Chapter 3

Essential Dynamics of Proteins\textsuperscript{1}

3.1 Introduction

Functional proteins are generally stable mechanical constructs that allow certain types of internal motion to enable their biological function. The internal motions may allow the binding of a substrate or coenzyme, the adaptation to a different environment as in specific aggregation, or the transmission of a conformational adjustment to affect the binding or reactivity at a remote site, as in allosteric effects. Such functional internal motions may be subtle and involve complex correlations between atomic motions, but their nature is inherent in the structure and interactions within the molecule. It is a challenge to derive such motions from the molecular structure and interactions, to identify their functional role, and to reduce the complex protein dynamics to its essential degrees of freedom.

We investigate the correlations between atomic positional fluctuations in a protein, as derived from (nanosecond) molecular dynamics

\textsuperscript{1}Most of this chapter was published in: A. Amadei, A.B.M. Linssen and H.J.C. Berendsen, Essential Dynamics of Proteins, \textit{Proteins:} (1993), \textbf{17}, 412-425
(MD), both in vacuum and in aqueous environment. By diagonalizing the covariance matrix of the atomic displacements, we find that most of the positional fluctuations are concentrated in correlated motions in a subspace of only a few (not more than 1%) degrees of freedom, while all other degrees of freedom represent much less important, basically independent, Gaussian fluctuations orthogonal to the "essential" subspace. The motion outside the essential subspace can be considered as essentially constrained. This offers the possibility of representing protein dynamics in the essential subspace only.

Our treatment differs from a harmonic or quasi-harmonic normal mode analysis [27, 28, 29, 30, 31] in two ways. First, we do not analyze the motion but rather the positional fluctuations, without involving the atomic masses in the analysis. Our purpose is to identify an "irrelevant" subspace which may be considered essentially constrained. Second, we do not attempt to describe the motion in the "essential" subspace as harmonic, or even as mutually uncoupled, because it is neither harmonic nor uncoupled and such a treatment would restrict the mechanics of a protein to the level of uninteresting vibrations. The projection of a MD trajectory onto normal mode axes as carried out by Horiuchi and Gō [28] bears a resemblance to our analysis of displacements in the essential subspace, be it that the spaces onto which the motion is projected are not the same: they consider a dihedral angle subspace defined by normal modes of low frequency; we retain Cartesian coordinates and define the subspace from the covariance matrix. They find that the motions in the lower modes are restricted in narrower ranges than those derived by the harmonic approximation; we find a similar restriction due to nonlinear behaviour and the presence of nonlinear constraints within the essential subspace. Our analysis is in fact identical to the one described by Garcia [32]. He found that the largest linearly correlated motions in a protein are defined by the eigenvectors of the covariance matrix with the largest eigenvalue. Also he recognized that these motions are far from harmonic. In fact, his and our approach correspond to Principal Component Analysis (PCA) of the configurational space.
3.2 Theory

We consider the dynamics of a protein in equilibrium in a given environment at temperature $T$. Assume that a trajectory in phase space is available from a reliable MD simulation. We first eliminate the overall translational and rotational motion because these are irrelevant for the internal motion we wish to analyze. The precise method of eliminating the overall motion is not important: either the linear and angular moments are removed every step in the simulation, or the molecular axes are constructed each step by a least-squares translational and rotational fit. The result in any case is a Cartesian molecular coordinate system in which the atomic motions can be expressed. The internal motion is now described by a trajectory $\mathbf{x}(t)$, where $\mathbf{x}$ can represent a subset of atoms. The correlation between atomic motions can be expressed in the covariance matrix $C$ of the positional deviations:

$$
C = \text{cov}(\mathbf{x}) = \langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^T \rangle
$$

(3.1)

where $\langle \rangle$ denote an average over time. The symmetric matrix $C$ can always be diagonized by an orthogonal coordinate transformation $T$:

$$
\mathbf{x} - \langle \mathbf{x} \rangle = T \mathbf{q} \text{ or } \mathbf{q} = T^T(\mathbf{x} - \langle \mathbf{x} \rangle)
$$

(3.2)
which transforms $C$ into a diagonal matrix $\Lambda = \langle qq^T \rangle$ of eigenvalues $\lambda_i$:
\[
C = T\Lambda T^T \quad \text{or} \quad \Lambda = T^T C T
\]  
(3.3)

The $i^{th}$ column of $T$ is the eigenvector belonging to $\lambda_i$. When a sufficient number of independent configurations (at least $3N + 1$) is available to evaluate $C$, there will be $3N$ eigenvalues, of which at least 6 representing overall translation and rotation are nearly zero. When a number of configurations, $S$, less than $3N + 1$, is analyzed, the total number of nonzero eigenvalues is at most $S - 1$ since the covariance matrix will not have full rank (see appendix B).

