PROTEINS: Structure, Function, and Genetics 35:283–292 (1999)

RESEARCH ARTICLES

A Kinetic Model for the Internal Motions of Proteins: Diffusion Between Multiple Harmonic Wells

A. Amadei,[1,2] B.L. de Groot,[1] M.-A. Ceruso,[2] M. Paci,[3] A. Di Nola,[2] and H.J.C. Berendsen [1]*

[1]*Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Biophysical Chemistry, the University of Groningen, Groningen, The Netherlands*

[2]*Department of Chemistry, University of Rome "La Sapienza," Rome, Italy*

[3]*Dipartimento di Scienze e Tecnologie Chimiche, Università di Roma "Tor Vergata," Rome, Italy*

ABSTRACT **The dynamics of collective protein motions derived from Molecular Dynamics simulations have been studied for two small model proteins: initiation factor I and the B1 domain of Protein G. First, we compared the structural fluctuations, obtained by local harmonic approximations in different energy minima, with the ones revealed by large scale molecular dynamics (MD) simulations. It was found that a limited set of harmonic wells can be used to approximate the configurational fluctuations of these proteins, although any single harmonic approximation cannot properly describe their dynamics.**

Subsequently, the kinetics of the main (essential) collective protein motions were characterized. A dual-diffusion behavior was observed in which a fast type of diffusion switches to a much slower type in a typical time of about 1–3 ps. From these results, the large backbone conformational fluctuations of a protein may be considered as "hopping" between multiple harmonic wells on a basically flat free energy surface. Proteins 1999;35:283–292.

Key words: collective motions; conformational fluctuations; molecular dynamics; essential dynamics

INTRODUCTION

The mechanical and dynamical characterization of the conformational space of proteins is a challenging task for molecular biophysics since it is a prerequisite for the understanding of protein behavior and folding processes. Up to now, such investigations have been limited even from a computational point of view, since the high dimensionality of a protein's configurational space, and the complexity of the model potential which is necessary to describe the protein-protein and protein-solvent interactions, make the computational costs enormous. Recently, a set of equivalent theoretical methods have been proposed[1,2] to analyze protein molecular dynamics (MD) trajectories in order to separate the mechanical-dynamical constraints from the "essential degrees of freedom" that are responsible for all the relevant structural transitions in these molecules. In these methods the covariance matrix of the atomic positional fluctuations obtained from a MD trajectory is constructed. Every linear constraint in the configurational space is associated with an eigenvector of this matrix with a (nearly) zero eigenvalue, and the only relevant directions of fluctuation are the eigenvectors with large eigenvalues.[2] The orthonormal set of eigenvectors can be used as a new basis set for generalized coordinates. Typically, for a protein, only about ten eigenvectors (essential eigenvectors) are sufficient to describe the most large concerted motions in the system, and the rest can be considered as approximated constraints responsible for small "harmonic" fluctuations. This procedure is equivalent to a principal components analysis of the protein trajectory in configurational space. In contrast with quasi-harmonic analysis[3–6] we do not require a harmonic approximation at all, and any subset of atoms can be used.

This approach (often referred to as essential dynamics) has been applied to several different proteins[2,7–9] always revealing a low-dimensional essential subspace, and in some cases the essential motions obtained could be related to functional properties of the protein. The existence of such a low-dimensional essential subspace proved to be useful in the design of a new procedure (essential dynamics sampling) able to provide a conformational sampling which is more efficient than the one obtained by usual MD.[10–12] In the present paper, we investigate two small proteins, initiation factor 1[13] and B1 domain of protein G,[14] characterizing the mechanics and the kinetics in the essential subspace in detail.

The first part of the paper concerns the comparison of the configurational properties obtained from harmonic

Grant sponsor: Menarini Ricerche; Grant sponsor: Istituto Pasteur Fondazione Cenci Bolognetti; Grant sponsor: Ministero dell'Universita'e della Ricerca Scientifica (MURST), Project on Structural Biology.

*Correspondence to: H.J.C. Berendsen, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Biophysical Chemistry, the University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands. Email: berendsen@chem.rug.nl

