps (see Fig. 2). This protein plays a critical role in the maturation of the HIV-1 virus and has been the focus of intensive research in both academic and pharmaceutical communities. As a result, there is a large quantity of structural information on this system. This protein is known (Collins *et al.*, 1995) to undergo a large conformational rearrangement during the binding process consisting of the opening and closing of the flaps over its binding site. The conformations of the open and closed forms are overlapped in Fig. 2.

There are approximately 150 experimental structures available in public databases (*http://srdata.nist.gov/hivdb/* and *http://www.rcsb.org/pdb/*) for this system, and this number is continually growing due to its pharmaceutical importance. Of these structures, many are bound to different ligands and, depending on the size and shape of the ligand, display widely different conformations of the residues in the binding site.

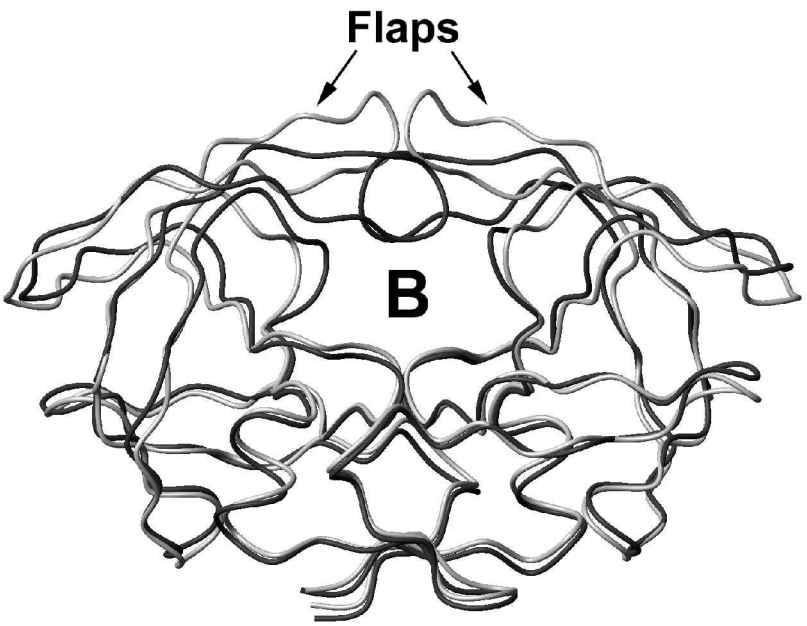

FIG. 2. Backbone representation of HIV-1 protease for the unbound (gray) and bound forms (black). The arrows indicate the flaps region where large conformational changes take place. The binding site is the location indicated by B.

It has been observed that the volume of the cavity can approximately double depending on the ligand. In Fig. 3, we show a tube representation of HIV-1 protease bound to two different inhibitors. The structures shown correspond to PDB access code 4HVP and 1AID. As can be seen in the figure, the protease is able to change its shape mostly in the flaps region to accommodate for different ligands which vary considerably in terms of shape and volume. HIV-1 protease provides an excellent demonstration of protein plasticity and underlines the importance of understanding and modeling protein flexibility for binding purposes.

One of the advantages of using the PCA methodology to analyze protein flexibility is that it can be used at different levels of detail depending on what kind of information we are interested in obtaining. For example, if we are interested in the overall motion of the backbone, then we can construct the matrix A defined in Section 3 using only the coordinate information from the α-carbons. This reduces the number of degrees of freedom to 3×198 and makes the computation of the SVD faster. Alternatively, we can include all the atoms of the protein for a total of $3 \times 3{,}120$ degrees of freedom and be able to observe the simplified flexibility of the protein as a whole. As an intermediate case, we can include only the atoms that constitute the binding site to study how it can change shape during the binding of different ligands. This intermediate solution would probably be better for a drug design study. Below, we describe the results obtained for all three cases above.

In the first experiment, we determined the modes of motion generated from MD data. The initial structure used for the simulation was determined by x-ray crystallography (Miller *et al.*, 1989) (PDB access code 4HVP). Molecular dynamics simulations for this system were carried out with the program NAMD2 (Kalé *et al.*, 1999) using the charmm22 force field (MacKerell *et al.*, 1998). Since we were mostly interested in conformational changes in the binding site of this protein, we did not include any inhibitor in the simulation in order to be able to observe a larger range of conformational motions in this region. The simulations were carried out in a box of TIP3 water using periodic boundary conditions, particle mesh Ewald full electrostatic integration, and pressure and temperature coupling using the Berendsen algorithm (Berendsen *et al.*, 1984). After an equilibration period of 200 picoseconds, the simulations were carried out for an extra 1.4 nanoseconds each at a temperature of 300K, and structures were saved to disk every 100 femtoseconds. The resulting 14,000 structures were used in the dimensionality reduction procedure.

In Fig. 4, we show the fraction of the total variance explained by the 20 most significant left singular vectors for the SVD analysis of the coordinate data for all 198 α-carbons in HIV-1 protease. The largest singular value accounts for 35% of the total variance. The first 3 and first 20 account for 53% and 80%, respectively. In practice, this means that we can approximate a system of 596 dimensions using only 20 dimensions and are still able to retain 80% of the variance in the original data. This results in a drastic reduction of complexity with only a small reduction in our ability to represent different conformations of HIV-1 protease.