The matrix $C$ has the property of being always connected to the holonomic constraints of the system. In appendix A we show that a subspace which is forbidden (or almost forbidden) for the motion is always fully defined by a subset of eigenvectors of the matrix $C$ with zero (or approximately zero) eigenvalues. It is also important to note that the probability distribution of the displacements along the eigenvectors, although linearly uncorrelated, is not necessarily statistically independent. On the other hand, if a linear orthogonal transformation defines a subset of statistically independent generalized coordinates, then the unit vectors corresponding to this subset will always be eigenvectors of the covariance matrix $C$. The total positional fluctuation $\sum_i \langle (x_i - \langle x_i \rangle)^2 \rangle$ can be thought to be built up from the contributions of the eigenvectors:

\[
\sum_i \langle (x_i - \langle x_i \rangle)^2 \rangle = \langle (x - \langle x \rangle)^T (x - \langle x \rangle) \rangle = \langle q^T T^T T q \rangle = \langle q^T q \rangle = \sum_i \langle q_i \rangle^2 = \sum_i \lambda_i
\]  
(3.4)

We choose to sort $\lambda_i$ in order of decreasing value. Thus the first eigenvectors represent the largest positional deviations, and most of the positional fluctuations reside in a limited subset of the first $n$ eigenvalues, where $n$ is small compared to a total of $3N$. 
3.3 Methods

Analysis was performed on the trajectories of two distinct simulations of hen eggwhite lysozyme.

A simulation in vacuum was performed by the authors, using the GROMOS simulation package and the GROMOS force field [15]. A starting structure was taken from the Brookhaven Protein Data Bank [36], entry 3LYZ. Including polar hydrogens, the system contained 1258 atoms. Nonpolar hydrogens were incorporated implicitly by the use of united atoms. In total a simulation of 1 ns was performed, with a step size of 2 fs. The temperature was kept at 298 K by coupling to an external temperature bath [37], with a coupling constant \( \tau = 0.01 \) ps. Bond lengths were constrained using the procedure SHAKE [16]. Rotational motion around, and translational motion of the center of mass was removed every 0.5 ps to prevent conversion of thermal motions into overall rotational and translational ones. Nonbonded interactions were evaluated using a short cutoff range of 0.8 nm, within which interactions were calculated every time step. Interactions in the range of 0.8-1.2 nm were updated every 20 fs. During the simulation, configurations were saved every 0.5 ps.

A. Mark kindly offered a 900 ps (100-1000 ps) trajectory of a simulation of lysozyme. This simulation, which included 5,345 water molecules, was performed at 300 K, also using the GROMOS package and the corresponding force field [15]. Here configurations were saved every 0.05 ps. For further details concerning this calculation we refer to Smith et al [38].

Before the covariance matrix was built, all configurations were fitted to the first configuration by first fitting the center of mass and next performing a least square fit procedure [39] on the \( C_\alpha \) coordinates. Covariance matrices \( C \) were constructed from the position coordinates of the atoms (all atoms or \( C_\alpha \) atoms only) according to:

\[
C_{ij} = \frac{1}{S} \sum_t \{x_i(t) - \langle x_i \rangle \} \{x_j(t) - \langle x_j \rangle \}, \tag{3.5}
\]
where $S$ is the total number of configurations, $t = 1, 2, ..., S$, $x_i(t)$ are the position coordinates with $i = 1, 2, ..., 3N$, $N$ is the number of atoms from which $C$ is constructed and $\langle x_i \rangle$ is the average of coordinate $i$ over all configurations. The system contained 129 C$_\alpha$ atoms, having 387 position coordinates. Eigenvalues and their corresponding eigenvectors were calculated using the QL algorithm [40]. Diagonalization of the C$_\alpha$ matrices, of size 387 by 387, required 24 s of CPU time on a single processor of a CONVEX 240, while diagonalizing the all atom covariance matrix of the solvent simulation of size 3792 by 3792 required 20.4 hr on the same machine.