Received 1 July 1998; Accepted 6 November 1998

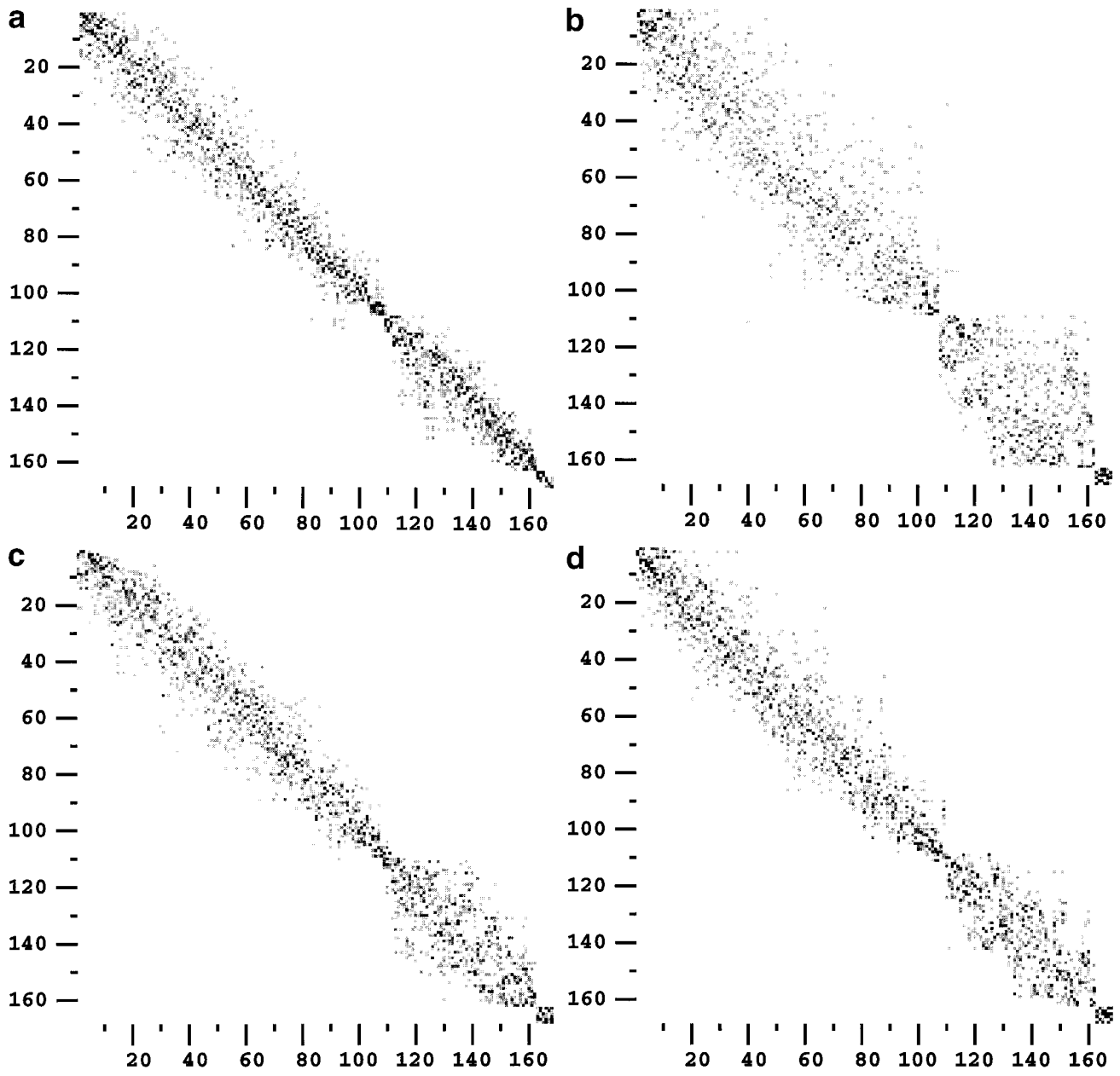

Fig. 1. Squared inner product matrices (all plots are for the B1 domain of protein G). **a:** Eigenvectors obtained from two halves of the 1.2 ns MD trajectory compared to each other. **b:** Eigenvectors obtained from the Hessian matrix built at the initial position compared to eigenvectors obtained from the 1.2 ns MD trajectory. **c:** Eigenvectors obtained from the combination of all (20) Hessian calculations compared to eigenvectors obtained from the 1.2 ns MD trajectory. **d:** Eigenvectors obtained from the combination of a set of 10 closely related Hessian calculations compared to eigenvectors obtained from the 1.2 ns MD trajectory. **e:** Eigenvectors obtained from the combination of a set of 10 Hessian calculations started from MD structures, 100 ps apart, compared to eigenvectors obtained from the 1.2 ns MD trajectory.

analyses of multiple minima with those from MD. Similar to early studies of energy landscapes of Lennard-Jones fluids by Stillinger and coworkers,[15] configurations were extracted from MD simulations which were subsequently energy-minimized. We collected different energy minima of the two test proteins, located in different positions of the essential subspace. A local harmonic approximation as in Normal Modes Analysis (NMA)[16–18] was utilized to compare the different minima to each other. Such kind of comparison has been presented on proteins by other investigators,[19–21] but with two major differences. First, we use the true Hessian (not including the masses) to describe the local harmonic behavior of the system. Second, we include a shell of water in the Hessian calculation in order to approach a physical system more closely than calculations in vacuo do. It was found that the essential subspace obtained by MD showed a high degree of overlap with the one obtained by combining the harmonic dynamical behavior of multiple minima.

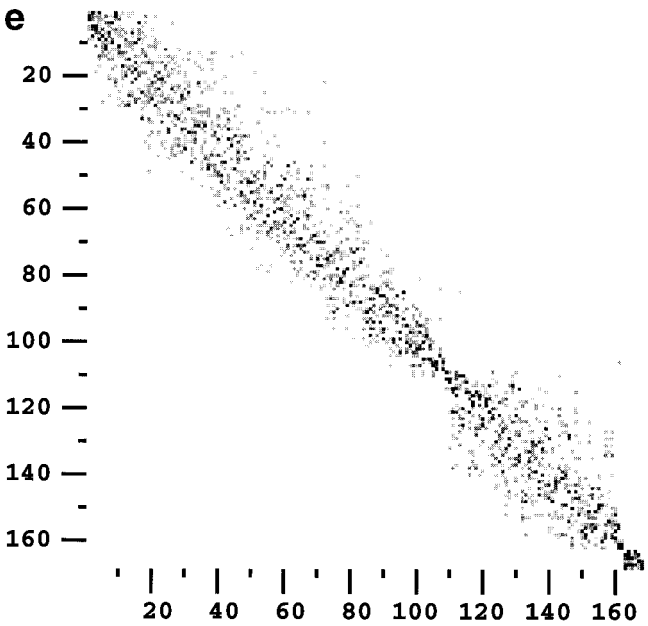

Figure 1. (Continued.)

The second part of this paper concerns the study of the kinetics in the essential subspace. Starting from a common position for the first three essential coordinates, many different runs of MD were produced in order to calculate the average square displacement in time along these coordinates. As previously observed,[10] we found a diffusion-like behavior for the essential coordinates in the time range studied (up to 30 ps) with a deviation within the first two picoseconds.

In this paper, we propose a simple diffusion model (just beyond the Einstein-Smolukowsky approximation) which seems to properly describe these data, and that in combination with the results on the local harmonic fluctuations, suggests as a possible simple physical model a slow diffusion between multiple harmonic wells for the kinetics in the essential subspace.

METHODS

All simulations were performed with the GROMACS simulation package.[22] A modification[23] of the GROMOS87[24] force field was used with additional terms for aromatic hydrogens[25] and improved carbon-oxygen interaction parameters.[23] SHAKE[26] was used to constrain bond lengths, allowing a time step of 2 fs. The essential dynamics analyses were performed always only on C-alpha coordinates.

B1 Domain of Protein G

For the B1 domain of protein G we have used a simulation of 1.2 ns. The initial configuration for this simulation was the crystallographic structure (PDB entry 1pgb[14]). Prior to simulation, the protein was solvated in a box of SPC[27] water molecules. Four sodium ions were added to compensate for the net negative charge of the molecule (the ions were added by replacing water molecules at the lowest electrostatic potentials), resulting in an electrically neutral box containing 1,555 water molecules adding up to a total of 5,231 atoms. One hundred (100) steps of energy minimization using a steepest descents algorithm and a MD simulation of 10 ps with position restraints on the protein (force constant of 1,000 kJ mol^{-1} nm^{-2}) were performed for structural regularization and an initial equilibration of the water molecules. A twin-range cut-off method was used for non-bonded interactions. Lennard-Jones and Coulomb interactions within 1.0 nm were calculated every step, whereas Coulomb interactions between 1.0 and 1.35 nm were calculated every ten steps. This simulation was performed at fixed volume after an isobaric equilibration. The temperature was regulated by weak coupling to a temperature bath[28] (τ = 0.1 ps) with a reference temperature of 300 K.