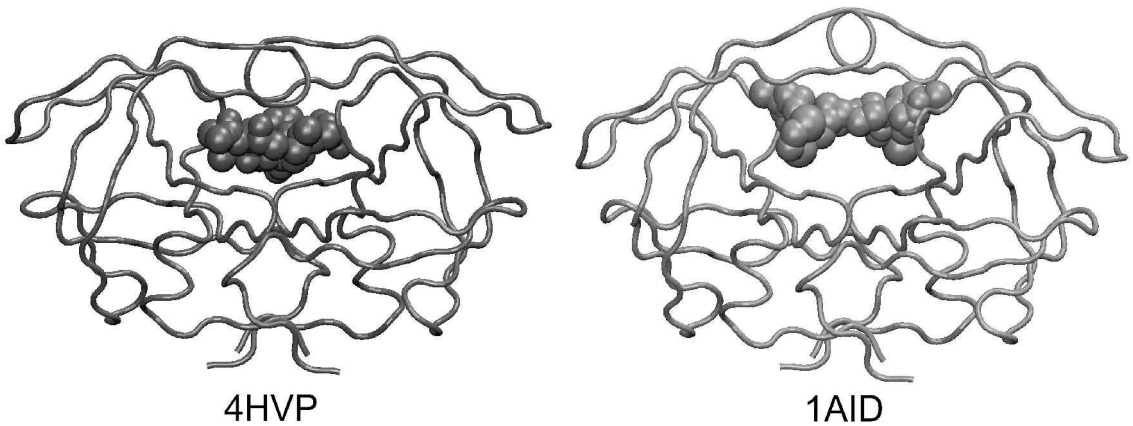

FIG. 3. Tube representation of HIV-1 protease (PDB access codes 4HVP and 1AID) bound to different inhibitors represented by spheres. The plasticity of the binding site of the protein allows the protease to change its shape in order to accommodate ligands with widely different shapes and volumes.

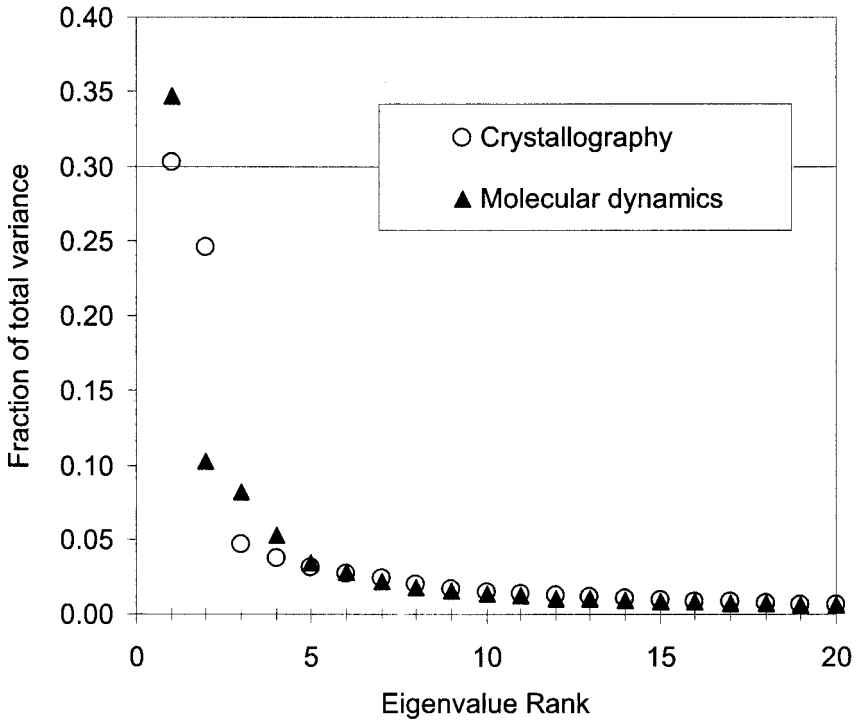

FIG. 4. Fraction of total variance represented by the most significant singular values for HIV-1 protease dimensionality reduction.

In Fig. 5, we show the motion of the backbone that was captured in the first principal component. This motion matches well with the opening and closing of the binding site, a fact that has been determined experimentally (Fig. 2). It also shows the strength of this method in isolating the most relevant biological motions from a large amount of high-dimensional input data where they were not clearly recognizable. The reason why it is difficult, even for an expert, to recognize these principal motions directly from the raw MD data is the enormous quantity of information generated by this technique. PCA also reveals the tendency of the protein to move in a certain direction even though this movement is not fully explored during the MD run. This constitutes one of the main advantages of this method since it enables the