### 3.4 Results and Discussion

Three different covariance matrices were diagonalized. The corresponding eigenvalues are shown in fig. 3.1, plotted in descending order against the corresponding eigenvector indices. Fig 3.1a shows the eigenvalues from the matrix that was constructed from (387) C$_\alpha$ coordinates in the vacuum simulation. In fig. 3.1b we show the eigenvalues as obtained from the (387) C$_\alpha$ coordinates from the solvent simulation. Finally, fig. 3.1c shows the eigenvalues obtained by analyzing the covariance matrix constructed from all atom coordinates (3792) of the protein from the solvent simulation. In this case the first few eigenvalues (mean square displacements) are one order of magnitude larger then in the previous plots, because the number of atoms involved in these displacements is approximately 10 times larger. Since the eigenvalues are mean square displacements, it is clear from fig. 3.1 that the configurational space of the protein is not a homogeneous space, in terms of the motion along the eigenvector directions. As can be seen from fig. 3.1a and b, the eigenvalues from the solvent simulation show a steeper decrease than those from the vacuum simulation. One reason for this may be the fact that the force fields used are not equivalent. In the vacuum force field, full charges have been replaced by dipoles. This produces a weakening of the electrostatic in-
interactions that, as we found, mainly affects the near constraints. As far as the methodology presented here is concerned there are no basic differences between vacuum and solvent simulation; so in the subsequent text we will show only the results obtained from the solvent simulation. At the end of this section the vacuum and solvent results will be compared.

The amount of motion associated to a subspace spanned by the first $n$ eigenvectors can be defined as the corresponding subspace positional fluctuation (eq. 3.4, for the summation over $n$) where the eigenvalues are ordered in descending order. In fig. 3.2 we show this relative subspace positional fluctuation (with respect to the total positional fluctuation) versus the increasing number of eigenvectors that span the subspace. In
Figure 3.2: (a) Relative positional fluctuation (see text) of the motions along the eigenvectors obtained from the $C_\alpha$ coordinates matrix matrix (solvent simulation). (b) Relative positional fluctuation of motions along the eigenvectors obtained from the all atom coordinates covariance matrix (solvent simulation).

Fig. 3.2a we show the results as obtained from the $C_\alpha$ matrix eigenvectors. Fig. 3.2b shows the results from the all atom analysis. From fig. 3.2 it can be seen that 90% of the total motion is described by the first 20 eigenvectors out of 387. If we analyze the motion due to all atoms (fig. 3.2b) we see that the first 35 eigenvectors out of 3792 contribute to 90% of the overall motion. This shows that most of the internal motion of the protein is confined within a subspace of very small dimension.

To have a closer look at the motion along the eigenvector directions one can project the trajectory onto these individual eigenvectors. In fig. 3.3 some projections of the $C_\alpha$ trajectory on the eigenvectors obtained from the $C_\alpha$ covariance matrix are plotted against time. It is clear from
Figure 3.3: Motions along several eigenvectors obtained from the $C_\alpha$ coordinates covariance matrix (solvent simulation)
Figure 3.4: Probability distributions for the displacements along several eigenvectors obtained from the $C_\alpha$ coordinates covariance matrix (solvent simulation). Solid line: Gaussian distributions derived from the eigenvalues of the corresponding eigenvectors. Dashed line: sampling distributions.
Figure 3.5: Motions along several eigenvectors obtained from the all atom coordinates covariance matrix (solvent simulation)
Figure 3.6: Probability distributions for the displacements along several eigenvectors obtained from the all atom coordinates covariance matrix (solvent simulation). Solid line: Gaussian distributions derived from the eigenvalues of the corresponding eigenvectors. Dashed line: sampling distributions.
this figure that all motions that have not yet reached their equilibrium fluctuation belong to the first 10 eigenvectors. Fig. 3.4 shows the observed distribution functions for the displacements along the same eigenvectors, as well as the corresponding Gaussian functions with the same variance and average value. Obviously the only non-Gaussian distributions are again found within the first 10 eigenvectors. Fig. 3.5 and 3.6 show the same, but now projections and distributions have been evaluated using the eigenvectors that were obtained from the all atom covariance matrix. Just as in the case for the Cα analysis we find that all the motions that have not yet reached equilibrium fluctuation are confined within the first 10 eigenvectors. Also the only non-Gaussian distributions appear within the same eigenvectors.

