BUTTERWORTH HEINEMANN

Use of TSAR as a new tool to analyze the molecular dynamics trajectories of proteins

Jacques Haiech,* Thierry Koscielniak,† and Gérard Grassy‡

**Laboratoire de Chimie Bactérienne, Marseille, France*

†Oxford Molecular, X-Pole, Ecole Polytechnique, Palaiseau, France

‡Centre de Biochimie Structurale, UMR C9955-INSERM U414, Université de Montpellier 1, Montpellier, France

There is a lack of tools to analyze simulations of protein molecular dynamics quantitatively. Our aim is to use calmodulin, a prototypical calcium-binding protein, to describe a strategy and some tools for extracting relevant information from dynamics calculations. Our main conclusions are as follows:

- *Autocorrelation vectors may be used to represent a 3D conformation in an* n*-dimensional space, where* n *is variable* ($n \leq 20$–30).
- *On such a transformation, classic statistical tools (PCA, clustering, etc.) may be used to differentiate or characterize dynamics trajectories quantitatively.*
- *TSAR, an integrated package used for quantitative structure–activity relationships, is well suited (after minor modifications) for such a purpose.*

Finally, this type of strategy is able to point out the effects of the solvent screening parameters of the Amber software on the dynamics trajectories of calmodulin.

Keywords: molecular dynamics, autocorrelation, statistical analysis, TSAR, calmodulin

INTRODUCTION

Molecular dynamics simulations create a huge amount of data. Molecular graphics presentation plays an important part in obtaining a global view of what is going on, as far as the protein structure is concerned, during a simulation. Such analyses have been mainly qualitative up to now. The most commonly used and impressive molecular graphics tools are animations. When one needs to simulate a trajectory, one must implement heuristics in order to simulate the solvent or to take into account the electrostatic interactions. Therefore, with the same protein and the same program, two users may end up with different dynamics simulations and be unable to compare their results in a quantitative way.

Furthermore, the analysis of trajectories of correlation of move of a given set of atoms or residues of a protein seems to be more quantitative but remains in fact qualitative. It appears that the amount of information embedded in a trajectory is more complete than what is really drawn even by an experienced user.

The aim of this work is to test the method of conformational autocorrelation first described by Broto et al.[1] to describe the dynamics of proteins globally and quantitatively. This method allows the description of a structural conformation as a set of vectors and, therefore, to describe the dynamics of a protein as a vectorial function of time. The dynamics simulation of a protein is a sequence of different conformations, each conformation being described by a vector in an n-dimensional space. With this type of representation, we now have the ability to use classic statistical techniques such as principal components analysis, discriminant analysis, and clusters analysis to compare and classify the different conformations that appear at any time. The software TSAR, mainly used in the pharmacochemical field, is particularly suited for this purpose.[2,3]

This article describes the method and use of TSAR to analyze the trajectories of calmodulin in order to illustrate this kind of analysis.[4]

METHODS

Starting molecule

Synthetic calmodulin (SYNCAM) has been used to illustrate the proposed method. A SYNCAM three-dimensional

Color plates for this article are on p. 59 and 60.

Address reprint requests to Dr. Grassy, Centre de Biochimie Structurale, UMR C9955-INSERM U414, Université de Montpellier 1, 15, Av. Charles Flahault, F-34060 Montpellier Cedex, France.
Received 20 September 1994; accepted 4 October 1994.

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(94)00012-H

(3D) model was obtained by mutating mammalian calmodulin into SYNCAM as described in Weber et al.[5]

Dynamics simulation protocol

Dynamics simulations were carried out with the AMBER package version 4.0,[6] using standard parameters set and polar hydrogens topology. The nonbonded interaction cutoff was 12 Å, with the shift function applied to the electrostatic term and the switch function applied to the van der Waals term from 10 to 12 Å. Distance-dependent dielectric permittivity was used (ϵ), with no explicit solvent. The nonbonded list was updated every 20 steps. The Velvet algorithm was used, with an integration step of 0.5 fs, using Shake. Each structure was first heated to 300 K in 15 ps, equilibrated for 25 ps, and the productive dynamics were continued for 250 ps.