For the study of diffusional properties, another 50 simulations of 100 ps were performed from conformations produced by a MD simulation of 50 ps, taking structures 1 ps apart. For each of these simulations, different initial velocities were chosen from a Maxwellian distribution to make sure that independent trajectories were generated. The last 50 ps from each of these simulations were added to pieces of trajectory collected from the 1.2 ns simulation and used for the analysis of the diffusional properties of the essential degrees of freedom (see Appendix B).

From the simulation of 1.2 ns ten structures were extracted, 100 ps apart. From these conformations, the protein and the closest 166 water molecules were selected. Another ten of such structures were extracted from a simulation of 10 ps, each 1 ps apart. Each structure was energy-minimized without constraints with a conjugate gradients algorithm until the maximum force was below 0.01 kJ mol^{-1} nm^{-1}. For each minimized structure a Hessian matrix was built and diagonalized, and alpha carbon-only eigenvectors were constructed as described in Appendix A. Subsequently, ensembles of alpha carbon structures were generated for each energy minimum from normally distributed displacements corresponding to the local harmonic behavior of the system at a temperature of 300 K. These ensembles of structures were fitted onto a common reference configuration, also used to fit the MD trajectories, in order to remove translational/rotational motions and obtain completely comparable evaluations of the eigenvectors/eigenvalues of different minima and of the simulations.

IF1

For IF1 protein (Initiation Factor 1) the first structure of the 1ah9 protein databank entry was selected as starting point. The protein was solvated in a pre-equilibrated box of SPC water and four water molecules with highest electrostatic potential were replaced by chloride ions, resulting in an electrically neutral cubic box (a = 4.962 nm) containing 3,468 water molecules and four counter ions for a total of 11,119 atoms. After 200 steps of energy minimization using a steepest descent algorithm, a constant volume molecular dynamics run was initiated using a non-bonded cutoff of 0.9 nm for both Lennard-Jones and Coulomb potentials. The pair lists were updated every ten steps. A constant temperature of 300 K was maintained by cou-

pling to an external bath using a coupling constant ($\tau = 0.002$ ps) equal to the integration time step. In this manner, 1.350 ns of simulation were produced, of which only the last 1.0 ns of trajectory was used for essential dynamics analysis. Structures close to one another in the 3-dimensional space defined by the first three principal eigenvectors were chosen as starting points for diffusional studies. From each of these configurations an additional 30 ps of simulation was performed using either a weak temperature coupling ($\tau = 0.1$ ps) or tight temperature coupling ($\tau = 0.002$ ps) to an external bath set at 300 K. After establishing that the type of temperature coupling had no measurable influence on the diffusional properties, a total of 101 pieces (48 produced with weak coupling and 53 with tight coupling) were combined, for a total of 3.03 ns. Each of the 101 runs used a different set of initial velocities, and for each configuration only the last 20 ps of trajectory were used to derive diffusional properties: i.e. the average mean square displacement along the first three principal eigenvectors.

To characterize the local harmonic behavior of the system, a set of configurations distributed in different regions of the 3-dimensional space defined by the first three eigenvectors was chosen. Each of the selected protein configurations was extracted from the trajectory along with a shell of 300 water molecules and minimized using a combination of steepest descent and conjugate gradients algorithms until the maximum force was less than 0.01 kJ mol^{-1} nm^{-1}.

RESULTS AND DISCUSSION

Local Harmonic Wells and MD Fluctuations

As a way to compare two eigenvector sets, A and B, obtained from two different covariance matrices, one can construct the inner product matrix, where each element corresponds to the inner product between the i^{th} eigenvector of set A and the j^{th} eigenvector of set B. A preliminary comparison between the sets corresponding to the two halves of the 1.2 ns trajectory of the B1 domain of protein G was carried out (Fig. 1a). As previously observed,[29] and also found here, no large inner products were found far from the diagonal, in particular the essential subspace (roughly the first 10 eigenvectors) and the far near constraints (eigenvectors beyond the 50*th* or the 60*th*) were completely orthogonal (zero inner product). Differences between the two sets were confined to subspaces defined by contiguous eigenvectors. This result shows that a few hundred picoseconds are enough to reach an approximate convergence for the subspaces definition although the individual eigenvectors require longer time to be properly defined. In what follows, the eigenvectors extracted from this 1.2 ns trajectory were used as an approximation of the true conformational space of the protein and as reference set of vectors in other comparisons.

Comparison of this set of eigenvectors with that obtained from the Hessian matrix of the initial configuration of the trajectory, the X-ray structure, showed that there was a much larger spread in the overlap between these two sets (Fig. 1b) than between the two halves of the trajectory. However, a significant overlap still remains concentrated around the diagonal, especially in the essential subspace. Similar results were found for the other energy minima implying that the dynamic behavior of a protein cannot be properly described by any individual local harmonic approximation, in agreement with previous observations.[1,2] To test whether the "dynamical" essential subspace could be approximated by an average over essential subspaces of various harmonic wells, we combined the C-alpha fluctuations from twenty harmonic wells. Comparison of the resulting eigenvectors with the reference set (Fig. 1c) showed a significant improvement in the overlap between the sets, which was now similar to that obtained in the control (Fig. 1a). Note that this combination of fluctuations (each with respect to the local average) excluded the difference between (local) averages.