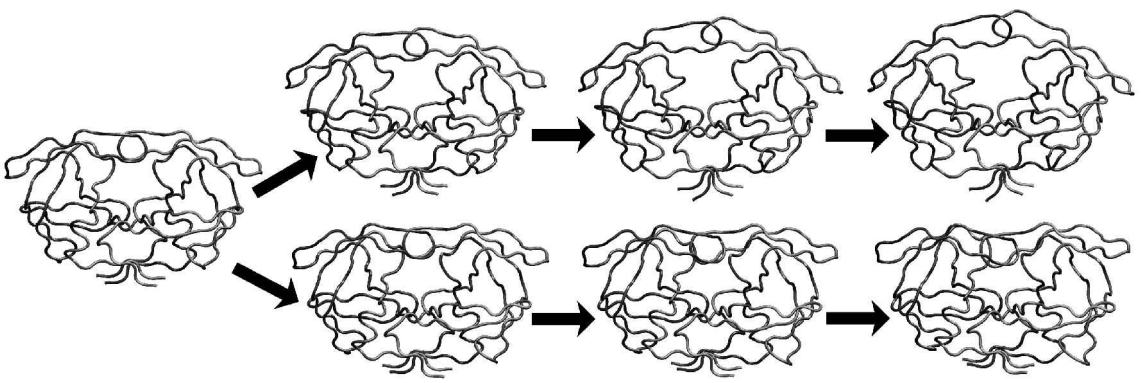

FIG. 5. HIV-1 protease backbone motion as defined by the first principal component. The structure shown at the **center left** corresponds to the bound reference structure (ligand not shown). As the structure moves along the first principal component the flaps either close more over the binding site (**bottom sequence**) or, if moving in the opposite direction, lead to an open conformation (**top sequence**) very similar to structures obtained by crystallography for the unbound conformation.

discovery of the important motions using shorter MD simulations with a consequent drastic reduction in computational cost.

It is important to emphasize at this point that no bias was introduced at any point in the calculation that would inevitably lead to the observed result. The input to the data reduction consisted uniquely of MD sampling data obtained from a short simulation starting from the bound conformation without the ligand. What is being captured is the opening caused by the removal of the ligand *without* driving the simulation directly to the final open conformation as is the case for the steered MD technique (SMD) (Israelewitz *et al.*, 2001). The reason we avoid this alternative is that we want to validate the utility of using MD simulations when only one experimental structure is known for the system of study. We do not want to introduce any bias to the system as SMD would do.

The results obtained from SVD for all atoms were similar to the α-carbon approximation but contain extra information about amino acid sidechain movements. The most dominant, the first three, and the first twenty left singular vectors account for 20%, 37%, and 68% of the total variance, respectively. These values are smaller than for the α-carbon carbon dimensionality reduction because the total number of degrees of freedom considered is much larger. It is again clear from these values that in the new basis only a few degrees of freedom account for most of the conformational variation.

One advantage of using the HIV-1 protease as a model system is the wealth of structural information publicly available for this protein. It is possible to carry out the same type of dimensionality reduction work using only laboratory-determined structures of HIV-1 protease bound to different ligands. Here, we present results of applying PCA to x-ray crystallographic data. A similar analysis is possible using families of structures derived from NMR data. We used 130 structures of HIV-1 protease deposited in the Protein Data Bank. The coordinates of the different ligands bound to the structures were not included in the calculations. The fraction of total variance represented by the most significant singular values for the PCA of the α-carbon coordinate information using exclusively laboratory derived data is also shown in Fig. 4. This result is similar to the result obtained from the MD data. The most significant left singular vector accounts for 30% of the total variance, and the first 20 account for 85%.

As a final validation of our method, we decided to investigate if, using the main modes of motion defined by the principal components and an experimental structure bound to a particular ligand, we could approximate the structure of HIV-1 protease bound to a different ligand. For this experiment, we were only concerned with variations in the shape of the binding site and computed the dimensionality reduction only for this part of the protein. We defined the binding site atoms as those that are part of amino acids that touch the ligand directly in any of the x-ray structures in the PDB. A total of 266 atoms were identified. As the initial reference, we chose the same structure we used for the MD simulation, and as a target structure, we used a complex with a large nonpeptide inhibitor (Rutenber *et al.*, 1993) (PBD access code 1AID). The binding site conformations as well as the inhibitors bound to these are considerably different as shown in Fig. 3. The root mean square deviation (RMSD) between the two proteins is 1.86 Å if we take into account only the atoms that constitute the binding site.

The next step was to calculate the coordinates of the target structure in the new basis. For this, we used the definition of the representation basis given by the principal components of the MD data, and we set the origin of the space to be our reference structure. The coordinates in each of the dimensions are given by the dot product of the atomic displacement vector and the left singular vector defining each dimension. The resulting coordinates will be a solution vector of the form $[w_1, w_2, w_3, w_4, \ldots, w_{3N}]$. We can now calculate what would be the RMSD between our target structure and our low-dimensional approximation. The approximation corresponds to $[w_1, 0, 0, 0, \ldots, 0]$ if we consider only the first collective mode, $[w_1, w_2, 0, 0, \ldots, 0]$ if we consider the first two, and $[w_1, w_2, w_3, w_4, \ldots, w_k, 0, \ldots, 0]$ if we consider the first k collective modes. The RMSD results for an increasing number of collective modes are shown in Fig. 6. When using the PCA basis, we are able to approximate the target structure to an RMSD of less than 1 Å using 40 principal components out of a total of 798. By contrast, if we used an approximation with a random orthonormal basis (Wolfram, 1999) defining the same space (shown by a broken line in Fig. 6), we would need more that 650 principal components to obtain the same accuracy. This shows the strength of our method in approximating other conformations of the same protein using a lower dimensional search space and validates the effectiveness of the PCA by comparing it with an approximation carried out using a random basis. It is also important to note that the values that we obtain for the approximation to the target can be further improved. Currently, we are using the projection of the target structure on the new

