The Effect of Box Shape on the Dynamic Properties of Proteins Simulated under Periodic Boundary Conditions

TSJERK A. WASSENAAR,[1] ALAN E. MARK[1,2]

[1]*Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands*

[2]*School of Molecular and Microbiological Sciences, University of Queensland, Brisbane, QLD 4072, Australia*

Received 1 May 2005; Accepted 10 June 2004
DOI 10.1002/jcc.20341
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: The effect of the box shape on the dynamic behavior of proteins simulated under periodic boundary conditions is evaluated. In particular, the influence of simulation boxes defined by the near-densest lattice packing (NDLP) in conjunction with rotational constraints is compared to that of standard box types without these constraints. Three different proteins of varying size, shape, and secondary structure content were examined in the study. The statistical significance of differences in RMSD, radius of gyration, solvent-accessible surface, number of hydrogen bonds, and secondary structure content between proteins, box types, and the application or not of rotational constraints has been assessed. Furthermore, the differences in the collective modes for each protein between different boxes and the application or not of rotational constraints have been examined. In total 105 simulations were performed, and the results compared using a three-way multivariate analysis of variance (MANOVA) for properties derived from the trajectories and a three-way univariate analysis of variance (ANOVA) for collective modes. It is shown that application of roto-translational constraints does not have a statistically significant effect on the results obtained from the different simulations. However, the choice of simulation box was found to have a small (5–10%), but statistically significant effect on the behavior of two of the three proteins included in the study.

© 2005 Wiley Periodicals, Inc.   J Comput Chem 27: 316–325, 2006

Keywords: periodic boundary conditions; near-densest lattice packing; molecular dynamics; simulation; MANOVA

Introduction

When molecular dynamics (MD) simulations of biological macromolecules such as proteins or DNA are performed in explicit solvent, periodic boundary conditions are commonly used to minimize edge and finite size effects. In such simulations the computational cost is normally dominated by the calculation of the (less interesting) solvent degrees of freedom. For this reason it is highly desirable to minimize the amount of solvent by optimizing the size and shape of the box for different solutes. For example, Bekker et al.[1] recently proposed a method to determine the periodic simulation box that yielded the optimal packing for a solute molecule with a specific geometry and a given minimal distance between the periodic images. This method is based on finding the near-densest lattice packing (NDLP) of the point set defined by the atom positions, dilated with a radius equal to the minimal distance to the wall. It was shown that for a set of proteins selected randomly from the protein databank[2] the NDLP box was, on average, 50% smaller than a rhombic dodecahedron box, the closest possible packing for a spherical molecule.

It should be noted that the NDLP box is not a box type in the general sense. Rather than being an Archimedean structure with a perfect packing, defining an infinite lattice, the NDLP box is inferred from the best possible packing of a given solute. The easiest definition is the triclinic box directly following from the lattice, bearing in mind that it is always possible to transform a simulation performed in any given box type, including the NDLP or molecular shaped box, into any of the other, more general, box types.[3]

A different definition of the box is the Voronoi region[4] of the solute in the lattice, which is the region of points closer to the solute than to any of its periodic images. In general, this will define a complex, nonconvex shape, which is referred to as the *molecular shaped box*. It is possible that the NDLP method will lead to a

Correspondence to: T. A. Wassenaar; e-mail: T.A.Wassenaar@rug.nl regular box shape. Spherical solutes, for example, will lead to a box definition closely resembling a rhombic dodecahedron.

Performing simulations of any nonspherical molecule in a minimal periodic box has one obvious consequence. If the box is not chosen such that the minimum box dimension is larger than the largest dimension of the solute plus the cutoff, rotation of the solute will lead to the direct interaction between neighboring periodic images. Thus, when performing any simulation of a nonspherical molecule in a nonspherically symmetric box such as a rectangular box or that generated using the NDLP method the rotational motion of the solute must be constrained, for example, using the method proposed by Amadei et al.[5] There is, however, another consequence of choosing a minimal box size in simulations under periodic conditions that has been largely ignored. The box shape itself will influence the nature of the motions the system can undergo by restricting sampling in certain directions and thus affect the outcome of the simulation. For boxes with a relatively large amount of solvent compared with the size of the solute molecule this effect will be small and can be neglected. However, when one is using a box containing only a minimal amount of solvent the effect may well be significant. No periodic box is truly spherically symmetric. All boxes will, to a greater or lesser degree, affect the outcome of a simulation. Nevertheless, the development of the NDLP box together with the application of methods to constrain rotational motion without affecting the statistical mechanical ensemble provides an opportunity to directly evaluate the effect of box shape and the degree of solvation on the dynamic properties of proteins.

In this study we compare the dynamic behavior of three proteins in each of four different box types (including the NDLP) with and without the application of rotational constraints. The aim of the study was to determine if it were possible to detect statistically significant differences between simulations performed under these different conditions, based on a comparison between a number of structural properties. These properties included the RMSD, the radius of gyration, the total number of intraprotein hydrogen bonds, and the number of hydrogen bonds associated with assigned secondary structure elements, the solvent-accessible surface, and the number of residues involved in α-helix and β-sheet secondary structure elements. Furthermore, an assessment was made of the statistical significance of differences in the directionality of sampling of conformational space between different sets of simulations. Specifically, principal component analysis[6] has been used to determine the collective motions in the simulations. The overlap of the conformational space sampled in the different simulations was estimated from the root-mean-square inner product (RMSIP)[7] of the eigenvectors corresponding to the dominant modes.

