In Proceedings of the Sixth Annual Conference on Computational Molecular Biology RECOMB 2002

A Dimensionality Reduction Approach to Modeling Protein Flexibility

Miguel L. Teodoro[1] George N. Phillips Jr[2] Lydia E. Kavraki[3]

mteodoro@rice.edu phillips@biochem.wisc.edu kavraki@rice.edu

[1] Department of Biochemistry and Cell Biology and Department of Computer Science, Rice University
[2] Department of Biochemistry and Department of Computer Science, University of Wisconsin-Madison
[3] Department of Computer Science and Department of Bioengineering, Rice University

ABSTRACT

Proteins are involved either directly or indirectly in all biological processes in living organisms. It is now widely accepted that conformational changes of proteins can critically 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 few structures, while computational methods such as Molecular Dynamics are too slow for routine use with large systems. A medium sized protein typically has a few thousands of degrees of freedom. This paper shows how to obtain a reduced basis representation of protein flexibility. We 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 the protein. 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.

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 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.

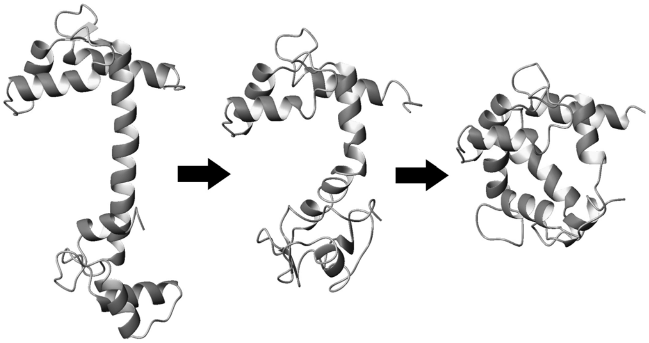

Figure 1. Conformations of calmodulin. The unbound form shown on the left bends the α- helix connecting its two main domains when it binds to a target (not shown). The final bound conformation is on the right.

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 [49]. 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 Figure 1. Many other proteins undergo conformational rearrangements during the course of their function [15].

The possibility of 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 step is known as molecular docking [33]. 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 know as induced fit. 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 [44]. 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. Current methods [42] for studying these interactions are also limited in accuracy and applicability because the molecules involved are modeled as rigid.

Current structural biology experimental methods are considerably restricted in the amount of information they can provide regarding protein motions. The two most common methods in use today are protein X-ray crystallography [36] and nuclear magnetic resonance (NMR) [52]. 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 [6] or quantum mechanics [17] 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 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 numerical integration timestep for 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 [13].

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 a 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 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 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 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 a m-dimensional manifold. The development of new methods is an active research area. The two most commonly used 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 [26, 41] 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 probably the most commonly used technique for dimensionality reduction. This method, which was first proposed by Pearson [35] and further developed by Hotelling [19], 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 non-linear these methods will represent the true reduced manifold in a subspace of higher dimension than necessary in order to cover non-linearity. To overcome this limitation several methods for non-linear dimensional reduction have been proposed in recent years. Among these are principal curves [18, 46], multi-layer auto-associative neural networks [25], local PCA [23], and generative topographic mapping [5]. More recently Tenenbaum et al proposed the isomap method [43] and Roweis and Saul proposed the locally linear embedding method [39]. 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 non-linear versus linear dimensionality reduction methods are increased computational cost, difficulty of implementation, and problematic convergence.

The application of dimensional reduction methods, namely PCA, to macromolecular structural data was first described by Garcia in order to identify high-amplitude modes of fluctuations in macromolecular dynamics simulations [14]. It as also been used to identify and study protein conformational substates [7, 24, 38], as a possible method to extend the timescale of molecular dynamics simulations [1, 2], and as a method to perform conformational sampling [10, 11]. The validity of the method has also been established by comparison with laboratory experimentally derived data [12, 48]. An alternative approach to determine collective modes for proteins uses normal mode analysis [16, 30, 31] and can also serve as a basis for modeling the flexibility of large molecules [53]. The PCA approach described in this article avoids some of the limitations of normal modes such as solvent modeling and existence of multiple energy minima during large conformational transitions. In this paper 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. PCA was the dimensionality reduction method we chose for our study because it is the most established method and efficient algorithms with guaranteed convergence for its computation are readily available. Furthermore, the quality of the dimensional reduction obtained using PCA can be seen as a 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 dimensional reduction technique and protein motion is in general non-linear [14]. Hence, it should be possible to obtain an even lower dimensional representation using non-linear methods. However, we wanted to test the overall approach before proceeding to more expensive methods. Furthermore 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 the 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, they relate to actual protein movements also known as modes of motion. By using the definition of the low dimensional subspace given by the principal components we can readily project the high dimensional data to a low dimensional space and also do it in the inverse direction recovering a representation of the original data with minimal reconstruction error. By contrast the latter operation 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. For non-linear methods, the inverse mapping needs to be obtained using for example a neural network approach but the feasibility and efficiency of this 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 1st principal component $PC_{(1)}$ is a linear combination of the initial variables A_j, with j=1, 2, … , n. That is