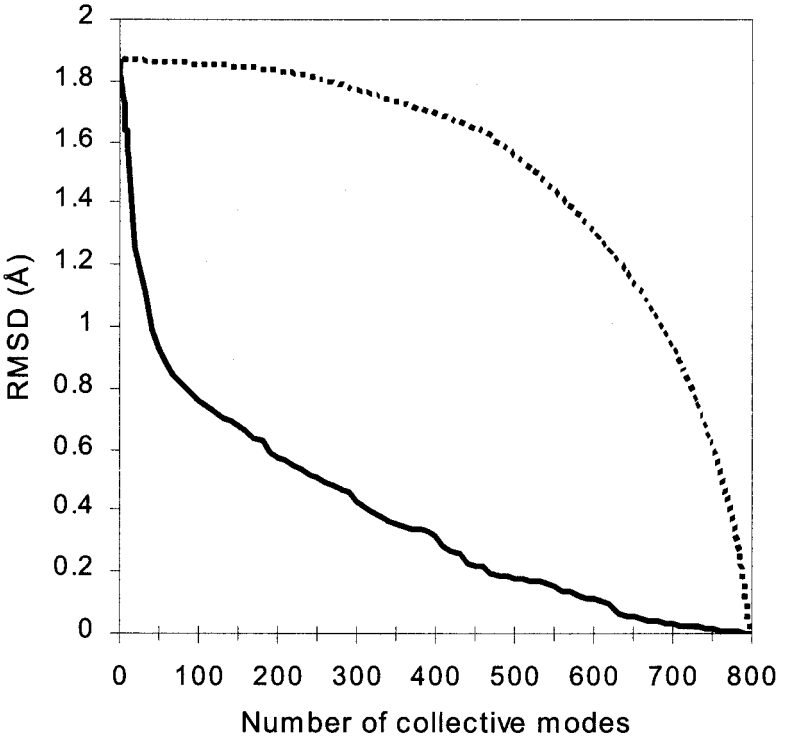

FIG. 6. RMSD between a reference (4HVP) and a target structure (1AID) for an approximation of the flexibility of HIV-1 protease using an increased number of collective modes. The solid uses the collective modes basis determined by PCA, and the broken line uses a random basis defining the same space.

basis to estimate a set of coordinates in the reduced space that approximate the target structure. However, the optimal approach is to search the low-dimensional space directly to look for the best match. In this way, we can search for alternative coordinate values along the most significant principal components that compensate for the approximation being introduced by the dimensional truncation of the representation basis. We are currently developing search techniques for the purpose of finding these solutions in the reduced space.

Aldose reductase

The second model system used in our study is aldose reductase. The biological function of this enzyme is yet not entirely known but it is believed to play a primary role in the development of severe degenerative complications of diabetes mellitus (Larson *et al.*, 1988). Finding new inhibitors for this protein could potentially lessen some of the complications of diabetes. Unfortunately, just like in the case of the previous example and like many other proteins, aldose reductase has the capacity to adjust the shape of its binding site depending on the ligand it is binding to. A small-molecule database screen for potential ligands would miss many potential candidates if it did not include the protein flexibility in the search process. Several experimental structures of aldose reductase have been solved using x-ray crystallography when bound to different ligands as well as in the unbound form. It was observed that with some inhibitors such as sorbinil (Urzhumtsev *et al.*, 1997), the structure is almost similar to the unbound form. For other inhibitors, such as the tolrestat (Urzhumtsev *et al.*, 1997) and zopolrestat (Wilson *et al.*, 1993), there is a formation of a specificity pocket resulting in significantly different binding site configurations. This conformational change is shown in Fig. 7 where we compare the shape of the binding site for the unbound form of the enzyme to the bound form with tolrestat. In the unbound form shown on the left of Fig. 7, there are a series of amino acids (represented by ball-and-stick models), which come together to close the specificity pocket. In the presence of tolrestat (represented on the right by a van der Waals sphere model), the amino