Molecular descriptors

Unweighted and weighted autocorrelogram. Given a 3D object (i.e., a protein), its 3D autocorrelogram is defined as follows. Choosing a distance step dx, the autocorrelation vector will be of dimension n with $n = d_{max}/dx$, where d_{max} is the distance between the two furthest atoms of the protein in a given conformation. Each component $A(i)$ of the vector will represent the number of couples of atoms separated by a distance between $(i)dx$ and $(i + 1)dx$. An autocorrelation vector can be weighted by some properties defined on each atom (e.g., atomic charges, lipophilic atomic contributions, and van der Waals radii). Therefore, the component $A(i)$ will be equal to $A(i) = \Sigma[P(k) \cdot P(l)]$, where $P(k)$ and $P(l)$ are the weighted properties on two atoms (k, l) with an interdistance contained between $(i)dx$ and $(i + 1)dx$. Usually $dx = 5$ Å for macromolecules; with such a step, the size of the matrix required to calculate the autocorrelogram is not too high.

Statistical analysis methodology.[7,8] *Principal component analysis.* The dynamics simulation provides us with a set of 250 conformations (one per picosecond). Each conformation is described by one autocorrelation vector. Therefore, a dynamics simulation may be described as a set of autocorrelation vectors in a 15-dimensional space (with a step equal to 5 Å, 15 components are sufficient to describe the whole set of conformations of calmodulin). The set of 250 conformations leads to a cloud of 250 points in a 15-dimensional space. Principal component analysis (PCA) is used to describe such a quantity of data geometrically. PCA is a method that can be used to reduce a large number of variables to a smaller number still containing the same information. Any set of weighted points in a multidimensional space has several orthogonal inertia axes (called principal components; the total number of principal components is the same as the total number of dimensions of the original space) that contain the major part of the information about the relative positions of these points. The projection of the points along one of the principal components brings up reduced information about these points, and for each component the fraction of the variance present in the original set of data explained by this component can be easily calculated (the larger the value, the more significant the component). The projection of the points on a principal factorial plane, defined by the two first principal components (or the projection of the data in the 3D space defined by the three first principal components), allows one to see if each conformation brings up any original information. The significance of the information related to this plane or 3D space is the cumulative fraction of the variance explained by the components involved: if two conformations are close to each other in such a plane, their informative contents are similar according to the descriptor used. The shape of the molecular cloud in the first factorial plane reflects the homogeneity or the heterogeneity of the sampled conformations during the dynamics simulation. If the structural variations are poor, the shape of the population will be quasilinear, showing that during the simulation a major part of the structural space has not been explored. When mixing the conformations provided by two different simulations, it will be possible to see if the distribution of conformations is similar or different in the two simulations.

Cluster analysis. Cluster analysis is a technique used to group a set of points into sets that consist of similar members, based on the distance between the points in a chosen parameter space. In this study we used the ascending hierarchical clustering algorithm: initially, each data point is considered to form a cluster, then, the closest pair of points is linked to form one new cluster and this step is repeated until all the points have been amalgamated in the same cluster. During a dynamics simulation, a protein may fluctuate between a finite set of conformations. Cluster analysis is well suited to test and describe such behavior. This analysis allows one to gather similar conformations (as measured by a Euclidean distance in the 15-dimensional space). The trajectory may then be described using this set of clusters and allows one to see how the protein explores the conformational space.

RESULTS AND DISCUSSION

Visualization of structures during the simulation

Two dynamics simulations have been performed for 250 ps on the SYNCAM, using two distance-dependent dielectric constants (respectively, $\epsilon = r$ and $\epsilon = 3r$) to simulate the presence of water macroscopically.[9]

Data issued from the dynamics run are stored using a PDB format (without hydrogens), which can be visualized using the current 3D representation (Color Plate 1), and then a file of files is generated (this is a text file containing a list of filenames, one per line) including, in the first column, the name of the file corresponding to a specific conformation and, in the second column, an ID code indicating the occurrence of this conformation.

From this file of files, TSAR stores information in the form of a spreadsheet consisting of a number of rows and columns. Each row of the spreadsheet holds information referring to a conformation. At the beginning of the row the corresponding entire 3D structure is displayed. (Color Plate 1).