$$PC_{(1)} = w_{(1)1}A_1 + w_{(1)2}A_2 + \ldots + w_{(1)n}A_n,$$

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

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

Other principal components $PC_{(p)}$ are similarly linear combinations of the observed variables which are uncorrelated with $PC_{(1)}, \ldots, PC_{(p-1)}$, 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 [21] 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.

SVD - In this work we use the singular value decomposition (SVD) as an efficient computational method to calculate the principal components [37]. 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×m. The left singular vectors of the SVD of A are equivalent to the principal components [37] 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 [38, 45]. 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 [28]. 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 [29].

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 dimensional 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 forcefield 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 [36]. The data that results from this structural determination process corresponds to a single set of coordinates of all atoms in the molecule. X-ray crystallography is sensitive to the experimental conditions under which it was performed and changes of these conditions change the result. Furthermore, it is a difficult, rather expensive, and time-consuming process. However, for a number of interesting molecules, there are several structures available. These structures were either obtained from different laboratories, or correspond to snapshots of the protein bound to different ligands. When such data is available, our analysis can be performed directly on the conformational vector set defined collectively by these structures. This is the case for proteins such as HIV-1 protease.

Nuclear Magnetic Resonance (NMR) - The second most common method of determining the structure of a protein is NMR [52]. This method is in general not as accurate as X-ray crystallography and its use is limited to small to medium sized proteins. However, it provides useful information about protein dynamics directly and avoids the two biggest problems in X-ray crystallography: crystallization and phase determination. 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 dimensional 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 this 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 X-ray crystallography or NMR are stored in major databanks [4].

Molecular Dynamics (MD) – An alternative to using experimental methods to derive structural data is using computational methods such as MD [6]. In fact, computational methods can be employed 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 forcefield [9, 32] to approximate the potential energy surface of a protein. The forcefield measures energy through a combination of bonded terms (bond distances, bond angles, torsional angles, etc.) and non-bonded terms (van der Walls 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 forcefield has been specified, the time evolution of the system at an atomic scale is determined by solving Newton's equations of motion. MD is a good data source for our purposes because it can provide a large number of conformations of a molecule. However, MD is less accurate due to approximations introduced in the computation in order to make the MD simulation computationally practical. Among these approximations is the lack of polarizability representation and the over simplified treatment of solvation effects.

When carrying out the dimensional 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 in part from structural genomics projects. In the next section we apply our method to two model systems using available sources of data.

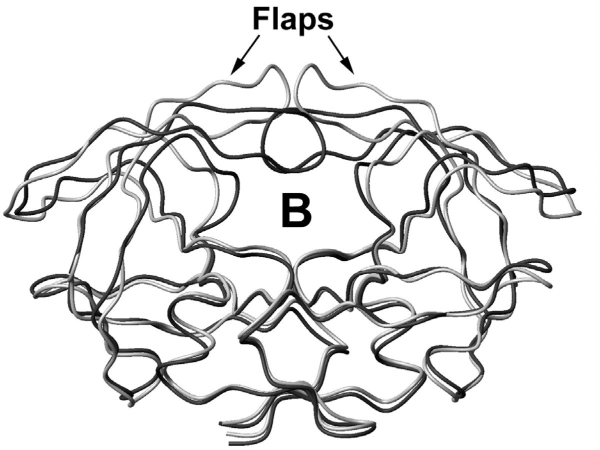

Figure 2. Backbone representation of HIV-1 protease for the unbound (gray) and bound forms (black). The arrows indicate the flaps region where large conformational changes take place. The binding site is the location indicated by B.

5. APPLICATION TO SPECIFIC SYSTEMS

5.1. 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 Figure 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 about this system. This protein is known [8] 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 Figure 2. There are approximately 150 experimental structures publicly available 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 ligand's size and shape, display widely different conformations of the residues in the binding site. It has been observed that the volume of the cavity can approximately double depending on the ligand. 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 3×3120 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.