To investigate whether this improvement in overlap depends on the distribution of the minima in configurational space, the same kind of comparison was performed between the 10 minima that are close in configurational space (starting conformations extracted from an MD simulation, at intervals only 1 ps apart) and the 10 others that are much further apart (100 ps intervals) with the reference set from MD. The corresponding inner product matrices (Fig. 1d and 1e) show that the two 10 minima sets have an overlap with the reference MD set which is for the essential subspace (first 10 or 20 eigenvectors) comparable to that of the combined 20 minima (Fig. 1c). This indicates that averaging over a limited set of minima increases the overlap with MD significantly, and that this increased overlap occurs for the more closely related minima and for the more distant minima. Finally, we compared (Table I) the different essential subspaces using as a measure of overlap between sets the average square norm u^2:

$$u^2 = \frac{\sum_{i=1}^{10}\sum_{j=1}^{10}(\eta_{Ai}\cdot\eta_{Bj})^2}{10}$$

with $\boldsymbol{\eta}_{Ai}$ and $\boldsymbol{\eta}_{Bj}$ the i and j eigenvectors of set A and B, respectively. Table I clearly shows that the different essential subspaces are rather similar except for the case of the essential subspace obtained for a single well, in which case the overlap is significantly lower. A similar analysis on IF1 protein showed, in the same way, that the combination of the essential subspaces of a limited set of energy minima largely reproduced the essential subspace obtained from the MD trajectory (data not shown). Hence, with respect to MD simulations with time lengths in the order of one nanosecond, a comparable convergence of the dynamical essential modes may be obtained from a combination of a limited number of local minima, even if they are not widely spread in configurational space.

Kinetics in the Essential Subspace

To investigate the dynamics of the configurational fluctuations obtained by MD simulations, we studied the kinetics of the essential coordinates. For both the B1 domain of protein G and IF1 protein we used a large set of very close configurations in the subspace defined by the first 3 eigenvectors. These configurations were then used

TABLE I. Cumulative Mean Square Inner Products Between the Ten Eigenvectors With Largest Eigenvalues Extracted From Different Sets of Structures[†]

Eigenvector-set	Mean cumulative square inner product						
	NM_1	NM_close	NM_far	NM_all	MD_firsthalf	MD_secondhalf	MD_all
NM_1	1.0	0.69	0.72*	0.73*	0.50	0.48	0.50
NM_close		1.0	0.79	0.94*	0.65	0.59	0.65
NM_far			1.0	0.93*	0.62	0.58	0.62
NM_all				1.0	0.66	0.61	0.66
MD_firsthalf					1.0	0.65	0.90*
MD_secondhalf						1.0	0.79*
MD_all							1.0

[†](NM_1: Hessian of x-ray structure; NM_close: collection of the 10 Hessian calculations of closely related structures; NM_far: collection of the 10 Hessian calculations of structures spread over configurational space; NM_all: collection of all 20 Hessian calculations; MD_firsthalf: eigenvectors extracted from the first half of the MD simulation of 1.2 ns; MD_secondhalf: eigenvectors extracted from the second half of the MD simulation of 1.2 ns. MD_all: eigenvectors extracted from the complete MD simulation of 1.2 ns.)
*These values are overestimated because the structures over which the pairs of eigenvectors were calculated partially overlap.

to start new short MD simulations from which we calculated the average square displacement from the initial point as a function of time for the first three essential coordinates. We were interested in characterizing the kinetic behavior of the system after the first fast relaxation when it really enters into the "diffusion regime." For these coordinates, this fast relaxation should be completed within 30–40 fs, as evaluated from the first fast decay of the velocity autocorrelation function to nearly zero (data not shown). The average square displacement was calculated using configurations sampled every 100 fs as shown in Figure 2a and 2b. Such a sampling frequency guarantees that in our analysis we skip completely any kinetic effect due to the initial shape and fast relaxation of the velocity autocorrelation function (e.g. the free flight effect). Hence we focused only on the kinetics behavior due to the tail of the velocity autocorrelation function. In order to increase the statistical sample we averaged the mean square displacements along the first three eigenvectors, assuming a similar diffusional behavior for these essential coordinates. The curves were fitted by a non-linear least-squares fitting procedure to the following functional form (see Appendix B):

$$\langle \Delta_q^2(t) \rangle - \langle \Delta_q^2(t_0) \rangle = 2D_\infty(t - t_0) + A_0\tau_c(1 - e^{-(t-t_0)/\tau_c}). \quad (1)$$

Here t_0 is the time at which the first displacement was calculated (100 fs), D_∞ is the long-time diffusion constant, τ_c the "relaxation time" and A_0 the amplitude of the short-time behavior. For a derivation of this formula see Appendix B.

As is clear from the figures, the theoretical model function reproduces the data obtained by the simulations in the time range investigated. For both proteins, the distribution of residuals was completely compatible with a Gaussian distribution centered on zero. For IF1 protein, the long-time diffusion coefficient for the first three principal eigenvectors was $D_\infty = 3.7 \times 10^{-4}$ nm^2 ps^{-1} and the relaxation time $\tau_c = 2.2$ ps. In the case of the B1 domain of protein G we found a long time diffusion coefficient $D_\infty = 1.9 \times 10^{-4}$ nm^2 ps^{-1} and a relaxation time $\tau_c = 2.8$ ps for the first three eigenvectors. The short time diffusion coefficient ($D_\infty + \frac{1}{2} A_0$), given by the slope at short time ($t < \tau_c$) was comparable for both proteins.

This double diffusion behavior with the long-time diffusion constant smaller than the short-time diffusion constant, is due to the slow decay (convergence to zero) of a negative tail in the velocity autocorrelation function (see Appendix B). Interestingly, a slowly decaying negative tail of the velocity autocorrelation function can also be observed for diffusion in dense liquids (viscoelastic effect), where in general such a fact is given by a slow structural rearranging of the liquid near the moving particle.[30–32] A similar basic physical process can be hypothesized for the essential coordinates diffusion in the medium of the other coordinates although in the dense liquid diffusion the tail decay of the velocity autocorrelation function is usually modeled with a power law which in general does not provide a double mode diffusion.[30–32] A possible simple physical model for the diffusion behavior of the essential coordinates can be the following: the double diffusion corresponds to a short-time diffusion within a configurational region that can be approximated by a single harmonic well, followed by diffusion between such regions, with a long-time diffusion constant D_∞. Transitions between local potential energy minima can be expected to dominate the large-scale dynamics of proteins, as previously observed for simple Lennard-Jones fluids.[15] Hence the negative tail in the velocity autocorrelation function may be interpreted as an increase, with respect to the single harmonic well, of the friction experienced by the essential coordinates in the "bath" of all the other coordinates, resulting in a time decay of the diffusion integral. Although the system is energetically excited and hence able to move between different harmonic well regions without really feeling free energy barriers, the motion from one minimum region to the other requires a longer relaxation time due to the equilibration of other degrees of freedom not contained in the subspace defined by the three essential eigenvectors.