The statistical significance of the differences between multiple simulations was determined using analysis of variance (ANOVA), which is a basic statistical method to compare multiple sets of data by comparing the variance observed within the sets with the variance observed between different sets.[8,9] Data resulting from the eigenvector analysis was analyzed using three way fixed effects (type I) ANOVA, whereas descriptive properties obtained from the simulations were analyzed using the multivariate equivalent, MANOVA.[10]

Methods

Simulations

Three proteins of varying shape and secondary structure content were used in this study: Chymotrypsin Inhibitor II (PDB code 2CI2[11]), the GAG polyprotein M-domain of the rous sarcoma virus (PDB code 1A6S[12]), and Lysozyme (PDB code 1AKI[13]). Starting structures were taken from the Protein DataBank.[2] For each protein five simulations were performed in each of four box types (rhombic dodecahedron, truncated octahedron, rectangular, and the appropriate NDLP box) either with or without the application of roto-translational constraints. Note that simulations in a NDLP box were performed exclusively with rotational constraints. This is because the NDLP box corresponds to a specific orientation of the solute in the box. Allowing rotational motion would rapidly and inevitably lead to direct interactions between periodic images. Simulation boxes were chosen such that the minimal distance to the wall was 1.0 nm in all box types, resulting in a minimal distance between periodic images of at least 2.0 nm. The method to find the NDLP has been described previously.[1] Production runs for 1A6S and 2CI2 were 10 ns each. Production runs for 1AKI were 20 ns for simulations performed in a NDLP box with rotational constraints and in a rhombic dodecahedron, a truncated octahedron or a rectangular box without constraints. Simulations of 1AKI performed in the latter three box types with rotational constraints were 5 ns in length. Note that although the longer simulations (20 ns 1AKI and 10 ns 1A6S and 2CI2) were required to allow sufficient convergence of principal components in terms of the analysis of structural properties it was found that 5 ns was sufficient to allow the statistical assessment. As this was the maximum time available for all systems, for consistency, only the analysis based on simulations of 4–5 ns is presented.

All simulations were performed using a modified version of the Gromacs 3.1.4 simulation package,[14–16] in which the roto-translational constraint algorithm of Amadei et al.[4] had been implemented. The interatomic interactions were described using the GROMOS96 43a2 united atom force field.[17,18] Water molecules were modeled explicitly using the Simple Point Charge (SPC) model.[19] The protonation state of ionizable groups was chosen appropriate for pH 7.0. Counterions were added to neutralize the net charge of the system. Nonbonded interactions were evaluated using a twin range cut off of 0.9 and 1.4 nm. Interactions within the shorter range cutoff were evaluated at every step, whereas interactions within the longer range cut off were evaluated every 10 steps. To correct for the neglect of electrostatic interactions beyond the longer range cut off, a Reaction Field (RF) correction[20] was used with $\varepsilon_{\mathrm{RF}} = 78.0$. In all simulations the system was kept at a constant temperature of 300 K by applying a Berendsen thermostat.[21] Protein and solvent molecules were independently coupled to the heat bath with a coupling time of 0.1 ps. Simulations were performed at constant volume. Note, the volume was set such that the pressure was on average approximately 1 atmosphere, ensuring that the water density was the same in all cases. The time step used for the integration of the equations of motion was 0.002 ps. The bond lengths and angle of the water molecules were constrained using the SETTLE algorithm.[22] Bond lengths within the protein were constrained using the SHAKE[23] algorithm. Starting velocities were randomly assigned from a Maxwellian distribution with different random seeds for each of the simulations. Each of the systems was equilibrated for 10 ps before starting production runs.

Analysis

The effect of the different box shapes and volumes on the result of the simulations was assessed by analyzing the similarity or otherwise of a number of properties derived from the trajectories and of the collective fluctuations in the system.

Properties that were included in the analysis were: (1) the root-mean-square deviation (RMSD) of the average structure against the experimentally determined structure (bbRMSD), (2) the RMSD, excluding flexible regions of the protein (i.e., residues not assigned to secondary structure elements; ssRMSD), (3) the average radius of gyration (Rgyr), (4) the number of intramolecular hydrogen bonds (totHbnd), (5) the number of hydrogen bonds associated with secondary structure elements defined in the experimental structure (ssHbnd), (6) the hydrophobic solvent accessible surface (SASphob), (7) the hydrophilic solvent accessible surface (SASphil), (8) the number of residues involved in β-sheet formation (β-Sheet), and (9) the number of residues involved in α-helical elements (α-helix). These properties were determined using routines available in the Gromacs package and are described in the documentation of Gromacs.[16] These properties were determined over a time window of 1 ns, from 4–5 ns for all trajectories.

The set of structural properties considered were expressed as observation vectors for all simulations and were used for a three-way fixed-effects multivariate analysis of variance (MANOVA) with provision for interactions between applied conditions. These conditions, or factors, were the protein, the type of box used for the simulation and the application of roto-translational constraints.

The first step in the eigenvector analysis was the generation of the positional covariance matrices for each of the individual simulations. These were diagonalized to obtain the eigenvectors and corresponding eigenvalues. The 10 eigenvectors with the largest associated eigenvalues were used to compare pairs of simulations. This was achieved by determining the root-mean-square inner product (RMSIP)[7] of the selected sets of eigenvectors from both simulations, which is a measure of the overlap of the subspaces sampled in these simulations. The approach used was to determine all pairwise overlaps. These were separated into two classes: from simulations performed under identical conditions, but with different starting configuration, and from simulations performed under different box conditions.

The assumption was made that RMSIP values for simulations performed under identical conditions were approximately normally distributed with a specific mean μ and variance σ^2. The mean and the variance were estimated from multiple simulations. If a and b denote two sets of simulations performed under different conditions, all pairwise RMSIP values within each set were determined (sets A and B) from which μ_A, μ_B, $\sigma^2{}_A$, and $\sigma^2{}_B$ could be estimated. All pairwise RMSIPs between simulations from set a and set b were then determined (set A × B). This provided the estimates for the mean RMSIP $\mu_{A\times B}$ and the corresponding variance $\sigma^2{}_{A\times B}$ between any two simulations performed under the different conditions. Samples of RMSIPs calculated from pairs of simulations performed under the same treatments will be called type I or *within* treatment RMSIPs, samples obtained from pairs of simulations from two different treatments will be called type II or *between* treatments RMSIPs.