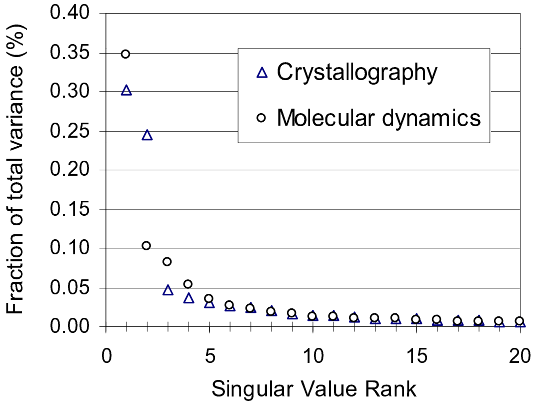

Figure 4. Fraction of total variance represented by the most significant singular values

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 [34] (PDB access code 4HVP). Molecular dynamics simulations for this system was carried out with the program NAMD2 [22] using the charmm22 forcefield [32]. 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.

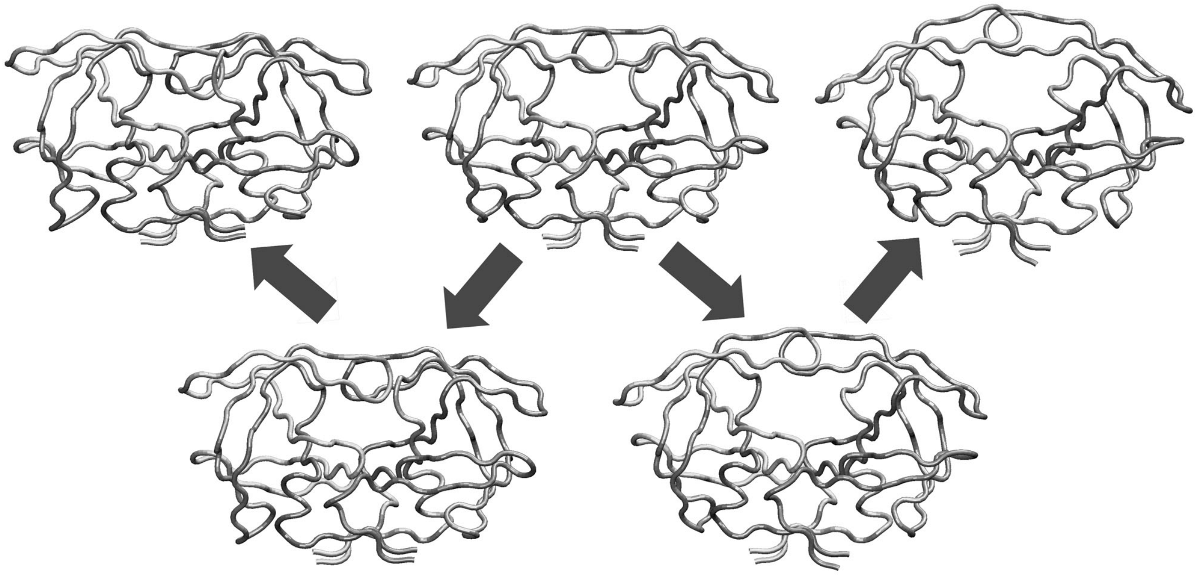

Figure 3. HIV-1 protease backbone motion as defined by the first principal component. The structure shown at the center corresponds to the bound reference structure (ligand not shown). As the structure moves along the first principal component the flaps either close more over the binding site (top left representation) or, if moving in the opposite direction lead, to an open conformation (top right representation) very similar to structures obtained by crystallography. The bottom representations correspond to intermediate states.

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 [3]. 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 14000 structures were used in the dimensional reduction procedure.

In Figure 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%, repectively.

In Figure 3 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 (Figure 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 reveals the tendency of the protein to move in a certain direction even though this movement is not fully explored during the MD run.

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 steered MD technique (SMD) [20]. 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 aminoacid sidechain movement. 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 dimensional 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 dimensional 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 Figure 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%.

