PROTEINS: Structure, Function, and Genetics 36:419–424 (1999)

On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations

Andrea Amadei,* Marc A. Ceruso, and Alfredo Di Nola
Department of Chemistry, University of Rome "La Sapienza," Rome, Italy

ABSTRACT **In this article we present a quantitative evaluation of the convergence of the conformational coordinates of proteins, obtained by the Essential Dynamics method. Using a detailed analysis of long molecular dynamics trajectories in combination with a statistical assessment of the significance of the measured convergence, we obtained that simulations of a few hundreds of picoseconds are in general sufficient to provide a stable and statistically reliable definition of the essential and near constraints subspaces, at least within the nanoseconds time range. Proteins 1999;36:419–424.** © 1999 Wiley-Liss, Inc.

Key words: essential dynamics; molecular dynamics; conformational fluctuations

INTRODUCTION

In this article we used two proteins (protein L and Cytochrome c551) to study in detail the convergence of the definition of the generalized coordinates obtained by the Essential Dynamics (ED) method.[1] The ED method was extensively developed in the group of Prof. H.J.C. Berendsen and proved to be a powerful tool to study protein conformational behavior. In the ED method the covariance matrix of the atomic positional fluctuations, obtained from molecular dynamics (MD) simulations, is constructed and from its eigenvectors a new (generalized) coordinates basis-set is obtained. The eigenvectors associated with large eigenvalues define the configurational subspace where most of the conformational fluctuations occur (essential subspace), while the other eigenvectors, associated with nearly zero eigenvalues, correspond to approximate mechanical-dynamical constraints defining the near constraints subspace. The noise in the definition of the eigenvectors arises from the insufficient configurational sampling of the finite time length simulations. At a given temperature the accessible phase space is filled in by the trajectory according to the specific equations of motion used, and the covariance matrix, built on the ensemble of points of the trajectory after a certain time (sampled covariance matrix), is an estimate of the covariance matrix which could be obtained with an "infinite" time length simulation of the folded protein (the expectation covariance matrix). The accuracy of this estimate depends on the statistical relevance of the configurational subspace sampled within the simulation. Previous articles[2–5] reported evidences that a few hundreds picoseconds simulation is in general sufficient to obtain a reasonable convergence of the essential and near constraints subspaces, but no investigations on the statistical significance of the measured convergence were given. Other articles[6,7] reported evidence of insufficient configurational sampling even in nanoseconds time scale simulations, suggesting that usual simulations could be unable to provide a reliable eigenvectors set. In this article we address in a quantitative manner the convergence and the stability of the eigenvectors, using pairs of independent trajectories of increasing time length for each protein. The statistical significance of the convergence is obtained from the comparison of the inner products distribution of the eigenvectors of one trajectory onto the eigenvectors of the other, with "random" inner products distributions obtained with two possible definitions of the random probability:

- Assuming that any pattern of square projections of one eigenvector of a set onto the eigenvectors of the other set, has an identical probability to occur.
- Assuming a homogeneous probability for the direction of one eigenvector of a set, in the space defined by the eigenvectors of the other set (homogeneous probability for the rotation angles).

In this way for each eigenvector we can decide whether its inner products distribution on the other eigenvectors set can be compatible with the random distribution or not, for a given statistical confidence, and hence evaluate the significance of the eigenvectors sets similarity for increasing simulation time. The results obtained for the two proteins used in this articles show that simulations of a few hundreds picoseconds can provide a stable and statistically reliable definition of the essential and near constraints subspaces, at least within the nanoseconds time range, and in the conclusions we will briefly discuss how this is not necessarily in contradiction with the incomplete configurational sampling observed.

THE RANDOM DISTRIBUTION

In this section we will derive two possible random probability distributions according to two different defini-

Grant sponsor: EC; Grant number: ERBFMRX-CT96-0013; Grant sponsor: Italian Ministero dell'Universita 'e della Ricerca Scientifica e Tecnologica (Progetto Nazionale "Biologia Strutturale").
*Correspondence to: Andrea Amadei, Department of Chemistry, University "La Sapienza," P.O. Box 34 Roma 62, P.le Aldo Moro 5, Rome 00185, Italy. E-mail: amadei@degas.chem.uniroma1.it
Received 5 January 1999; Accepted 18 March 1999

tions of the random basic event. First we will derive the distribution due to the assumption that for a vector any set of square projections onto the reference axis has an identical probability to occur. Secondly we will derive a more usual random distribution based on the assumption that any direction in space for a vector has the same probability. Of course each of these assumptions, although reasonable, is somewhat arbitrary and other possible definitions for the random distribution could be used. Interestingly for a high dimensional space, as in our case, the two random distributions used provide virtually identical results.