The resulting RMSIP values were analyzed using three-way fixed-effects ANOVA with provision for interactions between conditions. The conditions included in the model were the same as for the MANOVA. Two sets of simulations, a and b, were regarded to be equal (indistinguishable) with respect to the collective motions governing the dynamics if the corresponding RMSIP samples A and B were equal and were both equal to A×B. If two sets of RMSIPs were found to be equal at the 95% confidence interval, the corresponding simulations were considered to give rise to the same essential motions, and any differences in the simulation conditions were regarded as not significantly affecting the sampling of conformational space on the time scales probed.

Statistical Analysis

All statistical analysis was performed using the program R, a language and environment for statistical computing (http://cran.r-project.org/).[24]

ANOVA and MANOVA are based on the general linear model underlying statistical experimental design. The model for two way MANOVA with provision for interaction is:

$$x_{ijkh} = \mu_h + \alpha_{ih} + \tau_{jh} + \eta_{ijh} + \varepsilon_{ijkh}$$

Here, x_{ijkh} is the kth observation of the hth response. μ_h is a general-level effect for the hth response, comparable to the grand mean of all observations. α_{ih} and τ_{jh} are deviations of the hth response from μ_h to effects of external conditions i and j. η_{ijh} is the deviation from that response due to an interaction between external conditions i and j, and ε_{ijkh} is a random variable term of the hth response with mean zero. Two-way ANOVA is the special case of two-way MANOVA with only one variate. The model can be reduced (e.g., one way) or extended (three or multiway) by the exclusion of terms or the inclusion of extra terms.

The purpose of (M)ANOVA is to test whether the effects due to conditions or interactions between conditions exceed the residual variance on a given significance level. In other words, the probability, denoted p, that different samples are obtained from a common underlying distribution is determined. For all analysis the criterion for rejection of the hypothesis of equality was a probability of the hypothesis being true lower than 5% ($p < 0.05$). To assess the equality of several sets of observations the total variance is decomposed into components determined by the external conditions and an unexplained residual variance. The ratio of the condition-determined and residual variance in the univariate case has the variance ratio (F) distribution with the appropriate degrees of freedom. This ratio provides the probability that the condition-determined and residual variances are equal and that the given condition consequently has no effect. For the multivariate case several test statistics are available that are based on the distribution of the nonzero eigenvectors of the ratio of covariance matrices equivalent to the ratio of variances in univariate ANOVA. In this study Wilk's lambda[25] was used.

If the ANOVA and MANOVA results suggested a significant difference between at least two sets of simulations, multiple com-

Table 1. System Sizes and Simulation Speeds.

Protein	Size (atoms)	Box	Volume (nm^3)	SPC	Solvent density ($kg\ m^{-3}$)	Total (atoms)	Speed (h/ns)	Rel. speed
1A6S	805	DH	171.30	5205	970	16,420	14.5	1
		OH	186.49	5713	970	17,944	15.9	0.9
		RC	171.50	5062	940	15,991	13.9	1
		MB	102.02	2975	975	9730	8.4	1.7
2C12	650	DH	183.73	5729	977	17,837	16.1	1
		OH	200.02	6229	971	19,337	17.9	0.9
		RC	141.62	4308	965	13,574	11.6	1.4
		MB	73.75	2042	940	6776	5.7	2.8
1AKI	1321	DH	279.53	8592	974	27,097	27.6	1
		OH	304.32	9425	977	29,596	30.6	0.9
		RC	209.25	6267	970	20,122	18.9	1.5
		MB	115.69	3176	952	10,849	9.9	2.8

Proteins are referred to by their PDB entry codes. The following codes are used for box types: DH, rhombic dodecahedron; OH, truncated octahedron; RC, rectangular box; MB, molecular-shaped (NDLP) box. The solvent density was calculated by dividing the box in slices. The reported density was obtained by averaging over slices containing no protein atoms. The relative speed is calculated taking the rhombic dodecahedron box as reference.

parisons were made to investigate the source(s) of these differences. For the univariate case this was done using Tukey's Honest Significant Differences (HSD) method,[26] multivariate observations were further analyzed using the Roy union-intersection approach.[10,25,27]

Results

Simulations

The system sizes and the relative cost of the simulations for each of the proteins in the four different boxes are given in Table 1. Graphical representations of these are given in Figure 1A, in which the box types are shown as wire frame and the solvent is colored according to the shortest distance to any of the surrounding images of the protein. Clearly, the molecular shaped box is much smaller than the other box types, resulting in an increase in efficiency of 70, 180, and 180% for 1A6S, 2CI2, and 1AKI, respectively, compared to the rhombic dodecahedron.

The average values for the set of descriptive properties obtained from each of the simulations are given in Figure 2A–C. Figure 2A–C shows that each protein has a characteristic profile with respect to the properties included in the analysis. No large deviations due to either the box type or the use of rotational constraints were observed.

The plots in Figure 2D show that the RMSIP values for both type I and type II observations, are evenly distributed around the mean value for each protein. This indicates that the RMSIPs are more or less specific for a given protein and time scale, regardless of the type of box used.

Statistical Analysis

The results of the three-way fixed-effects MANOVA (Table 2) show that the main factors effecting the deviation from the common mean are the protein used in the simulation ($p < 2 \times 10^{-16}$), the box type used for the simulation ($p = 9 \times 10^{-6}$) and the interaction between the protein and the box type ($p = 7 \times 10^{-4}$). The application of rotational constraints did not give rise to statistically significant differences ($p = 0.87$). One-way MANOVA performed on each of the proteins separately, assessing the box effect only, revealed that the p-values are 0.81, 0.10, and 0.008, respectively, for 2CI2, 1AKI, and 1A6S.

It should be noted that diagnostic analysis of three-way MANOVA results (as well as for the three-way ANOVA results of the next section) showed that the samples were heteroscedastic, that is, of unequal variances. Because homoscedasticity is a basic assumption of (M)ANOVA, this can in some cases indicate that the results of the analysis may be invalid. However, the differences in variance were found to be due to the protein. Analysis of each of the proteins separately showed that each of these models itself was valid. Separate analysis did not lead to any change in the overall conclusions made and are therefore not presented.