Figure 2a also shows the average square displacement of eigenvectors triplets 9–11, 19–21, 49–51 and 99–101 for the B1 domain of protein G. In this case the initial positions were chosen in the vicinity of each eigenvector's average position, as for these "near-constraints" the free energy is roughly a quadratic function of the displacement, and hence the diffusion behavior is observed only close to the average position. The model curve (Eq. 1) was also

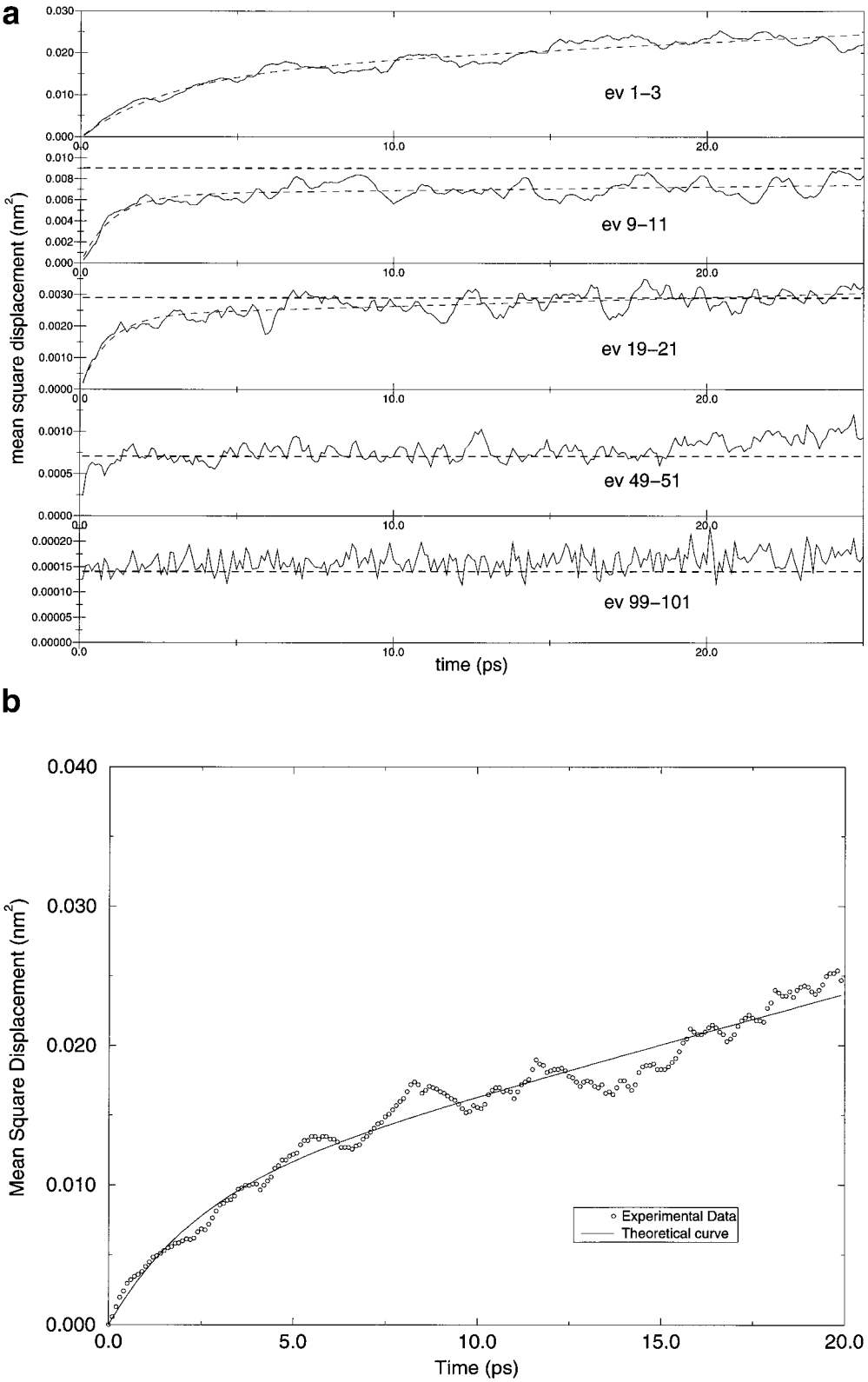

Fig. 2. Mean square displacement as a function of time averaged over the three principal eigenvectors. **a:** B1 domain of protein G. For eigenvectors 1–3 and 9–11 the theoretical curve derived in Appendix B was fitted to the data (dashed lines). The eigenvalues of the corresponding eigenvectors is drawn as a fat dashed line in the lower four panels (for each triplet the eigenvalue of the middle eigenvector was taken). **b:** Diffusion along the first three principal axes of Initiation Factor IF1.

fitted to the mean square displacement for eigenvectors 9–11, where we found a long-time diffusion constant $D_\infty = 1.9 \times 10^{-5}$ nm^2 ps^{-1} and $\tau_c = 0.9$ ps. After 25 ps these coordinates already approach their equilibrium value indicating that from this time on the eigenvector coordinates start to feel a relevant free energy gradient. The figures also show the average eigenvalues of each triplet which define the limit to which the average square displacement will converge. The time required for each triplet to reach the corresponding average eigenvalue, a rough measure of the equilibration time, was estimated, obtaining for the 9–11 eigenvectors $\Delta t \approx 50$ ps and for 19–21 $\Delta t \approx 15$ ps. From the 50^{th} eigenvector on, the full convergence to the eigenvalue (full equilibration) was reached within 1 ps. From these data it is evident that only the first 20 eigenvectors are responsible for the structural fluctuations which involve a time scale larger than 20 ps, and the kinetics of these coordinates can be largely described by the presented double diffusion model.