The Homogeneous Square Projections Distribution

Let's consider a complete set of eigenvectors which define the whole configurational space, and call these the reference set. In such a multidimensional space we can now consider a new unit vector and evaluate which is the probability distribution of its projection onto a specific subspace of the reference set, if any pattern of square projections of the unit vector has exactly the same probability. If the dimension of the space is M and the subspace dimension is m_1, the probability density of finding a value s of the square projection of the unit vector on the subspace is:

$$\rho(s, m_1, M) = A_0 I(s, n_1) I(1 - s, n_2) \tag{1}$$

where $n_1 = m_1 - 1$, $n_2 = M - m_1 - 1$, and

$$I(s, n_1) = \int_0^s dx_1 \cdot \int_0^{s-x_1} dx_2 \ldots \int_0^{s-x_1-x_2\ldots-x_{n_1-1}} dx_{n_1} = \frac{s^{n_1}}{n_1!} \tag{2}$$

$$I(1 - s, n_2) = \int_0^{1-s} dx_1 \cdot \int_0^{1-s-x_1} dx_2 \ldots \int_0^{1-s-x_1-x_2\ldots-x_{n_2-1}} dx_{n_2} = \frac{(1 - s)^{n_2}}{n_2!}. \tag{3}$$

A_0 is the normalization constant and the x variables are the square projections of the unit vector on the axis (eigenvectors) of the M dimensional space. Note that the multiple integrals provide the "number of ways" in which the unit vector gives a square projection s on the chosen m_1 dimensional subspace and hence a square projection $1 - s$ on its complement. From the previous equations it follows:

$$\int_0^1 \rho ds = A_0 \int_0^1 \frac{s^{n_1}(1 - s)^{n_2}}{n_1! n_2!} ds = \frac{A_0}{(n_1 + n_2 + 1)!} \tag{4}$$

hence $A_0 = (n_1 + n_2 + 1)!$ and:

$$\rho = (n_1 + n_2 + 1)! \frac{s^{n_1}(1 - s)^{n_2}}{n_1! n_2!}. \tag{5}$$

The mode of the distribution (s value corresponding to the maximum of the probability density) is clearly given by:

$$\frac{d\rho}{ds} = \frac{(n_1 + n_2 + 1)!}{n_1! n_2!} \cdot s^{n_1 - 1}(1 - s)^{n_2 - 1}[n_1 - (n_1 + n_2)s] = 0 \tag{6}$$

and hence

$$s_m = \frac{n_1}{(n_1 + n_2)} = \frac{m_1 - 1}{M - 2}. \tag{7}$$

Finally the average (expectation) value of s is:

$$\int_0^1 \rho s ds = \frac{(n_1 + n_2 + 1)!}{n_1! n_2!} \int_0^1 s^{n_1}(1 - s)^{n_2} s ds = \frac{n_1 + 1}{n_1 + 1 + n_2 + 1} = \frac{m_1}{M} \tag{8}$$

and to evaluate the value of s that define a tail of a given total probability, for instance 1 percent, we can use the integral of the probability density:

$$P(s') = \int_0^{s'} \rho ds = 0.99. \tag{9}$$

In the results section we will use this 1 percent criterion to decide whether the square projection of one eigenvector onto a subspace of the reference set is statistically significant ($s > s'$) or can still be considered compatible with the random distribution ($s < s'$).

The Homogeneous Rotation Angles Distribution

We can derive another random probability density of the square projection s if we assume an homogeneous probability for the rotation angles of the unit vector. If θ is the angle of the unit vector with respect to the m_1 dimensional subspace, the probability density in θ is:

$$\rho(\theta, m_1, M) = C|\cos\theta|^{m_1 - 1}|\sin\theta|^{M - m_1 - 1} \tag{10}$$

where C is the normalization constant. In this last equation we simply used the fact that the probability density in θ is proportional to the product of the surface areas of the m_1 and $M - m_1$ dimensional spheres with respectively the absolute values of $\cos\theta$ and $\sin\theta$ as radii. Hence transforming from θ to the square projection $s = (\cos\theta)^2$, via the usual relation for probability density transformations:

$$\rho(s, m_1, M) = \rho(\theta, m_1, M) \frac{d\theta}{ds}$$

we obtain the probability density in s due to Eq. (10):

$$\rho(s, m_1, M) = C' s^{(n_1 - 1)/2} (1 - s)^{(n_2 - 1)/2} \tag{11}$$

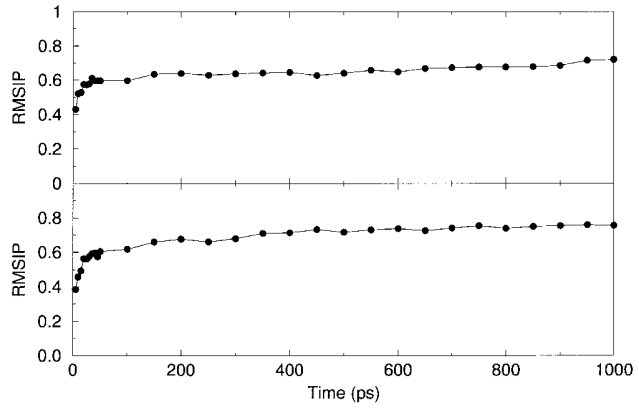

Fig. 1. Root mean square inner product of the essential subspaces (10 eigenvectors), obtained from two independent subparts of the simulation, for protein L (**upper panel**) and Cytochrome (**lower panel**).

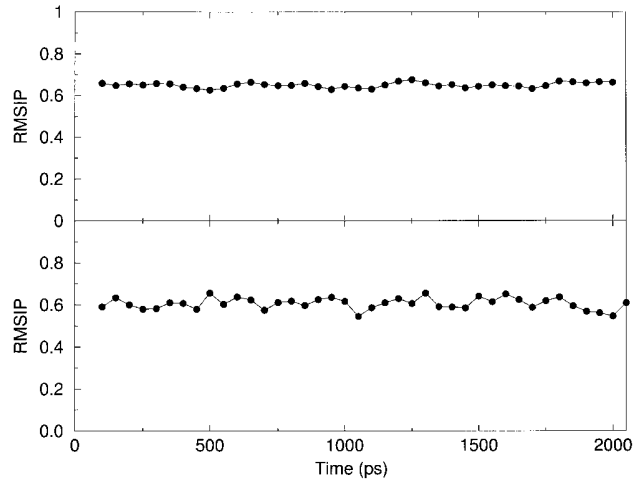

Fig. 2. Overlap between the essential subspaces of subsequent 50-ps slices, for protein L (**upper panel**) and Cytochrome (**lower panel**).

with again C' the normalization constant. It is interesting to note that this new probability density is equivalent to the previous one with n_1 and n_2 changed to $n'_1 = (n_1 - 1)/2$ and $n'_2 = (n_2 - 1)/2$ respectively. Hence substituting in the equations of the previous probability density n_1 and n_2 with n'_1 and n'_2, we have:

$$\langle s \rangle = \frac{n'_1 + 1}{n'_1 + 1 + n'_2 + 1} = \frac{m_1}{M} \tag{12}$$

$$s_m = \frac{n'_1}{n'_1 + n'_2} = \frac{m_1 - 2}{M - 4} \tag{13}$$

showing that the new random probability density has the same average s value of the previous probability density, but a slightly different mode. In practice for large M, as in our case, the two probability densities are very similar and