Results similar to those for the MANOVA were found for the analysis of variance of collective modes from the simulation (Table 3). The main determining factor was again the protein ($p < 2 \times 10^{-16}$). Statistically significant differences were also found for combinations of box types ($p = 2 \times 10^{-8}$) and for interaction terms between the protein and the combination of box types ($p = 9 \times 10^{-10}$). Again, no statistically significant effects were observed related to the application or not of rotational constraints ($p = 0.63$).

Given that there was interaction between the box type and the protein, the next step was to investigate the results of different proteins separately. The application of one way MANOVA on the chosen descriptive properties for each of the proteins yielded p-values for an effect of the box type of 0.008, 0.10, and 0.81 for proteins 1A6S, 1AKI, and 2CI2, respectively. For the RMSIP samples, the respective p-values from an individual one-way ANOVA were 4×10^{-9}, 2×10^{-5} and 6×10^{-2}.

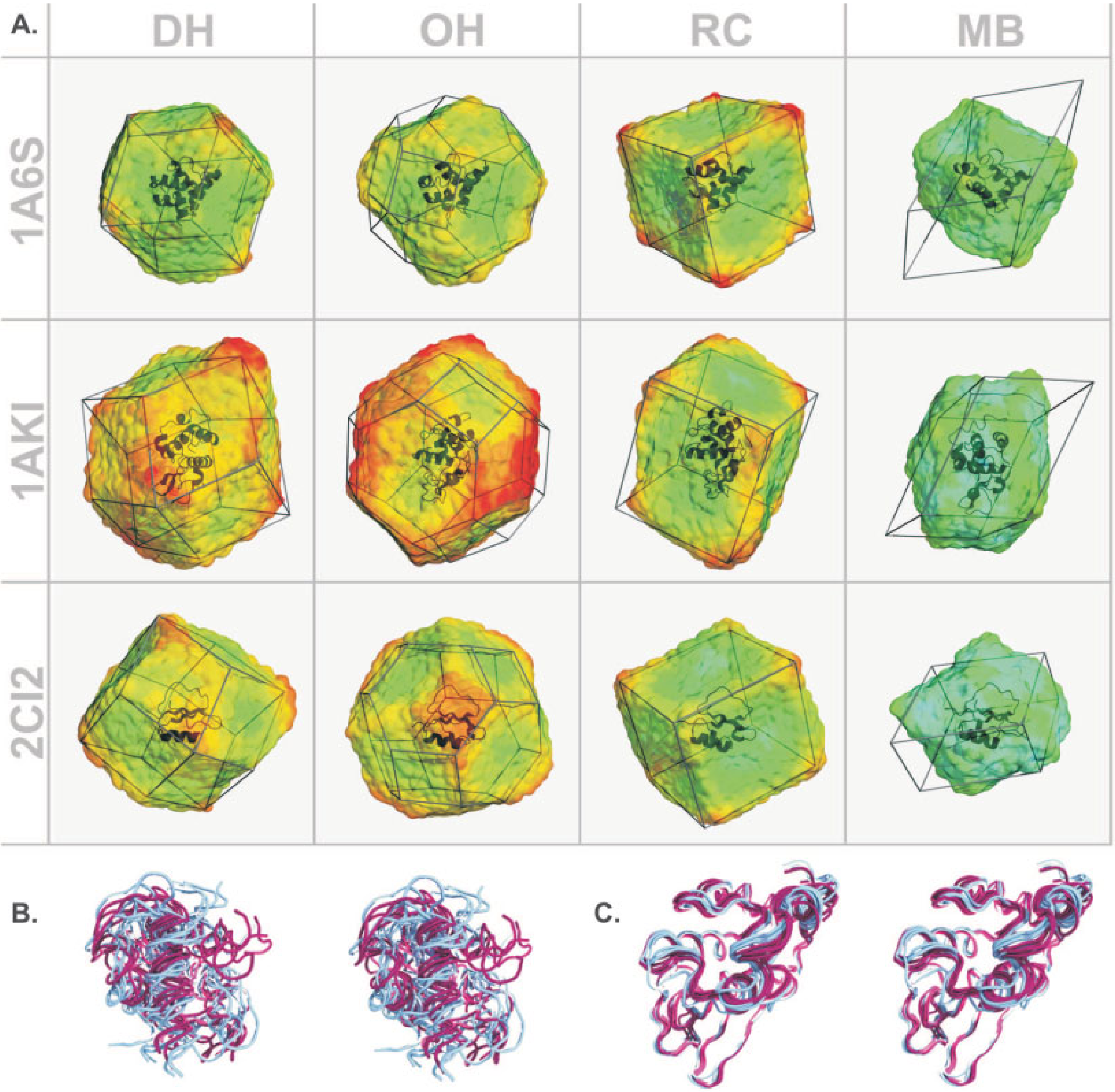

Figure 1. (A) Simulation systems for all proteins and all box types included in the study. Proteins are shown in cartoon representation, simulation boxes as wire frames. Solvent is arranged around the protein according to the shortest distance from the central image in the periodic system, revealing the Voronoi region of that protein in the lattice, defined by the box. Solvent is colored according to the shortest distance to any of the periodic images of the protein. Color range from blue/cyan (<1 nm) to red (>3 nm). (B) Stereoview of superimposed average structures from simulations of protein 1A6S in a rectangular (blue) and a NDLP (purple) box. (C) Stereoview of superimposed average structures from simulations of protein 1AKI in a rectangular (blue) and a truncated octahedron (purple) box. Images were created using PyMOL[28] and POV-Ray.[29]

The results of tests comparing the descriptive properties of box types for 1A6S and 1AKI are given in Tables 4 and 5. Because an effect of the box type on these properties was effectively ruled out for 2CI2, this protein was not included in this analysis. In addition, as only one out of the 35 simulations of 1A6S showed any β-sheet content (0.11 residues averaged over 1 ns), this property was excluded from the analysis of this protein. The results show the most significant differences (p = 0.002, 0.009, and 0.05) were between simulations performed in a rectangular box with respect to the other box types for 1A6S. For 1AKI, the most significant differences were found between simulations performed in a truncated octahedron box compared with the simulations performed in other box types (p = 0.02, 0.03, 0.11).