CONCLUSION

In the first part of this paper we compared the structural fluctuations of the C-alpha coordinates of single harmonic wells, for a protein surrounded by a water layer, with the ones obtained by MD of the same protein solvated in a box of water. The dynamical behavior of the protein, obtained by MD, cannot be described properly by a local harmonic approximation, although each local harmonic essential subspace has a certain degree of similarity with that derived from MD. On the contrary, we have found that the internal dynamics of a protein can be well approximated by a combination of the local dynamics of a limited set of harmonic wells. From the results in the second part (Fig. 2), it seems that it is possible to accurately describe the kinetics in the essential subspace using a double diffusion model, with a characteristic relaxation time τ_c of the faster component of about 1–3 ps. This fact, taken together with the conclusions drawn from the comparison of the structural fluctuations during MD and within local harmonic wells, suggests the following possible physical model: the first diffusion regime, up to a few picoseconds, probably corresponds to the diffusion of the essential coordinates in a single harmonic well, and is characterized by a higher diffusion constant due to the fact that the "bath" degrees of freedom do not have to equilibrate into a new local condition. The second diffusion mode is probably connected with the motions from one well to the other, with then a lower diffusion constant, caused by an increase in friction, resulting from the equilibration of the system when the essential coordinates move beyond the initial well. Finally from the study of the near-constraints kinetics it was found that only the first 20 eigenvector coordinates, responsible for the largest structural fluctuations, are associated with slow structural transitions, with a kinetics which can be largely described by the presented diffusion model.

REFERENCES

1. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68:2696–2699.
2. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins 1993;17:412–425.
3. Levy RM, Srinivasan AR, Olson WK, McCammon JA. Quasi-harmonic method for studying very low frequency modes in proteins. Biopolymers 1984;23:1099–1112.
4. Kitao A, Hirata F, Gō N. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. J Chem Phys 1991;158:447–472.
5. Hayward S, Kitao A, Hirata F, Gō N. Effect of solvent on collective motions in globular proteins. J Mol Biol 1993;234:1207–1217.
6. Hayward S, Gō N. Collective variable description of native protein dynamics. Annu Rev Phys Chem 1995;46:223–250.
7. Van Aalten DMF, Amadei A, Vriend G, Linssen ABM, Venema G, Berendsen HJC, Eijsink VGH. The essential dynamics of thermolysin—confirmation of hinge-bending motion and comparison of simulations in vacuum and water. Proteins 1995;22:45–54.
8. Van Aalten DMF, Findlay JBC, Amadei A, Berendsen HJC. Essential dynamics of the cellular retinol binding protein—evidence for ligand induced conformational changes. Protein Eng 1995;8:1129–1136.
9. Van Aalten DMF, Jones PC, De Sousa M, Findlay JBC. Engineering protein mechanics: inhibition of concerted motions of the cellular retinol binding protein by site-directed mutagenesis. Protein Eng 1997;10:31–38.
10. 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.
11. De Groot BL, Amadei A, Van Aalten DMF, Berendsen HJC. Towards an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. J Biomol Struct Dyn 1996;13:741–751.
12. 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.
13. Sette M, Van Tilborg P, Spurio RR, Kaptein MP, Gualerzi CO, Boelens R. The structure of the translational initiation factor IF1 from *E. coli* contains an oligomer-binding motif. EMBO J 1997;16:1436–1443.
14. Gallagher T, Alexander P, Bryan P, Gilliland GL. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 1994;33:4721–4729.
15. Stillinger FH, Weber TA. Dynamics of structural transitions in liquids. Phys Rev A 1983;28:2408–2416.
16. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics—trypsin-inhibitor, crambin, ribonuclease, and lysozyme. J Mol Biol 1985;181:423–447.
17. Gō N, Noguti T, Nishikawa T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 1983;80:3696–3700.
18. Brooks BR, Karplus M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 1983;80:6571–6575.
19. Janezic D, Brooks BR. Harmonic analysis of large systems. II. Comparison of different protein models. J Comp Chem 1995;16:1543–1553.
20. Janezic D, Venable M, Brooks BR. Harmonic analysis of large systems. III. Comparison with molecular dynamics. J Comp Chem 1995;16:1707–1713.
21. Kitao A, Hayward S, Gō N. Energy landscape of a native protein: jumping-among-minima model. Proteins 1998;33:496–517.
22. Van der Spoel D, Berendsen HJC, Van Buuren AR, Apol E, Meulenhoff PJ, Sijbers ALTM, Van Drunen R. Gromacs User Manual. Nijenborgh 4, 9747 AG Groningen, The Netherlands. Internet: http://rugmd0.chem.rug.nl/~gmx 1995.
23. Van Buuren AR, Marrink SJ, Berendsen HJC. A molecular dynamics study of the decane/water interface. J Phys Chem 1993;97:9206–9212.
24. Van Gunsteren WF, Berendsen HJC. Gromos manual. BIOMOS, Biomolecular Software, Laboratory of Biophysical Chemistry, The Netherlands: University of Groningen; 1987.
25. Van Gunsteren W, Billeter S, Eising A, Hünenberger P, Krüger P, Mark A, Scott W, Tironi I. Biomolecular simulation: the GRO-

MOS96 manual and user guide. Biomos b.v., Hochschulverlag ETH, Zürich; 1996. 903 p.
26. 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.
27. Berendsen HJC, Postma JPM, Van Gunsteren WF, Hermans J. Interaction models for water in relation to protein hydration. In: Pullman B, editor. Intermolecular forces. Dordrecht: D. Reidel Publishing Company; 1981. p 331–342.
28. Berendsen HJC, Postma JPM, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690.
29. 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.
30. Wainwright T, Alder BJ, Gass DM. Decay of time correlations in two dimensions. Phys Rev A 1971;4:233–237.
31. Holian B, Evans D. Shear viscosities away from the melting line: a comparison between equilibrium and non-equilibrium molecular dynamics. J Chem Phys 1983;78:5147–5150.
32. Erpenbeck J. Shear viscosity of the hard-sphere fluid via non-equilibrium molecular dynamics. Phys Rev Lett 1984;52:1333–1335.

APPENDIX A

In this appendix we show how it is possible to obtain the Cα fluctuations for one harmonic well, and hence to derive the eigenvectors and eigenvalues of the Cα covariance matrix, in such a way that these can be directly compared with the fluctuations obtained by MD. For this purpose we show that the probability density of the internal Cα fluctuations is a multivariate Gaussian distribution, and derive the eigenvectors and eigenvalues of that distribution.