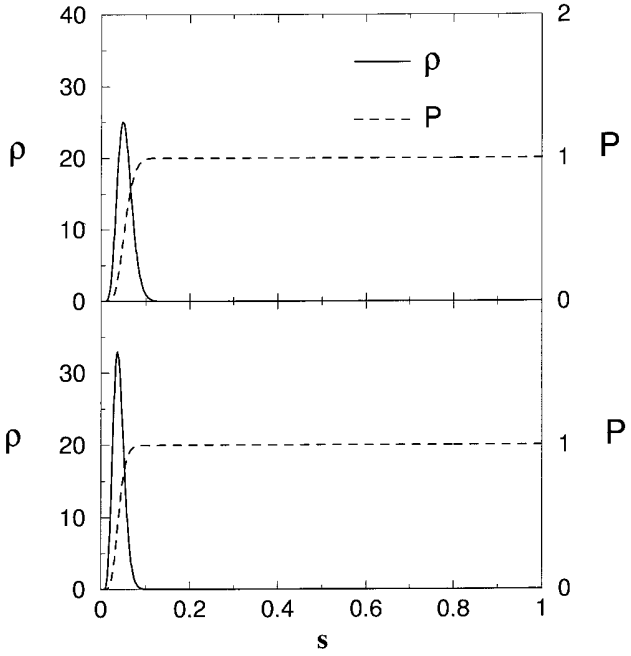

Fig. 3. Probability density (full lines) and total probability (dashed lines) obtained assuming a random square projection distribution of an eigenvector onto a 10-dimensional subspace, for protein L (**upper panel**) and Cytochrome (**lower panel**).

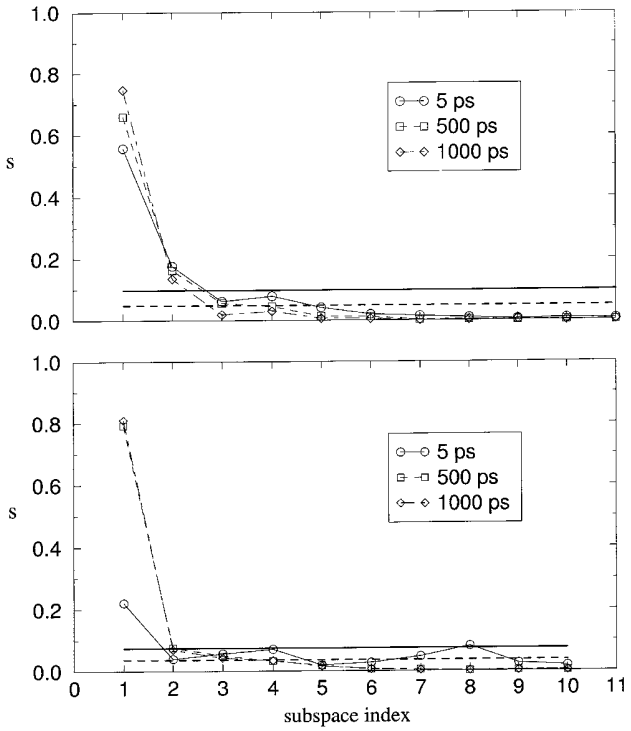

Fig. 4. Square projection, s, of eigenvector 1, obtained from subparts of 5, 500, and 1000 ps, onto the subsequent 10-dimensional subspaces of the reference set, for protein L (**upper panel**) and Cytochrome (**lower panel**). The values s′ (full line) and s_m (dashed line) are also shown.

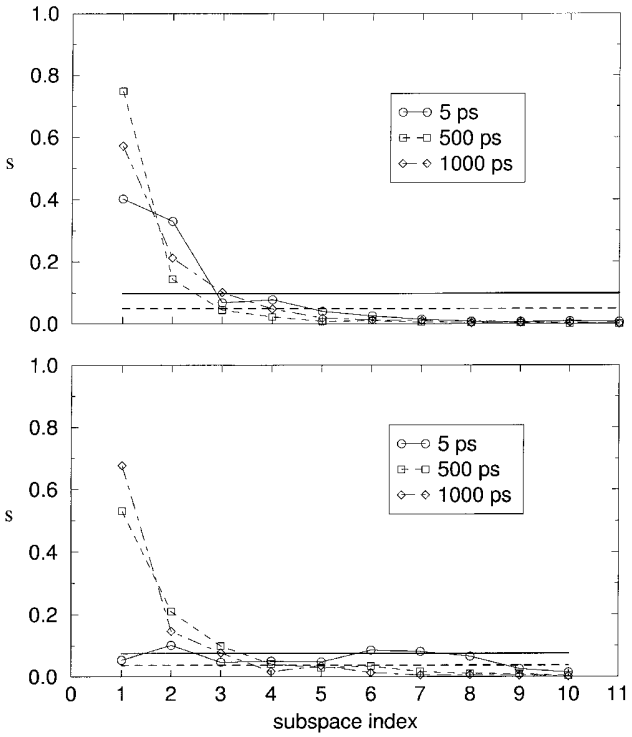

Fig. 5. Square projection, s, of eigenvector 2, obtained from subparts of 5, 500, and 1000 ps, onto the subsequent 10 dimensional subspaces of the reference set, for protein L (**upper panel**) and Cytochrome (**lower panel**). The values s′ (full line) and s_m (dashed line) are also shown.