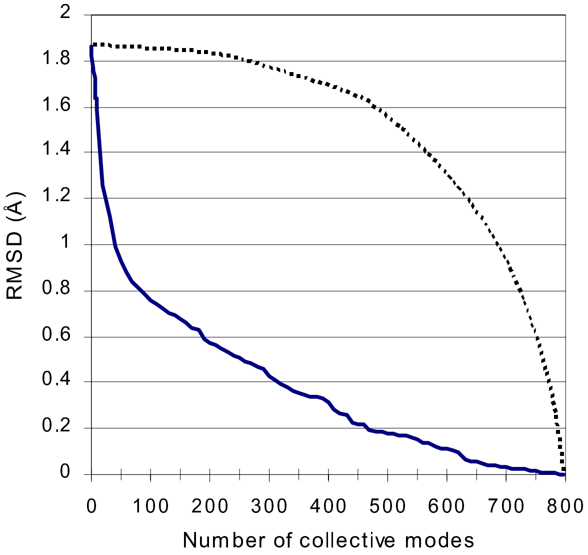

Figure 5. RMSD between a reference and a target structure for an approximation of the flexibility of HIV-1 protease using an increasing number of collective modes. The solid line uses the collective modes basis determined by PCA and the broken line uses a random basis defining the same space.

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 dimensional reduction only for this part of the protein. We defined the binding site atoms as those that are part of aminoacids 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 non-peptide inhibitor [40] (PBD access code 1AID). The inhibitors bound to these structures are considerably different. 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 is shown in Figure 5. 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 [51] defining the same space (shown by a broken line in Figure 5) 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.

5.2. 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 [27]. 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 [47] the structure is almost similar to the unbound form. For other inhibitors, such as the tolrestat [47] and zopolrestat [50], there is a formation of a specificity pocket resulting in significantly different binding site configurations. This conformational change is shown in Figure 6 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 Figure 6 there are a series of aminoacids (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 Walls sphere model) the aminoacids 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 dimensional 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 vs. 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 Figure 6). This changed the dimensionality of the input data from 3×5121

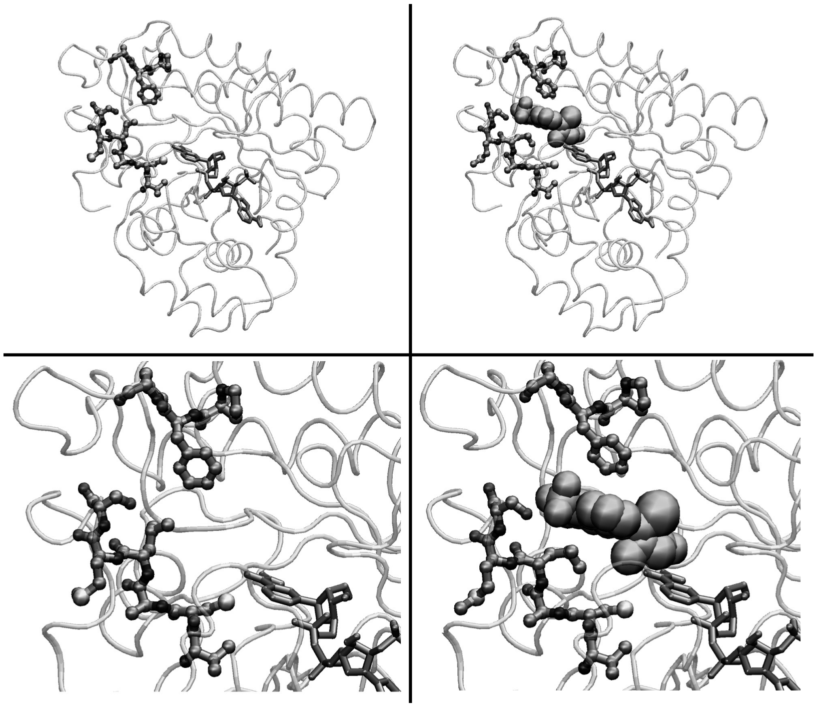

Figure 6. The unbound form of aldose reductase is shown on the left while the bound from is shown on the right. The bottom figures zoom in the corresponding top figures to show aminoacids whose rearrangements open a pocket that make the binding possible.

to 3×2544.

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. If we consider the new reduced basis of 40 dominant left singular vectors (versus the initial 3×2544) as a representation for 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 between the bound (shown in light gray) and unbound form (shown in black) is 1.74Å. Using a 40 degrees of freedom approximation we can reduce this value to 1.07Å. 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. The bottom part of the specificity pocket does not show a match of similar quality but also shows a trend in the right direction.

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. 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 its function. We envision that protein databases would be annotated in the future with principal modes of motion for proteins allowing rapid and detailed analysis of biomolecular interactions.