The potential energy for a single harmonic well is given by a quadratic function of the coordinates with respect to the local potential minimum:

$$V(\Delta \boldsymbol{x}) = V_0 + \frac{1}{2}\Delta \boldsymbol{x}^{\mathsf{T}}\mathbf{H}\Delta \boldsymbol{x} \tag{2}$$

with

$$\Delta \boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}_{min}. \tag{3}$$

Here $\boldsymbol{x}$ is the full Cartesian coordinates column vector and $\mathbf{H}$ the Hessian matrix. Since the Hessian is by definition a symmetric matrix, its eigenvectors can always be chosen as an orthonormal set. Six eigenvalues are zero, corresponding to overall translation and rotation. Hence we can define (for an unconstrained molecular cluster) $3N - 6$ orthonormal coordinates ξ_i such that

$$V(\boldsymbol{\xi}) = V_0 + \frac{1}{2}\boldsymbol{\xi}^{\mathsf{T}}\mathbf{K}\boldsymbol{\xi} \tag{4}$$

$$= V_0 + \frac{1}{2}\sum_{i=1}^{3N-6} k_i\xi_i^2. \tag{5}$$

Here

$$\mathbf{K} = \mathbf{B}^{\mathsf{T}}\mathbf{H}\mathbf{B} \tag{6}$$

is the diagonal eigenvalue matrix with diagonal elements k_i, with $\mathbf{B}$ the linear orthogonal transformation which diagonalizes $\mathbf{H}$, and $\boldsymbol{\xi}$ the coordinates defined by the eigenvectors of $\mathbf{H}$:

$$\Delta \boldsymbol{x} = \mathbf{B}\boldsymbol{\xi}. \tag{7}$$

Since the Jacobian of an orthogonal transformation is equal to 1, the canonical distribution function of ξ_i, $i = 1, \cdots, 3N - 6$, is given by the Boltzmann distribution

$$\rho(\boldsymbol{\xi}) = \prod_{i=1}^{3N-6}\left(\frac{\beta k_i}{2\pi}\right)^{1/2} e^{-1/2\beta k_i \xi_i^2}, \tag{8}$$

where $\beta = 1/k_B T$. This means that the probability density of the fluctuations along each nonzero eigenvalue eigenvector is an independent Gaussian distribution with variance $\langle \xi_i^2 \rangle = 1/\beta k_i$ and $\langle \xi_i \xi_j \rangle = 0$ for $i \neq j$. The distribution function of ξ_i, $i = 3N - 5, \cdots, 3N$, corresponding to the null eigenvectors, is irrelevant, and these coordinates can be arbitrarily set to zero.

We note that in the presence of n_c internal constraints Eq. 8 (but now for $i = 1, \ldots, 3N - 6 - n_c$) still gives the correct probability density when the usual atomic coordinates are replaced by generalized coordinates which are defined by a local orthogonal set of axes on the constraint surface. The computation of the Hessian in the presence of constraints requires the calculation of the potential energy second derivatives in the generalized coordinates defined on the constraint surface. To avoid such complications we have represented covalent bonds by harmonic terms rather than by constraints.

The fluctuations of the $\boldsymbol{x}$ coordinates without any translation and rotation can be obtained from

$$\boldsymbol{x} = \boldsymbol{x}_{min} + \Delta \boldsymbol{x} \tag{9}$$

$$\Delta \boldsymbol{x} = \mathbf{B}\boldsymbol{\xi}, \tag{10}$$

where $\boldsymbol{\xi}$ now has the 6 translational/rotational ξ_i, $i = 3N - 5, \ldots, 3N$, exactly set to zero. The $\boldsymbol{x}$ covariance matrix for the system with no translation and rotation is

$$\mathbf{C} = \langle \Delta \boldsymbol{x} \Delta \boldsymbol{x}^{\mathsf{T}} \rangle$$

$$= \langle \mathbf{B}\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}} \rangle = \mathbf{B}\langle \boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}} \rangle \mathbf{B}^{\mathsf{T}} = \mathbf{B}\begin{pmatrix} \mathbf{K'}^{-1}/\beta & 0 \\ 0^{\mathsf{T}} & \mathbf{C}^0 \end{pmatrix}\mathbf{B}^{\mathsf{T}}, \tag{11}$$

or, for any element,

$$C_{ij} = k_B T \sum_{l=1}^{3N-6} k_l^{-1} B_{il} B_{jl}. \tag{12}$$

Here $\mathbf{K'}$ is identical to the $(3N - 6) \times (3N - 6)$ sub-matrix of $\mathbf{K}$ corresponding to the non zero-eigenvalues, $\mathbf{C}^0$ is a

6×6 null matrix and $\mathbf{0}$ is a $(3N - 6) \times 6$ null matrix. Equation 11 or 12 means that from the full-coordinate Hessian $\mathbf{H}$ we can evaluate the covariance matrix for (any subset of) the atomic coordinates due only to the internal motions in the molecule.

Since, according to Eq. 10, each Δx_i can be constructed as linear combination of ξ_j's which have independent Gaussian distributions, the distribution of Δx_i will be a convolution of Gaussians and hence will be Gaussian itself. Thus the equilibrium distribution of any subset coordinate vector $\boldsymbol{\Delta x}$ must be a multivariate Gaussian, completely determined by its covariance matrix $\mathbf{C}$. This property allows us to generate an equilibrium ensemble for the subset from knowledge of its covariance matrix only. The procedure is simple: first diagonalize the covariance matrix:

$$\boldsymbol{\Delta x} = \mathbf{T}\boldsymbol{q} \tag{13}$$

$$\mathbf{C} = \mathbf{T}\Lambda\mathbf{T}^{\mathrm{T}}; \quad \Lambda_{ij} = \lambda_i \delta_{ij}, \tag{14}$$

then sample each q_i from a normal distribution with variance λ_i, and finally construct $\boldsymbol{\Delta x}$ from Eq. 13. The structures thus generated for the C_α subset were fitted onto a common reference structure also used for the configurations obtained by MD. In this way we could directly compare local harmonic C_α fluctuations in different minima to each other and to the fluctuations observed in MD.

We note that the covariance matrix also determines the free energy Hessian $\mathbf{H}'$ in the subset of coordinates, normally obtained by integration over the equilibrium distribution of all other coordinates. If overall translation and rotation have been removed from the coordinate set, $\mathbf{H}' = k_B T \mathbf{C}^{-1}$.

APPENDIX B

In this section we describe the theory used in this paper to model the kinetics in the essential subspace.