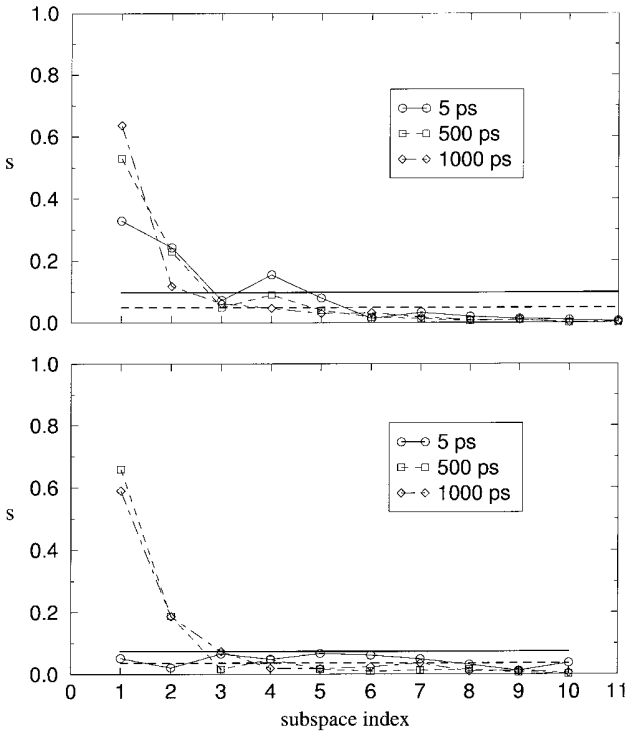

Fig. 6. Square projection, s, of eigenvector 3, obtained from subparts of 5, 500, and 1000 ps, onto the subsequent 10-dimensional subspaces of the reference set, for protein L (**upper panel**) and Cytochrome (**lower panel**). The values s′ (full line) and s_m (dashed line) are also shown.

provide s′ values for the 1 percent criterion, which are virtually identical. For this reason in the analysis described in the results section we only used the probability density given in Eq. (5).

SIMULATION METHODS

The initial proteins configurations were taken from PDB entries 2ptl[8] and 351C for protein L and Cytochrome c 551, respectively.[9] All simulations were performed with the GROMACS simulation package.[10] A modification of the GROMOS87[11] force field was used with additional terms for aromatic hydrogens[12] and improved carbon-oxygen interaction parameters.[13] The SHAKE algorithm[14] was used to constrain bond lengths. Each system was immersed in a pre-equilibrated box of simple point charge (SPC) water.[15] There were 1,971 water molecules for protein L and 2,955 for Cytochrome c 551. Both simulations were 2.3 ns long, and the first 0.30 ns of each simulation were discarded in order to ensure equilibration. Molecular dynamics simulations were initiated as follows: using a restraining harmonic potential, all heavy atoms of the protein were constrained to their initial positions, while surrounding SPC water molecules were first minimized and then submitted to 5 ps of constant volume MD at 300 K. The resulting system was then minimized, without any constraints, before starting constant temperature and constant volume molecular dynamics. A nonbonded cutoff of 1.2 nm was used for both Lennard-Jones and Coulomb potentials. The pair lists were updated every ten steps. A constant temperature of 300 K was maintained by coupling to an external bath[16] using a coupling constant ($\tau = 0.002$ ps) equal to the integration time step.

RESULTS AND CONCLUSIONS

For each of the two proteins (protein L and Cytochrome c 551) we used a 2-ns simulation from which different types of analyses were performed. The eigenvectors were always obtained from the covariance matrix of the alpha carbons only. To evaluate the convergence of the essential eigenvectors in time we divided the whole trajectory into two halves, from which pairs of subparts of increasing time length were taken. In order to prevent any correlation we always compared subparts from the first half starting from the initial point of the trajectory, with the subparts of the second half ending in the last point of the trajectory. In Figure 1 the overlap of the essential subspaces obtained from pairs of subparts ranging from 5 ps to the whole half (1 ns), is shown for the two proteins. As usual the essential subspace of each subpart was defined by the ten eigenvectors with the largest eigenvectors, and the overlap between the essential subspaces was obtained from the root mean square inner product (RMSIP) of the essential eigenvectors of one subpart with the essential eigenvectors