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 non-linear dimensionality reduction techniques to the same problem. For example local PCA [23], locally linear embedding [39], and multi-layer auto-associative neural networks [25] 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. Carrying out this step efficiently may be difficult and constitutes an open research question. Nevertheless, the advantage of reduced complexity may outweigh the increased computational cost of the non-linear dimensionality reduction.

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. Initial experiments showed that an angle-based analysis of conformational data is very sensitive to noise and prone to error. However our results are not conclusive and further work in this area is needed. Last but not least, we investigate how to effectively explore conformational flexibility of a protein in the reduced basis representation and make approximate but fairly accurate predictions for protein-protein and protein-ligand interactions.

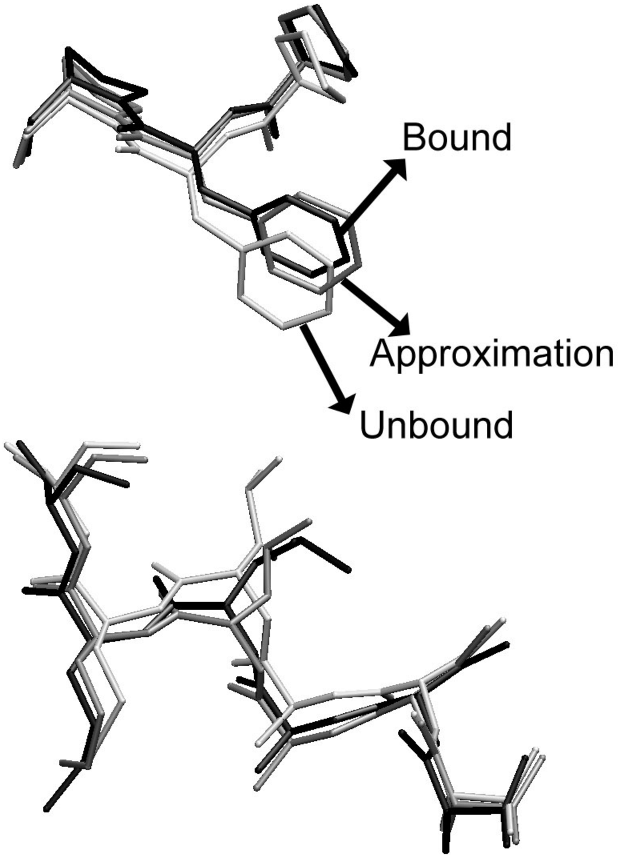

Figure 7. Approximation (gray) of the bound conformation (black) using the unbound conformation (light gray) as a starting point and searching along main modes of collective motion.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Amadei, A., Linssen, A.B. and Berendsen, H.J. Essential dynamics of proteins. *Proteins*, *17* (4). 412-425, 1993

[2] Amadei, A., Linssen, A.B., de Groot, B.L., van Aalten, D.M. and Berendsen, H.J. An efficient method for sampling the essential subspace of proteins. *J Biomol Struct Dyn*, *13* (4). 615-625, 1996

[3] Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. and Haak, J.R. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, *81* (8). 3684-3690, 1984

[4] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res*, *28* (1). 235-242, 2000

[5] Bishop, C.M., Svensen, M. and Williams, C.K. GTM:the Generative Topographic Mapping. *Neural Computation*, *10* (1). 215 -234, 1998

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

[7] Caves, L.S., Evanseck, J.D. and Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci*, *7* (3). 649-666, 1998

[8] Collins, J.R., Burt, S.K. and Erickson, J.W. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nat Struct Biol*, *2* (4). 334-338, 1995

[9] 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. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.*, *117*. 5179-5197, 1995

[10] de Groot, B.L., Amadei, A., Scheek, R.M., van Nuland, N.A. and Berendsen, H.J. An extended sampling of the configurational space of HPr from E. coli. *Proteins*, *26* (3). 314-322, 1996

[11] de Groot, B.L., Amadei, A., van Aalten, D.M. and Berendsen, H.J. Toward an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. *J Biomol Struct Dyn*, *13* (5). 741-751, 1996

[12] de Groot, B.L., Hayward, S., van Aalten, D.M., Amadei, A. and Berendsen, H.J. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins*, *31* (2). 116-127, 1998

[13] Finn, P. and Kavraki, L. Computational Approaches to Drug Design. *Algorithmica*, *25*. 347-371, 1999

[14] Garcia, A.E. Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, *68* (17). 2696-2699, 1992

[15] Gerstein, M. and Krebs, W. A database of macromolecular motions. *Nucleic Acids Res*, *26* (18). 4280-4290, 1998