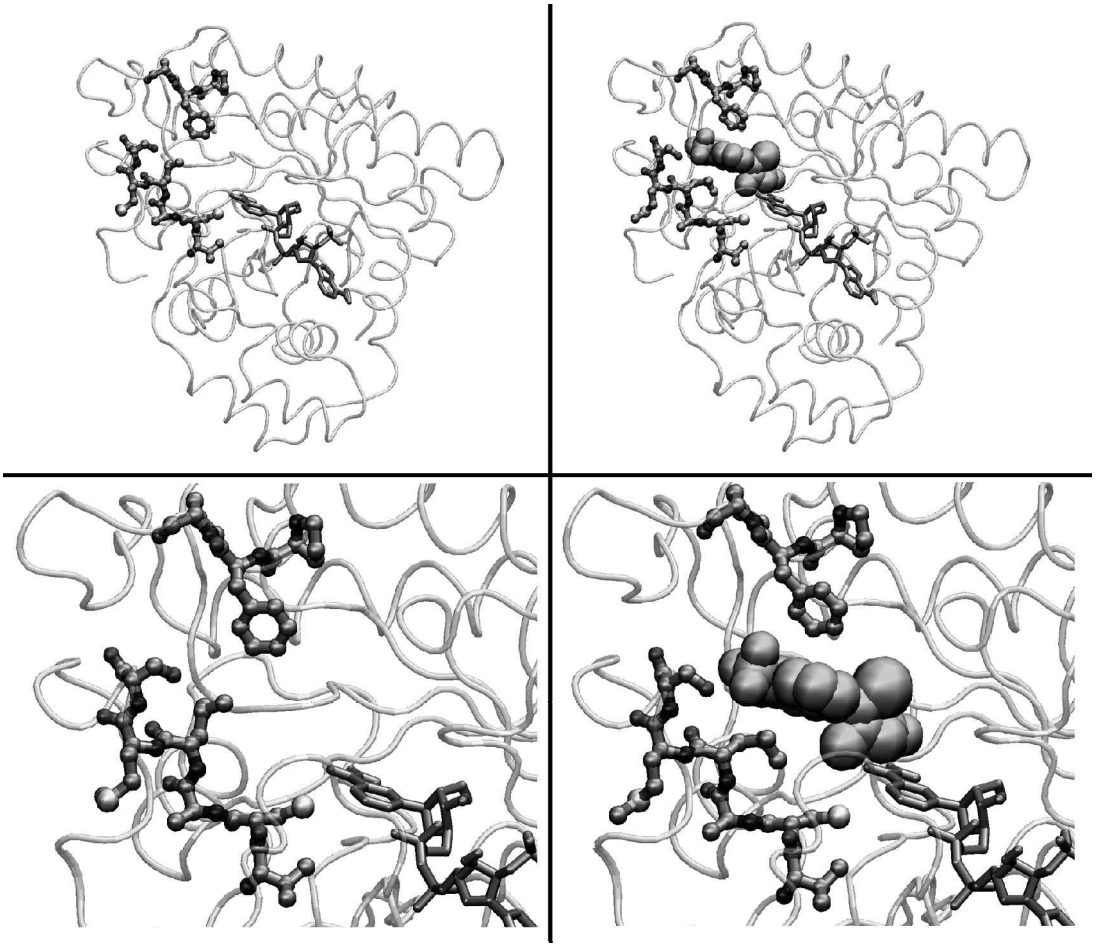

FIG. 7. The unbound form of aldose reductase is shown on the **left** while the bound from is shown on the **right**. The **bottom figures** zoom in the corresponding **top figures** to show aminoacids whose rearrangements open a pocket that make the binding possible.

acids at the top and bottom of the binding site separate and open the extra cavity. The movement is caused by both side chain and backbone rearrangements.

Although there are currently 16 structures of aldose reductase deposited in the PDB, we cannot use this data exclusively as input for the dimensionality reduction technique as we did for HIV-1 protease. The reason is that the data does not contain much variability, as only two ligands are significantly different from the rest. In this case, we have to complement laboratory obtained structures with data obtained from an MD simulation. This will be typically the case with most protein systems of interest for drug design, as few structures have been determined under different sets of experimental conditions and with different ligands. The starting structure for the MD simulation was the unbound form with PDB code 1AH4. The MD and PCA procedures used for this example were similar to what was described above for HIV-1 protease. However, since the system is larger (315 versus 198 aminoacids for HIV-1 protease) and it is known from crystallography observations that only the binding site region changes its shape, we decided to apply our dimensionality reduction technique only to this region. We conservatively defined the binding site to be a sphere of radius 20 Å around the center of the tolrestat specificity pocket (see Fig. 7). This changed the dimensionality of the input data from $3 \times 5{,}121$ to $3 \times 2{,}544$.

Again, as it was observed with HIV-1 protease, there is a large dominance of only a small fraction of the left singular vectors. The 40 most dominant vectors account for 81% of the total variance (Fig. 8), with the first three accounting for 22%, 11%, and 6% of the total variance, respectively. If we consider the new reduced basis of 40 dominant left singular vectors (versus the initial $3 \times 2{,}544$) as a representation for

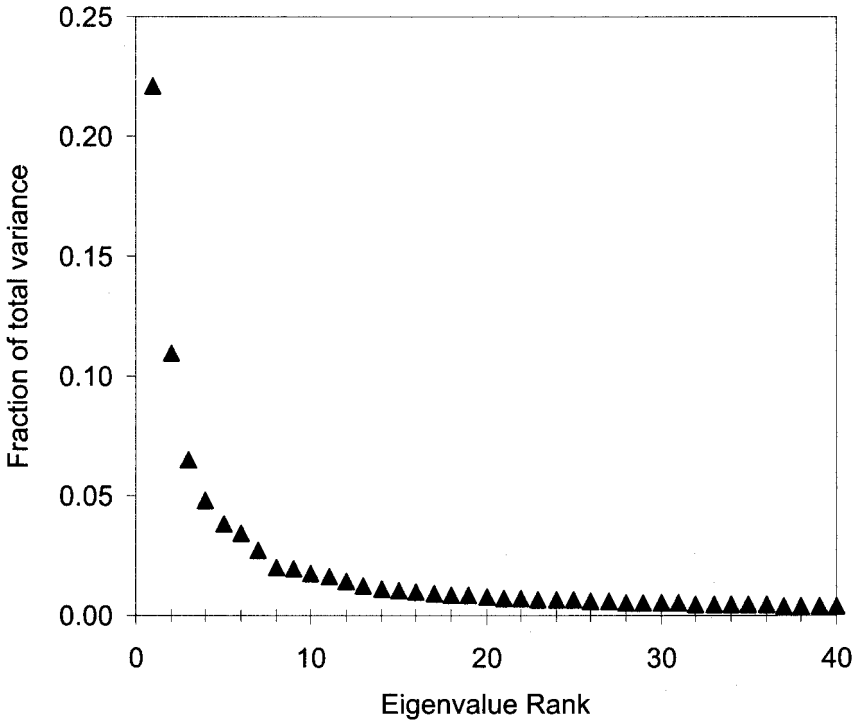

FIG. 8. Fraction of total variance represented by the most significant singular values for aldose reductase dimensionality reduction.