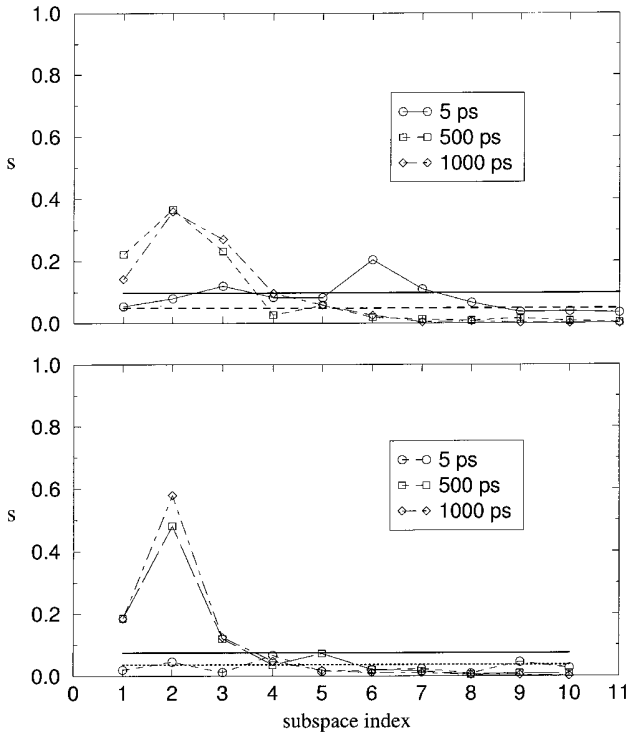

Fig. 7. Square projection, s, of eigenvector 15, obtained from subparts of 5, 500, and 1000 ps, onto the subsequent 10-dimensional subspaces of the reference set, for protein L (**upper panel**) and Cytochrome (**lower panel**). The values s′ (full line) and s_m (dashed line) are also shown.

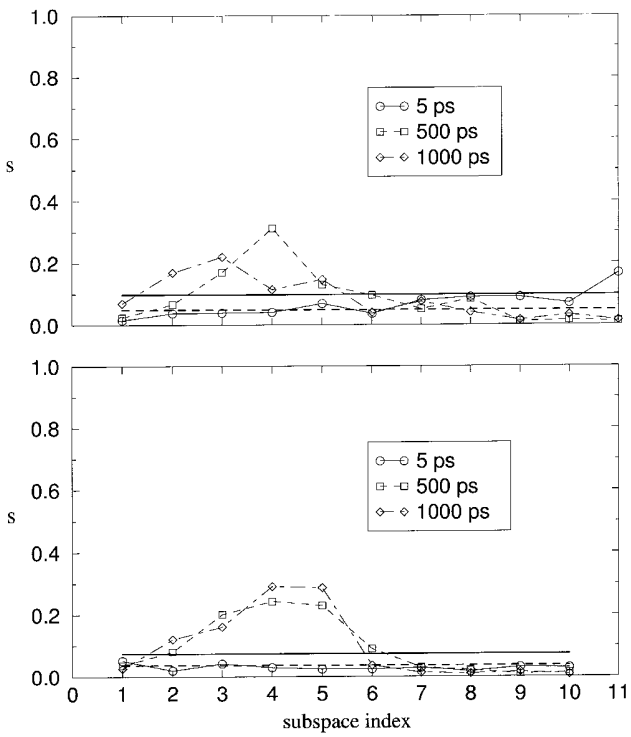

Fig. 8. Square projection, s, of eigenvector 35, obtained from subparts of 5, 500, and 1000 ps, onto the subsequent 10-dimensional subspaces of the reference set, for protein L (**upper panel**) and Cytochrome (**lower panel**). The values s′ (full line) and s_m (dashed line) are also shown.

of the other subpart

$$\mathrm{RMSIP} = \left(\frac{1}{10}\sum_{i=1}^{10}\sum_{j=1}^{10}(\boldsymbol{\eta}_i \cdot \boldsymbol{\nu}_j)^2\right)^{1/2}$$

where $\boldsymbol{\eta}_i$ and $\boldsymbol{\nu}_j$ are the eigenvectors of the two subparts. Figure 1 clearly shows that within 50 to 100 ps a relevant convergence was reached with an overlap of about 0.6, while in the range from 100 ps to 1 ns a much slower further convergence was present reaching a final overlap of about 0.75. Interestingly for both proteins 5 ps were enough to obtain an overlap of about 0.4 meaning that for the essential eigenvectors the initial convergence was extremely fast. This is in agreement with recent results on the kinetics of the essential coordinates.[17]