Multiple comparisons of the RMSIP sets for 1A6S and 1AKI, using Tukey's Honest Significant Differences, revealed that for 1AKI seven pairs of RMSIP samples differ from each other on the 95% level, whereas for 1A6S 13 pairs differ. The results for 1A6S show that the overlap between simulations performed in a rectangular box or between simulations performed in a rectangular and a rhombic dodecahedron box tend to be higher than for simulations from other pairs of box types. For simulations of 1AKI, RMSIP samples from simulations in a truncated octahedron and either a rectangular or a molecular-shaped box are lower than samples involving simulations performed in other box types. These results are in good agreement with the results obtained from the analysis of configurational properties.

Discussion

Although multivariate analysis of variance is a very common technique in biological, social, and economic sciences, it has rarely if ever been used to analyze molecular dynamics trajectories. The main reason for this is that in the past incomplete sampling and the cost of performing multiple simulations of the same system has limited the use of rigorous statistical approaches. With advances in computing resources, however, these problems are slowly being overcome. The present application of (M)ANOVA on a number of properties obtained from simulation trajectories demonstrates the possibility to use such statistical techniques in the evaluation of results from molecular dynamics simulations.

The current results show that the application of roto-translational constraints using the method of Amadei et al.[5] does not give rise to statistically significant differences in either descriptive properties or the main collective modes, as reflected by the RMSIP

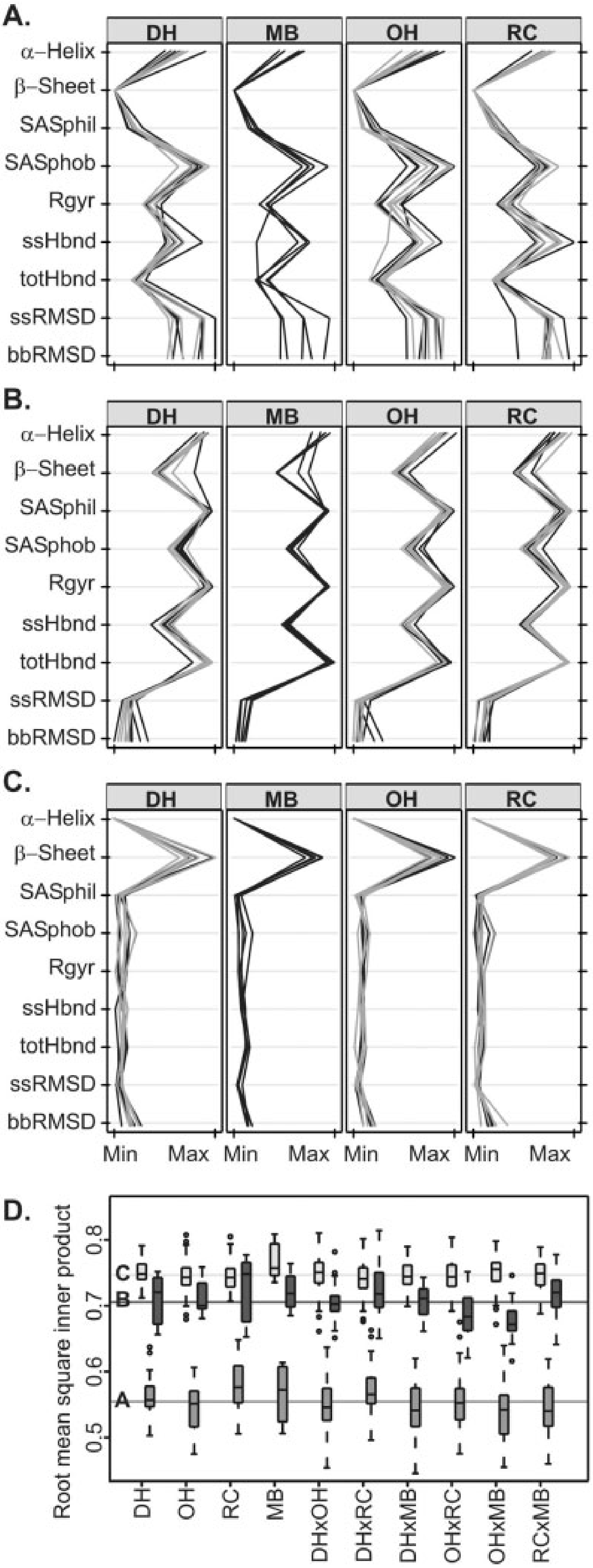

Figure 2. (A–C) Profiles of descriptive properties for proteins 1A6S (A), 1AKI (B), and 2CI2 (C), as time averages from 4–5 ns. Results from simulations performed without the application of roto-translational constraints are given in light gray, those obtained with such constraints are shown in black. *min* and *max* denote the minimum and maximum values for each of the properties measured as determined from the complete data set. α-Helix: number of residues involved in α-helical formation, β-Sheet: number of residues involved in β-sheet formation, SASphil: hydrophilic solvent accessible surface, SASphob: hydrophobic solvent accessible surface, Rgyr: radius of gyration, ssHbnd: number of hydrogen bonds in structures denoted secondary structure elements in the starting structure, totHbnd: total number of intramolecular hydrogen bonds, ssRMSD: root mean square deviation of backbone atoms belonging to secondary structure elements as determined from the starting structure, bbRMSD: root-mean-square deviation of all backbone atoms against the starting structure. (D) Boxplots of overlaps of conformational subspaces sampled as determined by the root-mean-square inner products of pairs of simulations from all combinations of box types for each of the proteins included in the study. Horizontal bars indicate the grand mean for each protein. Sets *A*, *B*, and *C* correspond to the preceding graphs.

Table 2. MANOVA Results.