the flexibility of the binding site, we can now model on our reduced basis some of the flexibility that is experimentally observed. For this, we calculated an approximation of the form $[w_1, w_2, \ldots, w_{40}, 0, \ldots, 0]$ as we did for HIV-1 protease using the unbound form as the reference structure (no pocket present) and the bound structure to tolrestat as our target (pocket is present). The RMSD for the binding site residues represented in Fig. 9 between the bound (shown in light gray) and unbound form (shown in black) is 1.74 Å. Using a 40 degrees of freedom approximation, we can reduce this value to 1.07 Å. From the figure, it is clear that we can obtain a good approximation on the residues that form the top of the specificity pocket with our approximate structure (shown in gray) matching almost exactly the experimental structure for the bound form. It is important to note that the approximation is able to capture not only the movement of amino acid sidechains, such as the rotation of the phenylalanine ring shown in the center, but also global displacements caused by a movement at the backbone level, such as shown in the top of Fig. 9. This contrasts with complexity reduction methods that consider the important flexibility of the protein as being represented only by movements of sidechains and which are unable to represent induced fit conformational changes caused by backbone movements. The bottom part of the specificity pocket does not show a match of similar quality but does show a trend in the right direction. In fact, the approximated structure already displays the specificity pocket and is large enough to accommodate ligands such as tolrestat or zopolrestat.

6. CONCLUSION

In this paper, we showed how to obtain a reduced basis representation of protein flexibility. Proteins typically have a few hundreds to a few thousands of degrees of freedom. Starting with data obtained from laboratory experiments and/or MD simulations, we demonstrate that we can compute a new set of degrees of freedom which are combinations of the original ones and which can be ranked according to significance. Depending on the level of accuracy desired, the k most significant of these new degrees of freedom can be used to model the flexibility of the system. We have observed, in multiple occasions, that the reduced basis representation retains critical information about the directions of preferred motion of the protein.

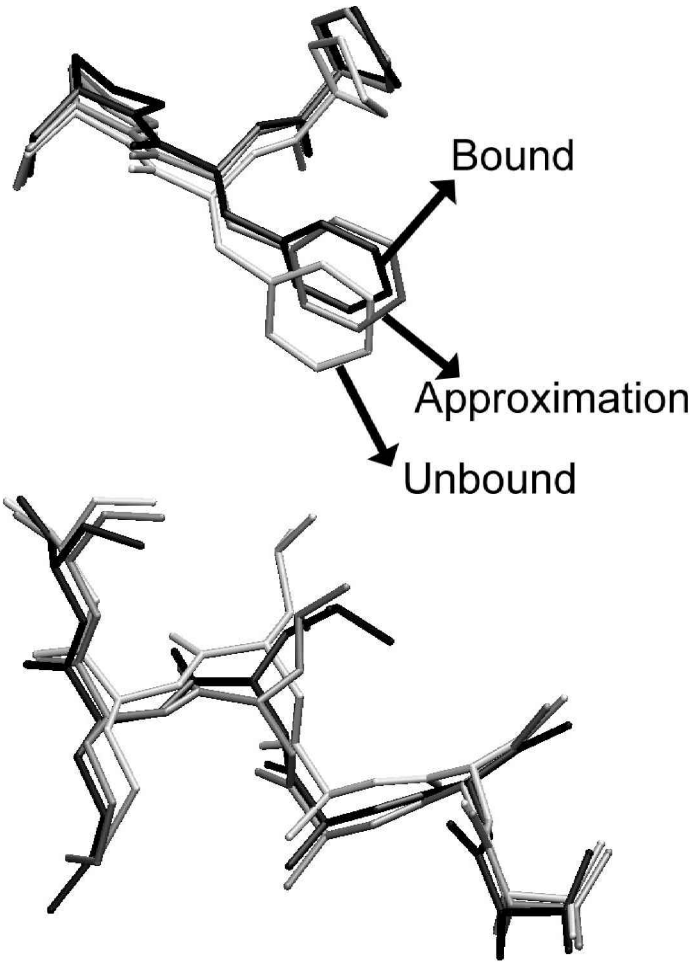

FIG. 9. Approximation (gray) of the bound conformation (black) using the unbound conformation (light gray) as a starting point and searching along main modes of collective motion.

It can thus be used to compute conformational rearrangements of the protein that can further be studied for interaction with novel ligands or other proteins. Our work contributes to the better understanding of how changes in the conformation of a protein affect its ability to bind other molecules and hence of its function. We envision that protein databases, such as the Protein Data Bank, would be annotated in the future with principal modes of motion for proteins allowing rapid and detailed analysis of biomolecular interactions. This annotation would allow researchers to analyze not only the static structure of a protein but also its motions and relations between structure, motion, and function. The process of determining collective modes of motion could be automated using the methods described in the present work, and the information would complement other structural databases, such as the Database of Macromolecular Movements (Gerstein and Krebs, 1998).

