Structural bioinformatics

Wordom: a program for efficient analysis of molecular dynamics simulations

Michele Seeber, Marco Cecchini, Francesco Rao, Giovanni Settanni and Amedeo Caflisch*
Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Received on May 9, 2007; revised on July 11, 2007; accepted on July 15, 2007
Advance Access publication August 23, 2007
Associate Editor: Alfonso Valencia

ABSTRACT
Summary: Wordom is a versatile program for manipulation of molecular dynamics trajectories and efficient analysis of simulations. Original tools in Wordom include a procedure to evaluate significance of sampling for principal component analysis as well as modules for clustering multiple conformations and evaluation of order parameters for folding and aggregation. The program was developed with special emphasis on user-friendliness, effortless addition of new modules and efficient handling of large sets of trajectories.
Availability: The Wordom program is distributed with full source code (in the C language) and documentation for usage and further development as a platform-independent package under a GPL license from http://www.biochem-caflisch.unizh.ch/wordom/
Contact: caflisch@bioc.unizh.ch

1 INTRODUCTION
Molecular dynamics (MD) simulations can produce several terabytes of data (Rueda et al., 2007) whose manipulation and analysis is problematic in the absence of proper tools. As an example, sampling in the millisecond time-scale generates on the order of $10^7$ to $10^8$ coordinate frames (Settanni et al., 2005). In this communication, a program is presented that can effectively access MD trajectory data and make them available for conversion, manipulation and analysis by different modules. In this way, it is possible to implement new procedures by focusing on the relevant equations and algorithms without having to deal with the data access module. The program provides general tools for analysis as well as special purpose modules which are not available in other programs (e.g. the efficient clustering of large sets of conformations and the evaluation of order parameters for monitoring polypeptide aggregation).

2 DATA MANIPULATION AND OUTPUT
Wordom can perform simple conversion and manipulation tasks, such as the extraction of one or more snapshots from a CHARMM (Brooks et al., 1983), GROMACS (van der Spoel et al., 2005), or NAMD (Kalé et al., 1999) trajectory into a pdb (Berman et al., 2000) or crd file (Brooks et al., 1983), adding a single coordinate set from a pdb or crd file to a trajectory, merging segments of trajectories, calculating the average structure along a trajectory, extracting xyz coordinates of selected atoms, showing/modifying the content of a trajectory file header, etc. For most of these tools an atom selection mechanism is implemented. Notably, all of these functions are accessible with a single command line. Moreover, the coordinate and trajectory files generated by Wordom can be visualized by widely distributed graphics packages, e.g. VMD (Humphrey et al., 1996) and Pymol (DeLano, 2002).

3 ANALYSIS OF MD SIMULATIONS
Wordom provides modules for the analysis of geometrical features (distances, contacts, dihedrals, hydrogen bonds), as well as evaluation of radius of gyration, DRMS (distance root mean square) and RMSD (root mean square deviation) from a reference structure. RMSD can also be computed with respect to the preceding structure, to highlight conformational changes along the trajectory. More complex types of analysis include a principal component analysis (PCA) module, the computation of quasiharmonic entropy, several clustering algorithms and the evaluation of order parameters to monitor folding (i.e. number of native and non-native contacts) and amyloid aggregation.

3.1 PCA analysis
PCA is a useful technique for separating large amplitude motions from irrelevant fluctuations along an MD trajectory (Amadei et al., 1993). The approach consists of diagonalizing the covariance matrix of atomic fluctuations with respect to the average structure, thus obtaining a set of eigenvectors and corresponding eigenvalues which represent the axes of maximal variance in the protein motion. An original feature in the PCA module is a procedure for iterative PCA evaluation along trajectory intervals of increasing length (e.g. snapshots 1–1000, snapshots 1–2000, etc.) to evaluate the significance of sampling. The eigenvectors computed at each time interval are projected into the ones of the preceding (i.e. shorter) time interval. A value close to 1 along the diagonal of the projection matrix indicates convergence of the corresponding eigenvector(s) suggesting that trajectory elongation is not necessary for the
calculation of those eigenvectors. The PCA module in Wordom can also be used to generate pdb files where the last field of each atom entry (which usually contains the crystallographic \( \beta \)-factor) is set equal to the length of the vector defined by the components of the selected eigenvector along the corresponding atom. This quantitative information on the protein segment(s) involved in the most relevant concerted motion can be displayed directly with available graphics programs (Fig.1, left).