Source	*dof*	Wilks λ	*F* (appx)	*df1*	*df2*	*p*
Protein	2	$1.3 \cdot 10^{-5}$	1877	22	148	2e-16*
Box	3	0.37	2.70	33	219	9e-06*
RTC	1	0.93	0.54	11	74	0.87
Protein × box	6	0.62	1.75	66	401	7e-04*
Protein × RTC	2	0.72	1.23	22	148	0.24
Box × RTC	2	0.72	1.21	22	148	0.25
Protein × box × RTC	4	0.59	0.96	44	285	0.55
Total	84					

Results from three-way MANOVA on descriptive properties from simulations, with conditions *Protein,* box type (*Box*), and rotational constraints (*RTC*). Interaction terms are indicated with a multiplication sign. *dof* is the number of degrees of freedom for the specific term, *F* is the approximate *F*-statistic determined from Wilk's lambda, *df1* and *df2* are the numerator and denominator degrees of freedom, respectively. *p* is the probability that the variance due to the source is equal to the residual variance. An asterix marks effects which are statistically significant at the 95% confidence level ($p < 0.05$).

values. This is reassuring, but was expected, as the method should not affect the statistical mechanical ensemble.

The main aim of the present study was to assess the effect of box type on the results of the simulations. The results show that the choice of box type can have a statistically significant effect on the outcome of a simulation. The magnitude of the effect is protein dependent. The box type in effect can act as a constraint on the dynamic behavior of the solute, restricting the conformational ensemble sampled. This could be a direct consequence of the orientation of the solute in the box or its relation to its periodic images. It might also arise from indirect effects due to constraints imposed on the solvent distribution around the solute in the specific box type (lattice) chosen. We note that both the analysis of the descriptive properties, which mainly reflect the structural properties of a mean configuration, as well as analysis of the degree of overlap of motions between different simulations, gave comparable results.

Table 3. ANOVA Results.

Source	*dof*	SS	MS	*F*	*p*
Protein	2	11.48	5.74	5815	2e-16*
Box	9	0.05	0.01	6.15	2e-08*
RTC	2	0.00	0.00	0.46	0.63
Protein × box	18	0.08	0.00	4.55	9e-10*
Protein × RTC	2	0.00	0.00	0.58	0.56
Box × RTC	13	0.01	0.01	0.93	0.52
Protein × box × RTC	13	0.01	0.00	0.97	0.48
Residuals	1320	1.30	0.00		

Results from three-way ANOVA on RMSIP values determined from pairs of simulations, with conditions *Protein, box* type (*box*), and rotational constraints (*RTC*). Interaction terms are indicated with a multiplication sign. *dof* is the number of degrees of freedom for the specific term, *SS* is the sum of squares and *MS* the mean of squares, equal to the sum of squares divided by the degrees of freedom. *F* is the *F*-statistic, equal to the ratio between the mean of squares and the mean of squares of the residuals, *p* is the probability that this ratio is equal to 1. An asterix marks effects that are statistically significant at the 95% confidence level ($p < 0.05$).

The protein 2CI2 was largely unaffected by the type of box in which it was simulated, reflecting the intrinsic stability of this protein. For 1A6S and 1AKI the results show a different picture. For 1A6S, the simulations performed in a rectangular box were significantly different from simulations performed in other box types. This was reflected in higher average α-helix content and the number of hydrogen bonds related to secondary structure elements as well as for the hydrophilic solvent-accessible surface. On the other hand, the average hydrophobic solvent-accessible surface was smaller. The rectangular box favored the formation of secondary structure and suppressed fluctuations. The RMSIP analysis suggested that simulations performed in a rectangular box showed more overlap than in other box types again implying restricted motility, the differences observed were in the range of 5–10%. The differences between simulations of 1A6S in a rectangular and a NDLP box are illustrated in Figure 1B, which shows a stereo image of superimposed average structures from the simulations. From this it is seen that the protein was flexible, as indicated by the large spread in the averages obtained from one box type. This corresponds well with the finding that the RMSIP values from simulations performed in the same box types were relatively low. However, it is also possible to distinguish consistent differences between the sets. Especially with regard to the terminal regions, the behavior of the protein seems to depend on the box type.

1AKI (lysozyme) simulated in a truncated octahedron differed significantly from simulations performed in other box types, with the largest (and statistically significant) differences found between this box and a rectangular or a molecular shaped box. Again, this difference is reflected in both analysis of descriptive properties and RMSIPs. For 1AKI, most of the properties showed differences less than 5%. The superimposed average structures from simulations in a rectangular and a truncated octahedron box, given in Figure 1C, show that the spread is much smaller than for 1A6S. The difference between the sets of averages is found to be small, but consistent, and is in accordance with the results from the statistical

Table 4. Averages and Contrasts (Differences) of the Various Descriptive Properties Described in Figure 2 for 1A6S.

Property	Average	DH − MB	DH − OH	DH − RC	MB − OH	MB − RC	OH − RC
bbRMSD	0.435	0.036	0.006	0.002	−0.030	−0.033	−0.004
ssRMSD	0.405	0.047	0.011	0.011	−0.036	−0.036	0.000
totHbnd	60.167	−3.254	−1.462	0.194	1.792	3.449	1.656
ssHbnd	31.845	0.098	−0.522	−3.258	−0.620	−3.356	−2.737
Rgyr	1.179	0.010	0.006	0.011	−0.004	0.001	0.005
SASphob	28.606	0.364	0.064	1.134	−0.301	0.770	1.070
SASphil	24.391	0.378	0.506	−0.871	0.127	−1.249	−1.377
α-Helix	37.468	2.696	−1.744	−2.585	−4.440	−5.281	−0.842
p-value		0.16	0.82	0.05	0.10	0.00	0.01

analysis. Lysozyme is known to undergo well-defined domain motions, which are believed to be functionally important. In this case, it is clear that the shape of the box could constrain such motions. This will also be true in other proteins that undergo allosteric changes.

Although the cause of the difference is here suggested to be the different distribution of solvent around the solute in the various box types, one can also think of other sources. For example, a difference in the volume will result in a different solute concentration or, when using counterions, in a different concentration of these. Both effects may influence the solute dynamics. In the present study, such effects were implicitly present as part of the difference between the box types. The intention of the work was to compare different box types as they would commonly be simulated. In this regard, we note that no statistically significant differences were detected between the rhombic dodecahedron and NDLP box types with regard to the results obtained. The difference in both the solute and ion concentrations was larger between these two box types than between box types that did yield statistically significant differences, suggesting that at least for the solute and ion concentrations involved in this study the effects are not significant.