Computation of weighted and unweighted autocorrelogram

The 3D unweighted autocorrelogram is then computed using a step size of 5 Å. It can be automatically added to the

table and will be displayed using a 2D graph, showing on the x axis the interatomic distance of atoms in angstroms and on the y axis the number of atoms for each separation band (Color Plates 2 and 3). The weighted 3D autocorrelograms might be computed in the same way using atomic contribution to log P, van der Waals radius, and so on as weighting factors.

Principal component analysis of each trajectory

Principal component analysis was performed successively on $\epsilon = r$ and $\epsilon = 3r$ trajectories (Color Plates 4 and 5). In our case, in which the contents of all the columns involved are equivalent (distribution of distance), we do not need any standardization. The 3D plot defined by the three first eigenvectors of the PCA explains, respectively, 93% ($\epsilon = r$) and 91% ($\epsilon = 3r$) of the total variance. Examining the component distribution of the autocorrelation vectors (Color Plates 3 and 6) points out an obvious difference between the two trajectories, therefore suggesting that a modification of the distance dependency factor for ϵ induces a strong effect on the dynamic simulation.

Cluster analysis of one trajectory

As PCA projections may perturb the real distances between points (each point belonging to a conformation in the dynamics) we have also performed a cluster analysis in order to visualize the different sets of close conformations. Results of this process are represented schematically by a dendrogram, in which each row (corresponding to a 3D conformation) is represented by a horizontal line at the left-hand side of the dendrogram (Color Plates 7 and 8), a link between two points is represented by a vertical bar joining these two points, and the distance between the different points is shown on the horizontal axis. A color coding based on the different clusters found (95% confidence limit) is added. Again, a clear difference between the two simulations is exhibited through this technique. A quantitative evaluation of the differences between the conformations produced during the molecular dynamic runs is then possible, in order to evaluate quantitatively the effects of the different parameters used in the simulation.[10]

CONCLUSION

Up to now, tools with which to extract information and to compare molecular dynamics trajectories of macromolecules such as proteins or DNA have been missing. Using SYNCAM as a prototypical protein, we have shown the following:

- The representation of protein conformations by means of auto correlation vectors allows one to describe a geometric shape as a vector in an n-dimensional space. Such description allows one to use classic tools to describe the temporal evolution of the 3D conformation of macromolecules during a dynamics run.
- The integrated package TSAR, which has been developed to investigate interactively the quantitative structure–activity relationships (QSARs) of small molecules, is well suited to this kind of analysis with minor modifications, such as increasing the limit number of atoms in order to store full trajectories.
- This kind of analysis may now allow the quantitative determination of the effect of the various parameters used by the dynamics simulation software.

Such a strategy may be used to compare dynamic simulations obtained by different authors or under different conditions on the same proteins. Future works will be aimed at optimizing the TSAR package for use with proteins or with DNA.

REFERENCES

1 Broto, P., Moreau, G., and Vandycke, C. *Eur. J. Med. Chem.* 1984, **19,** 61–84
2 TSAR: Tools for Structure Activity Relationships. Oxford Molecular, Ltd., The Magdalen Centre, Oxford Science Park, Oxford OX4 4GA, U.K.
3 Grassy, G. *Acta Chim. Ther.* 1987, **14,** 191–204
4 Haiech, J. *Acta Chim. Ther.* 1993, **20,** 295–308
5 Weber, P., Lukas, T., Craigh, T., Wilson, E., King, M., Kwiatowski, A.,. and Watterson, D. *Proteins* 1989, **6,** 70–85
6 Pearlman, D.A., Case, D.A., Caldwell, J.C., Seibel, G.L., Chandra Sing, U., Werner, P., and Kollman, P.A. Amber 4.0. University of California at San Francisco, San Francisco, CA, U.S.A.
7 Cailliez, F. and Pages, J.P. In *Intoduction à l'Analyse de Données,*SMASH Ed., Paris, 1976
8 Benzecri, J.P. In *L'Analyse de Données,* Dunod Ed., Paris, 1973
9 Schiffer, C.A., Caldwell, J.W., Kollman, P.A., and Stroud, R.M. *Proteins* 1990, **8,** 30–43
10 Yasri, A., Haiech, J., and Grassy G. 1994 (to be published)