To investigate the reproducibility of these results we calculated the overlap between the essential subspaces of contiguous 50-ps slices of the whole trajectory (comparison of not contiguous 50-ps slices was also done giving identical results). In Figure 2 these overlaps (RMSIP) are shown for both proteins, demonstrating that the amount of overlap between two 50 ps simulations is stable at about 0.6, as also obtained comparing the first and the last 50 ps of the trajectory (see Fig. 1). These results show very clearly that the definitions of the essential and near constraints subspaces converge at least up to the nanoseconds time range.

As described in the previous section, in this article we addressed in a quantitative way the problem of the statistical significance of the observed convergence of the eigenvectors sets. We used for both proteins the complete first half of the trajectory (1 ns) to obtain a reference set, that we compared with the eigenvectors obtained from the subparts of increasing time length, taken from the second half. We analyzed the square projections of a single eigenvector of a subpart, on the subsequent 10 dimensional subspaces of the reference set. In Figure 3 the probability density, with $m_1 = 10$, and its integral (total probability) are shown for the two proteins. Note that the dimension M of the space is M = 186 and M = 246 for protein L and Cytochrome c551 respectively. In Figures 4–9 we show, for the two proteins, the square projections s of eigenvectors 1, 2, 3, 15, 35, and 65, obtained from subparts of 5,500, 1000 ps, onto consecutive 10-dimensional subspaces of the reference set. Note that the eigenvectors are ordered according to the size of the corresponding eigenvalue, and hence the first 10 dimensional subspace is the essential subspace. In the figures the value s′ corresponding to the 1 percent criterion, and the mode value s_m are also shown. From these figures it is evident that the eigenvectors of the 500- and 1000-ps subparts have always an s value much larger than s′ only in the corresponding subspace of the reference set, the 10 dimensional subspace which contains the reference set eigenvector with the same index of the eigenvector of the subpart, or in its neighbor subspaces. In all the other subspaces the

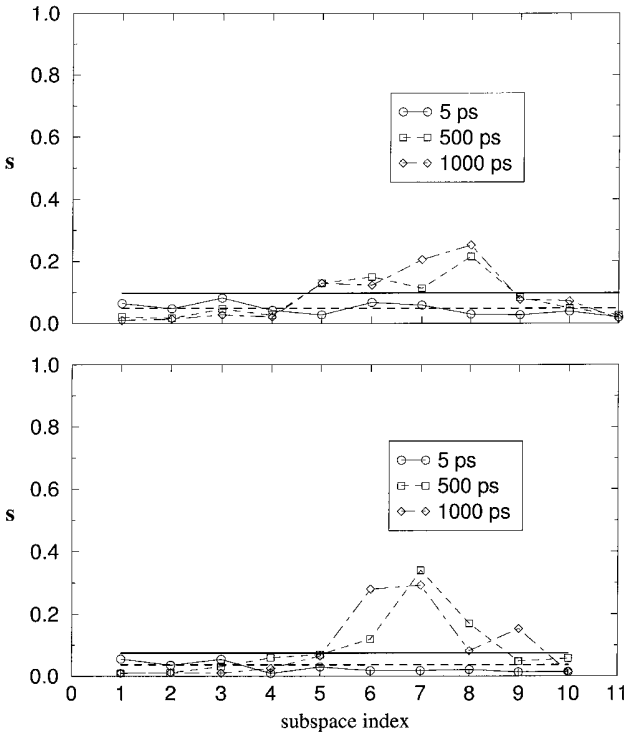

Fig. 9. Square projection, s, of eigenvector 65, obtained from subparts of 5, 500, and 1000 ps, onto the subsequent 10-dimensional subspaces of the reference set, for protein L (**upper panel**) and Cytochrome (**lower panel**). The values s′ (full line) and s_m (dashed line) are also shown.

s value is lower or very close to s′. This clearly means that using the 1 percent confidence, within the 10-dimensional subspace sensitivity, we must consider the convergence for simulations of 500 ps or more as statistically significant, and that simulations of a few hundreds of picoseconds can provide a statistically reliable approximation of the eigenvectors which can be obtained by a simulation of a few nanoseconds. Interestingly, for the subparts of only 5 ps most of the eigenvectors have s values on all the subspaces around the mode s_m.