Another area of difference between the simulations, which was brought to our attention by one of the reviewers, is that although the protocol used to set up the simulations was identical in all cases, there are significant differences in the water density in specific cases as shown in Table 1. This artefact is due to slight differences in the overlap between water molecules, the protein, and/or the edges of the box due to differences in the orientation of the protein and the shape of the box. This effect can be severe when creating minimal box sizes and is normally corrected by equilibrating at constant pressure, which was not done. For example, in the case of 1A6S the rectangular box solvent density (±940 kg m^{-3}) is significantly lower than the other box types (±970 kg m^{-3}). In this case the possibility that density of the solvent may, in fact, be the dominant factor determining the differences in results cannot be excluded. For protein 2CI2, the differences in solvent density did not lead to differences in simulation results. Finally, differences in solvent density can be excluded as the primary cause for the differences found for 1AKI. Rather, the NDLP box shows a significantly lower density, but gives results similar to those obtained in a rectangular or rhombic dodecahedron box. The truncated octahedron, which has a solvent density similar to those in a rectangular and rhombic dodecahedron box type, is responsible for differences in the simulation results.

Note that together the results demonstrate that the use of an NDLP box can significantly improve the efficiency of simulations, without the introduction of major artefacts. The present results also show that the use of any box type including a rectangular box type or a truncated octahedron carries some risk of introducing arte-

Table 5. Averages and Contrasts (Differences) of the Various Descriptive Properties Described in Figure 2 for 1AKI.

Property	Average	DH − MB	DH − OH	DH − RC	MB − OH	MB − RC	OH − RC
bbRMSD	0.144	0.033	0.026	0.026	−0.007	−0.007	0.000
ssRMSD	0.115	0.009	0.032	0.006	0.023	−0.004	−0.026
totHbnd	109.937	−3.364	0.229	−0.944	3.593	2.420	−1.173
ssHbnd	28.786	−0.242	−0.583	−0.314	−0.341	−0.072	0.268
Rgyr	1.372	0.003	0.000	−0.002	−0.002	−0.004	−0.002
SASphob	26.485	0.479	0.449	0.546	−0.030	0.068	0.097
SASphil	40.257	0.108	−0.051	0.050	−0.160	−0.059	0.101
β-Sheet	9.778	−1.100	0.884	0.480	1.984	1.579	−0.404
α-Helix	46.288	0.404	−0.327	0.874	−0.731	0.470	1.201
p-value		0.48	0.11	0.62	0.03	0.70	0.02

facts. In particular, in a rectangular box rotational motion may also lead to direct interactions between periodic images. It should also be noted that it was found that the effect of the box type on the dynamics was independent of the use of rotational constraints. This shows that the effect of the box type is not caused by direct interactions between periodic images. As the NDLP box can allow a greater distance between periodic images, for the same total system size the use of the NDLP box may well minimize the possibility of periodicity artefacts and allow better sampling of the configurational space.

Conclusion

In this article an assessment has been made of the influence of the box type used in a given simulation on a number of properties that can be regarded as descriptive for the behavior of the proteins. A set of 105 simulations, corresponding to three proteins in four box types and five replicate runs, was analyzed with respect to sampling of conformational space and configurational properties, such as the number of hydrogen bonds, the solvent-accessible surface, and the number of residues involved in secondary structure elements. It has been shown that a small but statistically significant effect may be attributed to the box type used. The nature and magnitude of the effect are, however, strongly dependent on the protein studied. The study also shows that the use of a molecular shaped or NDLP box can increase the efficiency of a simulation without introducing major artefacts. In most cases the optimal box will be the rhombic dodecahedron. Because of its approximate spherical symmetry, this box minimizes effects related to the box or the resulting distribution of solvent. In addition, the results demonstrate the possibility of using ANOVA and MANOVA to compare sets of simulations. These methods enable one to evaluate conditions that may have a significant effect on a simulation. Together with multiple comparison tests, ANOVA and MANOVA allow one to make assessments on the effects of a wide range of conditions.

A basic assumption underlying the analysis in the present study is that external effects may cause a redistribution of conformational densities, which can be expressed in terms of descriptive properties and atomic fluctuations. Such redistribution can also be seen as the mechanism underlying allosteric effects.[30] Thus, the methods of analysis of molecular simulations proposed here may also be applicable in a much broader context, where the binding of a ligand to a target protein, for example, could be treated as an external factor exerting its influence on the system.

Acknowledgments

The authors wish to thank Andrea Amadei, Henk Bekker, and Geert Joosten for helpful discussions. The authors also wish to thank the reviewer for bringing to our attention the fact that the systems setup lead to differences in the density of solvent in certain cases.

Appendix A: Multiple Comparisons Using the Roy Union Intersection Method in R

The program R[24] is a powerful, freely available open-source package for the statistical processing and analysis of data. The standard tests, including ANOVA and MANOVA applied in this work, are available in the standard distribution. Other tests, such as the multiple comparisons using Tukey's Honest Significant Differences,[26] were performed using the package multcomp, acquired from the R Web site. However, the program at present does not provide a method to test multiple contrasts for the multivariate ANOVA model. For this reason an implementation is given for tests of contrasts using a method based on the Roy union-intersection method. This implementation, which was generously provided by Dr. Yves Rosseel from the University of Ghent, Belgium, is given as a function definition below, which can be loaded (sourced) from within R.