[16] Go, N., Noguti, T. and Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci U S A*, *80* (12). 3696-3700, 1983

[17] Gogonea, V., Suarez, D., van der Vaart, A. and Merz, K.M., Jr. New developments in applying quantum mechanics to proteins. *Curr Opin Struct Biol*, *11* (2). 217-223, 2001

[18] Hastie, T. and Stuetzle, W. Principal curves. *Journal of the American Statistical Association*, *84*. 502 -516, 1989

[19] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*. 441, 1933

[20] Isralewitz, B., Gao, M. and Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol*, *11* (2). 224-230, 2001

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

[22] Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K. and Schulten, K. NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, *151*. 283-312, 1999

[23] Kambhatla, N. and Leen, T.K. Dimension reduction by local principal component analysis. *Neural Computation*, *9* (7). 1493 -1516, 1997

[24] Kitao, A. and Go, N. Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol*, *9* (2). 164-169, 1999

[25] Kramer, M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, *37* (2). 233 -243, 1991

[26] Kruskal, J.B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*. 115-129, 1964

[27] Larson, E.R., Lipinski, C.A. and Sarges, R. Medicinal chemistry of aldose reductase inhibitors. *Med Res Rev*, *8* (2). 159-186, 1988

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

[29] Lehoucq, R.B. and Sorensen, D.C. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Analysis and Applications*, *17* (4). 789-821, 1996

[30] Levitt, M., Sander, C. and Stern, P.S. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol*, *181* (3). 423-447., 1985

[31] Levy, R.M. and Karplus, M. Vibrational Approach to the Dynamics of an alpha-Helix. *Biopolymers*, *18*. 2465-2495, 1979

[32] MacKerell, A.D., Bashford, D., Bellot, M., Karplus, M. and al, e. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, *102*. 3586-3616, 1998

[33] Martin, Y.C. and Willett, P. (eds.). *Designing bioactive molecules: three dimensional techniques and applications*. American Chemical Society, Washington D.C., 1998.

[34] Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B. and Wlodawer, A. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science*, *246* (4934). 1149-1152, 1989

[35] Pearson, K. 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, 1901

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

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

[38] Romo, T.D., Clarage, J.B., Sorensen, D.C. and Phillips, G.N., Jr. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins*, *22* (4). 311-321, 1995

[39] Roweis, S. and Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290* (5500). 2323--2326, 2000

[40] 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. and et al. Structure of a non-peptide inhibitor complexed with HIV-1 protease. Developing a cycle of structure-based drug design. *J Biol Chem*, *268* (21). 15343-15346, 1993

[41] Shepard, R.N. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, *27* (2). 125-140, 1962

[42] Ten Eyck, L.F., Mandell, J., Roberts, V.A. and Pique, M.E., Surveying molecular Interactions with DOT. in *Supercomputing '95*, (San Diego, California, USA, 1995), IEEE Press.

[43] Tenenbaum, J.B., de Silva, V. and Langford, J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290* (5500). 2319-2323, 2000

[44] Teodoro, M., Phillips, G.N.J. and Kavraki, L.E., Molecular Docking: A Problem with Thousands of Degrees of Freedom. in *IEEE International Conference on Robotics and Automation*, (Seoul, Korea, 2001), IEEE Press.

[45] Teodoro, M.L., Phillips, G.N. and Kavraki, L.E. Singular Value Decomposition of Protein Conformational Motions. in Satoru, M., Shamir, R. and Tagaki, T. eds. *Currents in Computational Molecular Biology*, Universal Academy Press, Inc., Tokyo, 2000, 198-199.

[46] Tibshirani, R. Principal curves revisited. *Statistics and Computing*, *2*. 183 -190, 1992

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

[48] van Aalten, D.M., Conn, D.A., de Groot, B.L., Berendsen, H.J., Findlay, J.B. and Amadei, A. Protein dynamics derived from clusters of crystal structures. *Biophys J*, *73* (6). 2891-2896, 1997

[49] Van Eldik, L.J. and Watterson, D.M. (eds.). *Calmodulin and Signal Transduction*. Academic Press, London, 1998.

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

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

[52] Wuthrich, K. *Nmr of Proteins and Nucleic Acids*. J. Wiley & Sons, New York, 1986.

[53] Zacharias, M. and Sklenar, H. Harmonic Modes as Variables to Approximately Account for Receptor Flexibility in Ligand]Receptor Docking Simulations: Application to DNA Minor Groove Ligand Complex. *Journal of Computational Chemistry*, *20* (3). 287-300, 1999