In this paper, we used PCA as our dimensionality reduction technique. The results obtained are biologically meaningful. Clearly, it is worth investigating the application of nonlinear dimensionality reduction techniques to the same problem. For example, local PCA (Kambhatla and Leen, 1997), locally linear embedding (Roweis and Saul, 2000), and multilayer auto-associative neural networks (Kramer, 1991) might be able to provide us with the same kind of information as PCA while using an even further reduced number of degrees of freedom. The application of these dimensionality reduction methods to protein structural data is only practical for modeling if we are able to obtain an inverse mapping from the lower to the higher dimensional space. This mapping can in principle be obtained using machine learning techniques such as neural networks. Carrying out this step efficiently is very difficult and constitutes an open research question.

Preliminary work in our group indicates that some of the advantages obtained by performing a nonlinear dimensional reduction are outweighed by an increased computational cost and a loss in accuracy due to the approximated inverse mapping. Nevertheless, the advantage of reduced complexity in the representation of molecular motion may justify the increased computational cost of the nonlinear dimensionality reduction depending on the application.

All our work was done using the Cartesian coordinates of atoms in the protein. An interesting idea is to perform the dimension reduction in the dihedral and the bond angle space of the system. The advantage of this approach is that the initial dimensionality of the problem is reduced because given certain constraints fewer parameters are necessary to uniquely define a protein structure. The first constraint is that bond lengths are fixed. At a second level of approximation, bond angles between three consecutively bonded atoms can also be considered fixed. As such, it is possible to represent a molecular conformation using only the set of dihedral angles corresponding to torsions around single bonds. The dimensionality of this space is in practice approximately almost an order of magnitude smaller than the Cartesian representation. Initial experiments showed that a dihedral angle-based analysis of conformational data is very sensitive to noise and prone to error. We are currently investigating this problem and improving the general methodology in order to apply the methods described in this work to a dihedral angle representation. Last but not least, we investigate how to effectively explore conformational flexibility of a protein in the reduced basis representation to make approximate but fairly accurate predictions for protein–protein and protein–ligand interactions. This work could be used to predict complex effects, such as the induced fit effect during ligand binding in a drug design study, in a computationally efficient manner.

ACKNOWLEDGMENTS

M. Teodoro has been partially supported by a PRAXIS XXI Predoctoral Fellowship from the Portuguese Ministry of Science, a Whitaker Biomedical Engineering Grant, an Autrey Fellowship Award from Rice University, and a Predoctoral Fellowship from the Keck Center for Computational Biology. Work on this paper by L. Kavraki has been supported in part by NSF IRI-970228, NSF 0114796 Grant, a Whitaker Biomedical Engineering Grant, a Sloan Fellowship, and a Texas ATP Award. The authors would like to thank Ara Hayrapetyan, John Olson, and Seiichi Matsuda for their comments.

REFERENCES

Amadei, A., Linssen, A.B., and Berendsen, H.J. 1993. Essential dynamics of proteins. *Proteins* 17, 412–425.

Amadei, A., Linssen, A.B., de Groot, B.L., van Aalten, D.M., and Berendsen, H.J. 1996. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn.* 13, 615–625.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., and Haak, J.R. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684–3690.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucl. Acids Res.* 28, 235–242.

Bishop, C.M., Svensen, M., and Williams, C.K. 1998. GTM: The Generative Topographic Mapping. *Neural Computation* 10, 215–234.

Brooks, C.L., Montgomery, B., and Karplus, M. 1988. *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, John Wiley, New York.

Byrne, B., and Iwata, S. 2002. Membrane protein complexes. *Curr. Opin. Struct. Biol.* 12, 239–243.

Caves, L.S., Evanseck, J.D., and Karplus, M. 1998. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Sci.* 7, 649–666.

Collins, J.R., Burt, S.K., and Erickson, J.W. 1995. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nat. Struct. Biol.* 2, 334–338.

de Groot, B.L., Amadei, A., van Aalten, D.M., and Berendsen, H.J. 1996b. Toward an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. *J. Biomol. Struct. Dyn.* 13, 741–751.

de Groot, B.L., Amadei, A., Scheek, R.M., van Nuland, N.A., and Berendsen, H.J. 1996a. An extended sampling of the configurational space of HPr from *E. coli*. *Proteins* 26, 314–322.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. 1995. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* 117, 5179–5197.

de Groot, B.L., Hayward, S., van Aalten, D.M., Amadei, A., and Berendsen, H.J. 1998. Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. *Proteins* 31, 116–127.

de Groot, B.L., van Aalten, D.M., Amadei, A., and Berendsen, H.J. 1996c. The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophys. J.* 71, 1707–1713.

Duan, Y., and Kollman, P.A. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282, 740–744.

Finn, P., and Kavraki, L. 1999. Computational approaches to drug design. *Algorithmica* 25, 347–371.

Garcia, A.E. 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68, 2696–2699.

Gerstein, M., and Krebs, W. 1998. A database of macromolecular motions. *Nucl. Acids Res.* 26, 4280–4290.

Go, N., Noguti, T., and Nishikawa, T. 1983. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* 80, 3696–3700.

Gogonea, V., Suarez, D., van der Vaart, A., and Merz, Jr., K.M. 2001. New developments in applying quantum mechanics to proteins. *Curr. Opin. Struct. Biol.* 11, 217–223.