The function is a general method to test multivariate linear hypotheses of the type **LBM = K**, where **L** is a contrast vector and **B** is the parameter matrix of the linear model underlying the MANOVA. **M** is usually the identity matrix and **K** the null matrix, both of appropriate dimensions. It should be noted that R reparameterizes linear models, such that the intercept includes the first level of each of the factors. Given the linear model for observations obtained with several levels of two different conditions

$$x_{ijh} = \mu + \tau_i + {}_{j} + \eta_{ij} + \varepsilon_{ijk}$$

reparameterization in R leads to the equation for the intercept

$$\mu_0 = \mu + \tau_1 + \nu_1 + \eta_{11}$$

and effects are represented as deviations from this intercept:

$$\tau_1{}^* = \tau_1 - \tau_1$$

Reducing the model to a one-way analysis of variance for explanatory purposes, a *contrast* of the parameters τ_i is defined as any linear function

$$\sum_{i=1}^{k} c_i \tau_i$$

where the coefficients have the property

$$\sum_{i=1}^{k} c_i = 0$$

These contrasts can be represented by vector notation as:

$$(c_1, c_2, \cdots, c_k)(\mu_0, \tau_2^*, \cdots, \tau_k^*)^T$$

From this it can be understood that the contrast for treatments 1 and 2 from a condition with four levels in R is given by (0, −1, 0, 0), whereas a contrast between treatments 3 and 4 is given by (0, 0, 1, −1).

The function to evaluate contrasts of this form in R is given by the following code, which was written by Dr. Yves Rosseel:

```
mlh <- function(fit, L, M)
{
  if ( !inherits( fit, ”maov” ) )
    stop( “object must be of class \”manova\” of \”maov\”” )
  if ( is.null( dim( L ) ) )
    L <- t( L )
  rss.qr <- qr( crossprod( fit$residuals %*% M ) )
  X <- as.matrix( model.matrix( fit ) )
  B <- as.matrix( fit$coef )
  LB <- L %*% B
  LXXL <- as.matrix( solve( L %*% solve( t( X ) %*% X )
%*% t( L ) ) )
  H <- t( M ) %*% t( LB ) %*% LXXL %*% LB %*% M
  eig <- Re( eigen( qr.coef( rss.qr, H ),
symmetric=FALSE)$values )
  q <- nrow( L )
  df.res <- fit$df.residual
  test <- prod( 1/( 1 + eig ) )
  p <- length( eig )
  tmp1 <- df.res – 0.5 * ( p – q + 1 )
  tmp2 <- ( p * q – 2 )/4
  tmp3 <- p^2 + q^2 – 5
  tmp3 <- if ( tmp3 > 0 ) sqrt( ( ( p * q )^2 – 4 )/tmp3 ) else
1
  wilks <- test
  df1 <- p * q
  df2 <- tmp1 * tmp3 – 2 * tmp2
  F <- ( ( test^( -1/tmp3 ) - 1 ) * df2 )/df1
  Prob <- pf( F, df1, df2, lower.tail=FALSE )
  out <- list(wilks=wilks, F=F, df1=df1, df2=df2,
Prob=Prob)
out
}
```

References

1. Bekker, H.; Van den Berg, J. P.; Wassenaar, T. A. J Comput Chem 2004, 25, 1037.
2. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
3. Bekker, H. J Comput Chem 1997, 18, 1930.
4. Voronoi, G. J Reine Angew Math 1908, 134, 198.
5. Amadei, A.; Chillemi, G.; Ceruso, M.A.; Grottesi, A.; DiNola, A. J Chem Phys 2000, 112, 9.
6. Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Proteins 1993, 17, 412.
7. Amadei, A.; Ceruso, M. A.; DiNola, A. Proteins 1999, 36, 419.
8. Miller, R. G. Beyond ANOVA: Basics of Applied Statistics; Chapman & Hall: Boca Raton, FL, 1997.
9. Chambers, J. M.; Hastie, T. J. Statistical models in S; Wadsworth & Brooks/Cole, Pacific Grove, CA, 1992.
10. Morrison, D. F. Multivariate Statistical Methods; McGraw-Hill Kogakusha Ltd.: Tokyo, 1976.
11. McPhalen, C. A.; James, M. N. G. Biochemistry 1987, 26, 261.
12. McDonnell, J. M.; Fushman, D.; Cahill, S. M.; Zhou, W.; Wolven, A.; Wilson, C. B.; Nelle, T. D.; Resh, M. D.; Wills, J.; Cowburn, D. J Mol Biol 1998, 279, 921.
13. Artymiuk, P. J.; Blake, C. C. F.; Rice, D. W.; Wilson, K. S. Acta Crystallogr B 1982, 38, 778.
14. Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. Comp Phys Commun 1995, 91, 43.
15. Lindahl, E.; Hess, B.; van der Spoel, D. J Mol Model 2001, 7, 306.
16. van der Spoel, D.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Hess, B.; Feenstra, K. A.; Lindahl, E.; van Drunen, R.; Berendsen, H. J. C. Gromacs User Manual Version 3.1; 2002. Nijenborgh 4, 9747 AG Groningen, The Netherlands. Internet: http://www.gromacs.org.
17. van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. Biomolecular Simulation: GROMOS96 Manual and User Guide; BIOMOS b.v.: Zürich, 1996.
18. Schuler, L. D.; van Gunsteren, W. F. Mol Sim 2000, 25, 301.
19. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In Intermolecular Forces; Pullman, B., Ed.; Reidel: Dordrecht, 1981, p 331.
20. Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. J Chem Phys 1995, 102, 5451.
21. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. J Chem Phys 1984, 81, 3684.
22. Miyamoto, S.; Kollman, P. A. J Comp Chem 1992, 13, 952.
23. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. J Comput Phys 1997, 23, 327.
24. Development Core Team, R: A language and environment for statistical computing; http://www.R-project.org, 2004.
25. Roy, S. N.; Bose, R. C. Ann Math Stat 1953, 24, 513.
26. Miller, R. G. Simultaneous Statistical Inference; Springer: Berlin, 1981.
27. Roy, S. N. Some aspects of multivariate analysis; John Wiley & Sons, Inc.: New York, 1957.
28. DeLano, W. L. The PyMOL Molecular Graphics System; DeLano Scientific: San Carlos, CA, 2002.
29. POV-Ray: The Persistence of Vision Raytracer, www.povray.org.
30. Gunasekaran, K.; Ma, B.; Nussinov, R. Proteins 2004, 57, 433.