All the results described show that the eigenvectors converge in time toward a "stable" set, at least up to the nanoseconds time range, and that such a convergence is statistically significant. Whether this "stable" set is really stable beyond the nanoseconds time range, and coincides with the expectation set, is still an open question. However it seems reasonable to the authors that such a clear and statistically significant covergence of the eigenvectors up to the nanoseconds range, is indicative of the fact that simulations in the range from a few hundreds of picoseconds to a few nanoseconds, can already provide a reasonable definition of the essential and near constraints subspaces, valid beyond the nanoseconds range. Finally it is important to note that the observed convergence for the essential and near constraints subspaces does not imply a good sampling of the configurational space, but simply means that the trajectory had time enough for covering a large amplitude in the essential subspace. In fact simulations in the nanoseconds time scale provide a good sampling only for the near constraints subspace while the essential subspace still remains poorly filled in by the trajectory, and the total fluctuation (trace of the covariance matrix), as well as the exact definition (i.e., with a 1-dimensional subspace sensitivity) of each single eigenvector are still not well converged in the nanoseconds time range.[3,6,7] On the contrary the typical time for a simulation to discriminate between the essential and the near constraints subspaces is within a few hundreds of picoseconds as clearly shown in this article.

ACKNOWLEDGEMENTS

It is acknowledged that the Istituto Pasteur Fondazione Cenci Bolognetti provided financial support to one of the authors.

REFERENCES

1. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins 1993;17:412–425.
2. Amadei A, Linssen ABM, De Groot BL, Van Aalten DMF, Berendsen HJC. An efficient method for sampling the essential subspace of proteins. J Biomol Struct Dyn 1996;13:615–626.
3. De Groot BL, Amadei A, Scheek RM, Van Nuland NAJ, Berendsen HJC. An extended sampling of the configurational space of HPr from E. coli. Proteins 1996;26:314–322.
4. De Groot BL, Van Aalten DMF, Amadei A, Berendsen HJC. The consistency of large concerted motions in proteins in molecular dynamics simulations. Biophys J 1996;71:1554–1566.
5. Van Aalten DMF, De Groot BL, Berendsen HJC, Findlay JBC, Amadei A. A comparison of techniques for calculating protein essential dynamics. J Comp Chem 1997;18:169–181.
6. Clarage J, Romo T, Andrews BK, Montgomery B, Phillips GN. A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci USA 1995;92:3288–3292.
7. Hunenberger PH, Mark AE, Van Gunsteren WF. Fluctuations and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. J Mol Biol 1995;252:492–503.
8. Wikstrom M, Drakenberg T, Forsen S, Sjobring U, Bjorch L. Three-dimensional structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. Biochemistry 1994;33:14011–14017.
9. Matsuura Y, Takano T, Dickerson RE. Structure of cytochrome C551 from P. aeruginosa refined at 1.6 Å resolution and comparison of the two redox forms. J Mol Biol 1982;156:389–405.
10. Van der Spoel D, Berendsen HJC, Van Buuren AR, et al. Gromacs user manual. University of Groningen, The Netherlands. (Internet: http://rugmd0.chem.rug.nl/~gmx) 1995.
11. Van Gunsteren WF, Berendsen HJC. Gromos manual. BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, The Netherlands; 1987.
12. Van Gunsteren WF, Billeter SR, Eising AA, et al. Biomolecular simulation: the GROMOS96 manual and user guide. Biomos b.v., Zürich, Groningen, The Netherlands; 1996.
13. Van Buuren AR, Marrink SJ, Berendsen HJC. A molecular dynamics study of the decane/water interface. J Phys Chem 1993;97:9206–9212.
14. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints; molecular dynamics of n-alkanes. J Comp Phys 1977;23:327–341.
15. Berendsen HJC, Postma JPM, Van Gunsteren WF, Hermans J. Interaction models for water in relation to protein hydration. In: Pullman B, editor. Intermolecular Forces, D. Reidel Publishing Company, Dordrecht; 1981. p 331–342.
16. Berendsen HJC, Postma JPM, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690.
17. Amadei A, De Groot BL, Ceruso MA, Paci M, Di Nola A, Berendsen HJC. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. Proteins 1999;35:283–292.