### 3.2 Quasiharmonic entropy

The calculation of the quasiharmonic entropy is based on the diagonalization of the mass-weighted covariance matrix of the atomic fluctuations (Andricioaei and Karplus, 2001). The \( 3n - 6 \) non-zero eigenvalues \( \lambda_i \) are related to the quasiharmonic frequencies \( \omega_i = \sqrt{kT/\lambda_i} \) (where \( T \) is the simulation temperature and \( k \) the Boltzmann constant), which in turn yield the vibrational entropy \( S = k \sum_i \omega_i \lambda_i \ln \lambda_i - \ln(1 - e^{-\hbar \omega_0/kT}) \), the quasiharmonic vibrational energy \( E = \sum_i \omega_i \lambda_i \), and the vibrational specific heat \( C_V = \partial E/\partial T \). The entropy measured in this way is directly related to the conformational space explored by the system along the MD simulation and provides a wide range of applications from estimation of the entropic effects of mutations in the native state of proteins to the determination of the relative entropy of ligands in a binding pocket (Thorpe and Brooks, 2004).

### 3.3 Clustering algorithms

Clustering can be performed using two metrics (DRMS or RMSD) and three algorithms: hierarchical (Johnson, 1967), quality threshold (Heyer et al., 1999) or leader-like (Hartigan, 1975). These algorithms have different CPU-time requirements and yield different clustering performances, hierarchical and quality threshold being slower but more precise and leader-like being much faster but dependent on the frame order. For the former, it is possible to first take into account a subset of the snapshots and then cluster the remaining snapshots to reduce memory and CPU-time requirements. In the case of RMSD-based clustering, superposition can be carried out before computing the distance, depending on whether the relative position of the structures is relevant. It is also possible to superpose the structures according to a subset of atoms and cluster them relatively to another subset, which is useful for analyzing multiple poses obtained by ligand docking into a single protein structure.

### 3.4 Order parameters

Wordom efficiently evaluates the number of contacts, i.e. \( \alpha \)-pairs with a distance below a given threshold (input value). The subset of contacts which are present in the native structure is a useful variable to monitor the progress of (un)folding (Chan and Dill, 1998).

The polar and nematic order parameters are used to study amyloid formation (Cecchini et al., 2004). They are defined as

\[
P_1 = \frac{1}{N} \sum_{i=1}^{N} \mathbf{d} \cdot \mathbf{\hat{N}} \quad \text{(polar)}
\]

\[
P_2 = \frac{1}{N} \sum_{i=1}^{N} \mathbf{\hat{N}}^2 \quad \text{(nematic)}
\]

where \( \mathbf{d} \) is a unit vector pointing in the preferred direction of alignment, \( \mathbf{\hat{N}} \) is a suitably defined molecular vector and \( N \) is the number of peptides. These order parameters capture different orientational properties of the system and yield useful and complementary information. The polar \( P_1 \) describes the polarity of the system, i.e. how much the molecular vectors (\( \mathbf{\hat{N}} \)) point in the same direction, and discriminates between parallel, antiparallel or mixed ordered aggregates. The nematic \( P_2 \) describes the orientational order of the system and discriminates between ordered and disordered conformations.

The evaluation of \( P_1 \) and \( P_2 \) facilitates the analysis of simulations of peptide aggregation (Cecchini et al., 2004) by measuring the degree of structural order during the fibril formation process (Fig.1, right).

### 4 CONCLUSIONS

Analyzing the ever growing amounts of MD trajectory data can become the rate-limiting step in MD simulation studies. The Wordom program facilitates this task. Manipulations of multiple trajectory files and simple geometric analyses require a single line of input, while input files for clustering and PCA analysis consist of about five lines. Therefore, for efficiently extracting information from massive sets of MD data Wordom is better suited than other programs which require complicated input files and/or the use of a graphical interface. Moreover, thanks to low-level data access functions, adding new analysis modules requires only the coding of the core function allowing for a steadily growing library of algorithms introduced by the community of users.

### ACKNOWLEDGEMENTS

The authors thank F. Raimondi and F. Fanelli for Figure. 1, left and helpful discussions, A. Cavalli, P. Kolb and D. West for technical help and R. Friedman and F. Marchand for comments to the manuscript. This work was supported by grants from the SNF and NCCR-neuro to A.C.
Conflict of Interest: none declared.

REFERENCES