Guntert, P., Braun, W., and Wüthrich, K. 1991. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* 217, 517–530.

Hastie, T., and Stuetzle, W. 1989. Principal curves. *J. Am. Statist. Ass.* 84, 502–516.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educational Psychol.* 24, 441.

Isralewitz, B., Gao, M., and Schulten, K. 2001. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* 11, 224–230.

Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* 32, 922–923.

Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. 1999. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comp. Phys.* 151, 283–312.

Kambhatla, N., and Leen, T.K. 1997. Dimension reduction by local principal component analysis. *Neural Computation* 9, 1493–1516.

Kitao, A., and Go, N. 1999. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* 9, 164–169.

Kramer, M.A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* 37, 233–243.

Kruskal, J.B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29, 115–129.

Larson, E.R., Lipinski, C.A., and Sarges, R. 1988. Medicinal chemistry of aldose reductase inhibitors. *Med. Res. Rev.* 8, 159–186.

Lehoucq, R.B., and Sorensen, D.C. 1996. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Analysis and Applications* 17, 789–821.

Lehoucq, R., Sorensen, D.C., and Yang, C. 1998. *Arpack User's Guide: Solution of Large-Scale Eigenvalue Problems With Implicitly Restorted Arnoldi Methods*, SIAM, Philadelphia.

Levitt, M., Sander, C., and Stern, P.S. 1985. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* 181, 423–447.

Levy, R. M., and Karplus, M. 1979. Vibrational approach to the dynamics of an alpha-helix. *Biopolymers* 18, 2465–2495.

MacKerell, A.D., Bashford, D., Bellot, M., Dunbrack, R.L., Jr., Evanseck, J.D., Field, M.J., *et al.* 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102, 3586–3616.

Martin, Y.C., and Willett, P., eds. 1998. *Designing Bioactive Molecules: Three Dimensional Techniques and Applications*, American Chemical Society, Washington D.C.

Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B., and Wlodawer, A. 1989. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science* 246, 1149–1152.

Muegge, I., and Rarey, M. 2001. Small molecule docking and scoring. *Reviews in Computational Chemistry* 17, 1–60.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science* 2, 572.

Rhodes, G. 1993. *Crystallography Made Crystal Clear*, Academic Press, London.

Romo, T. 1998. Identification and modeling of protein conformational substates, p. 235, *Department of Biochemistry and Cell Biology*, Rice University, Houston.

Romo, T.D., Clarage, J.B., Sorensen, D.C., and Phillips, Jr., G.N. 1995. Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins* 22, 311–321.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326.

Rutenber, E., Fauman, E.B., Keenan, R.J., Fong, S., Furth, P.S., Ortiz de Montellano, P.R., Meng, E., Kuntz, I.D., DeCamp, D.L., Salto, R., *et al.* 1993. Structure of a non-peptide inhibitor complexed with HIV-1 protease. Developing a cycle of structure-based drug design. *J. Biol. Chem.* 268, 15343–15346.

Shepard, R.N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika* 27, 125–140.

Tenenbaum, J.B., de Silva, V., and Langford, J.C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323.

Teodoro, M., Phillips, G.N.J., and Kavraki, L.E. 2001. Molecular docking: A problem with thousands of degrees of freedom. *IEEE International Conference on Robotics and Automation.*

Teodoro, M.L., Phillips, G.N., and Kavraki, L.E. 2000. Singular value decomposition of protein conformational motions, 198–199, in Satoru, M., Shamir, R., and Tagaki, T., eds., *Currents in Computational Molecular Biology*, Universal Academy Press, Tokyo.

Tibshirani, R. 1992. Principal curves revisited. *Statistics and Computing* 2, 183–190.

Urzhumtsev, A., Tete-Favier, F., Mitschler, A., Barbanton, J., Barth, P., Urzhumtseva, L., Biellmann, J.F., Podjarny, A., and Moras, D. 1997. A 'specificity' pocket inferred from the crystal structures of the complexes of aldose reductase with the pharmaceutically important inhibitors tolrestat and sorbinil. *Structure* 5, 601–612.

van Aalten, D.M., Conn, D.A., de Groot, B.L., Berendsen, H.J., Findlay, J.B., and Amadei, A. 1997. Protein dynamics derived from clusters of crystal structures. *Biophys. J.* 73, 2891–2896.

Van Eldik, L.J., and Watterson, D.M., eds. 1998. *Calmodulin and Signal Transduction*, Academic Press, London.

Wilson, D.K., Tarle, I., Petrash, J.M., and Quiocho, F.A. 1993. Refined 1.8 A structure of human aldose reductase complexed with the potent inhibitor zopolrestat. *Proc. Natl. Acad. Sci. USA* 90, 9847.

Wolfram, S. 1999. *The Mathematica Book*, Cambridge University Press, New York.

Wüthrich, K. 1986. *NMR of Proteins and Nucleic Acids*, Wiley, New York.

Zacharias, M., and Sklenar, H. 1999. Harmonic modes as variables to approximately account for receptor flexibility in ligand–receptor docking simulations: Application to DNA minor groove ligand complex. *J. Comp. Chem.* 20, 287–300.

Address correspondence to:
Lydia E. Kavraki
Department of Computer Science
Rice University
MS 132
P.O. Box 1892
Houston, TX 77251-1892

E-mail: kavraki@rice.edu