We also noticed a great similarity between the motions along the first few eigenvectors of the Cα matrix and those along the first few eigenvectors derived from the all atom matrix. To investigate this similarity further, we extracted the components from the all atom eigenvectors that corresponded to the Cα coordinates and normalized the vectors that we obtained in this way. In fig. 3.7 the projections of these vectors on the eigenvectors of the Cα matrix are plotted. It is clear that the first 8 extracted vectors correspond to the first 8 Cα matrix eigenvectors. It should be mentioned that the length of these extracted vectors is approximately 20% of the whole length of the corresponding all atom eigenvector, whereas the total number of Cα atoms is about 10% of the total number of atoms in the protein. This indicates that the essential internal motion of the protein mainly involves the backbone atoms. We also noted that the displacements along the first 5 eigenvectors produced a large motion near the active site of the molecule. Fig. 3.8 shows a superposition of 10 sequential projections of the Cα motion onto the first eigenvector, each separated by 100 ps (compare with fig. 3.3). The catalytic site residues Glu-35 and Asp-52 are rigid, but the entrance to the active site cleft, including residues involved in substrate binding (59,62,63,101,107) [41] shows extensive flexibility. This motion, which also involves other loops in the protein, possibly affects the association and dissociation of sub-
Figure 3.7: Absolute value of the projections of the (normalized) extracted vectors coming from the first 10 eigenvectors of the all atom coordinate covariance matrix (see text) on the eigenvectors obtained from the $C_\alpha$ covariance matrix (solvent simulation)

strates and products.

Fig. 3.9 shows the trajectory projected on four planes, each defined by two all atom matrix eigenvectors. In the planes of fig 3.9a and b (respectively, eigenvectors 1 and 2 and eigenvectors 2 and 3) the trajectories are confined within narrower ranges than those expected from independent motions, suggesting the presence of a coupled force field. In fig. 3.9d (eigenvectors 20 and 50) the trajectories fill the expected ranges almost completely. This means that we are dealing with basically independent motions. We analyzed the vacuum simulation and compared the motion in the essential space with that of the solvent simulation. The motion in the vacuum simulation appears to be largely restricted to the carboxy terminal strand; the motion near the active site is no longer present.
Figure 3.8: Superposition of 10 configurations obtained by projecting the C$_\text{a}$ motion onto the first eigenvector. Configurations are separated by 100 ps. Residues involved in the catalytic reaction (35 and 52) and in the binding of the substrate (59, 62, 63, 101, and 107) are indicated.
Figure 3.9: Projection of the trajectory (solvent simulation) on the planes defined by two eigenvectors from the all atom coordinates covariance matrix. (a) Horizontal axis: displacement along first eigenvector. Vertical axis: displacement along second eigenvector. (b) Horizontal axis: displacement along second eigenvector. Vertical axis: displacement along third eigenvector. (c) Horizontal axis: displacement along first eigenvector. Vertical axis: displacement along 50th eigenvector. (d) Horizontal axis: displacement along 20th eigenvector. Vertical axis: displacement along 50th eigenvector. (note the difference in scale)
3.5 Conclusions

The analysis given in this article shows that the essential dynamics of lysozyme, and presumably of other globular proteins, can be described in a subspace of very small dimension (less then 1% of the original Cartesian space) consisting of linear combinations of Cartesian degrees of freedom defined in a molecule-fixed coordinate system. All other degrees of freedom can be considered as corresponding to irrelevant Gaussian fluctuations, behaving like near-constraints. The essential subspace itself is defined by the near-constraints, which are related to the mechanical structure of the molecule in a given conformation. We have strong evidence from inspection of a few proteins studied up to now (lysozyme, thermolysin, and a subtilisin analog) that these motions are related to the functional behavior of the proteins such as opening and closing of the active site and hinge-bending motions between two domains enclosing the active site. The analysis of this behaviour will be the subject of a subsequent study. A (major) conformational change to a different folded conformation may alter the characteristics of the essential subspace, while unfolding will lead to an increase of its dimensionality. The fact that active site motions are not present in the essential space of the vacuum simulation suggests strongly that vacuum simulations are not suitable for the study of biologically relevant motions.

3.6 Acknowledgments

We thank Dr. A.E. Mark from ETH in Zürich for offering a 900 ps trajectory of lysozyme in solvent. We are also grateful to Daan van Aalten who applied the method presented in this paper to a few other proteins. Finally, we gratefully acknowledge the Italian foundation ”Cenci Bolognetti”, Istituto Pasteur, for financially supporting Dr. A. Amadei in this project.