For a coordinate q we can express the average square displacement from an initial point, as a function of time, as:

$$\langle \Delta q^2(t) \rangle = \left\langle \int_0^t dt' \int_0^t \dot{q}(t')\dot{q}(t'')\, dt'' \right\rangle = \int_0^t dt' \int_0^t \langle \dot{q}(t')\dot{q}(t'') \rangle\, dt'' \tag{15}$$

where $\dot{q}$ is the time derivative of q, $\Delta q(t) = q(t) - q(0)$ and the angle brackets represent the ensemble average. Using standard derivations we can rewrite the previous equation as:

$$\langle \Delta q^2(t) \rangle = 2 \int_0^t I(\tau)\, d\tau, \tag{16}$$

where

$$I(\tau) = \int_0^\tau \gamma(\tau')\, d\tau'. \tag{17}$$

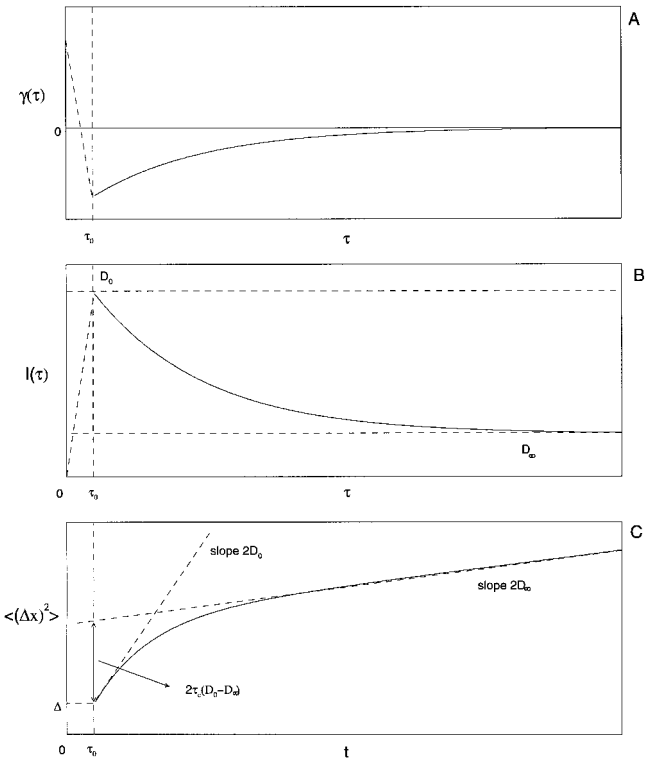

Fig. A1. Time behavior of diffusional functions. The details of the functions below the (short) time τ_0 are irrelevant and indicated by a dashed line. **1A:** Velocity autocorrelation function, with long negative tail. **1B:** The function $I(\tau)$ (see text). **1C:** The mean square displacement.

and $\gamma(\tau') = \langle \dot{q}(0)\dot{q}(\tau') \rangle$ is the velocity autocorrelation function of q. In a typical simple diffusion model the function $I(\tau)$ is a rapidly increasing function converging to a positive final value, reached approximately within a time interval τ_0:

$$\lim_{\tau \to \infty} I(\tau) = D \tag{18}$$

$$I(\tau) \cong D \quad \tau \geq \tau_0. \tag{19}$$

In this paper we use a more complex approach where we still consider the function $I(\tau)$ rapidly converging to a positive value within τ_0 (corresponding to a fast first relaxation), but we model the function $I(\tau)$ differently for $\tau > \tau_0$. In fact, we consider that the system undergoes a second, slower relaxation affecting $I(\tau)$ via a simple first order kinetics with time constant τ_c, leading to

$$I(\tau) = (D_0 - D_\infty) e^{-(\tau - \tau_0)/\tau_c} + D_\infty, \tag{20}$$

valid clearly only for $\tau \geq \tau_0$. In Figure A1 both the velocity autocorrelation function $\gamma(\tau')$ and its integral $I(\tau)$ have been sketched for this model. It is clear that for $D_0 > D_\infty$ the additional first order process corresponds to a slow negative tail in the velocity correlation function.

Inserting Eq. 20 into Eq. 16 we obtain

$$\langle \Delta q^2(t) \rangle = 2 \left(\int_0^{\tau_0} I(\tau)\, d\tau + \int_{\tau_0}^{t} I(\tau)\, d\tau \right) \tag{21}$$

$$= 2\Delta + 2D_\infty (t - \tau_0) + 2\tau_c (D_0 - D_\infty)[1 - e^{-(t-\tau_0)/\tau_c}] \tag{22}$$

with

$$\Delta = \int_0^{\tau_0} I(\tau)\, d\tau. \tag{23}$$

(see Fig. 1a). The slope is given by

$$\frac{d\langle \Delta q^2(t) \rangle}{dt} = 2D_\infty + 2(D_0 - D_\infty) e^{-(t-\tau_0)/\tau_c}, \tag{24}$$

which corresponds to the usual Einstein-Smolukowsky expression, when we distinguish two time regimes depending on the observation time in relation to τ_c: $t \ll \tau_c$: slope $2D_0$, and $t \gg \tau_c$: slope $2D_\infty$.

If the displacement for a small time $t_0 > \tau_0$ is subtracted in Eq. 22, the following behavior is found:

$$\langle \Delta q^2(t) \rangle - \langle \Delta q^2(t_0) \rangle = 2D_\infty (t - t_0) + A_0 \tau_c (1 - e^{-(t-t_0)/\tau_c}), \tag{25}$$

with

$$A_0 = 2(D_0 - D_\infty) e^{-(t_0 - \tau_0)/\tau_c}. \tag{26}$$

Equation 25 was used to evaluate the time behavior of $\langle \Delta q^2(t) \rangle$ in the time range: 100 fs–25 ps, for the essential coordinates and some of the first near constraints.

We note that the kinetics of any internal coordinate, including the essential ones, cannot be described by this model to full equilibration, because in this time limit the coordinates have sampled the whole available space and reached the free energy barriers which define the corresponding boundaries. This model, if appropriate, can be used in a time range where the coordinates do not encounter a relevant free energy gradient. If free energy boundaries are met, the mean square displacement will gradually level off for longer times and reach the limit given by the eigenvalue of the degree(s) of freedom concerned.