Software News and Updates

Simulaid: A Simulation Facilitator and Analysis Program

MIHALY MEZEI
Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, New York 10029

Received 28 October 2009; Revised 21 January 2010; Accepted 27 February 2010
DOI 10.1002/jcc.21551
Published online 13 May 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Simulaid performs a large number of simulation-related tasks: interconversion and modification of structure and trajectory files, optimization of orientation, and a large variety of analysis functions. The program can handle structures in PDB (Berman et al., Nucleic Acids Res 2000, 28, 235), Charmm (Brooks et al., J Comput Chem 4, 187) CRD, Amber (Case et al.), Macromodel (Mohamadi et al., J Comput Chem 1990, 11, 440), Gromos/Gromacs (Hess et al.), InsightII (InsightII. Accelrys Inc.: San Diego, 2005), Grasp (Nicholls et al., Proteins: Struct Funct Genet 1991, 11, 281) **.crg**, Tripos (Tripos International, S. H. R., St. Louis, MO) .mol2 (input only), and in the MMC (Mezei, M.; MMC: Monte Carlo program for molecular assemblies. Available at: http://inka.mssm.edu/~mezei/mmc) formats; and trajectories in the formats of Charmm, Amber, Macromodel, and MMC. Analysis features include (but are not limited to): (1) simple distance calculations and hydrogen-bond analysis, (2) calculation of 2-D RMSD maps (produced both as text file with the data and as a color-coded matrix) and cross RMSD maps between trajectories, (3) clustering based on RMSD maps, (4) analysis of torsion angles, Ramachandran (Ramachandran and Sasiskharan, Adv Protein Chem 1968, 23, 283) angles, proline kink (Visiers et al., Protein Eng 2000, 13, 603) angles, pseudorotational (Altona and Sundaralingam, J Am Chem Soc 1972, 94, 8205; Cremer and Pople, J Am Chem Soc 1975, 97, 1354) angles, and (5) analysis based on circular variance (Mezei, J Mol Graphics Model 2003, 21, 463). Torsion angle evolutions are presented in dial plots (Ravishanker et al., J Biomol Struct Dyn 1989, 6, 669). Several of these features are unique to Simulaid.

 J Comput Chem 31: 2658–2668, 2010

Key words: simulation; analysis; conversion; orientational optimization; RMSD; circular variance

Introduction

In the early years of molecular simulation, the time spent preparing and analyzing a molecular simulation was negligible compared with the time required for the simulation itself, However, with increased CPU power the balance shifted so that nowadays setup and analysis takes up a significant fraction of the time involved in computation. Thus, any tool that makes the pre- and post-processing more efficient is a welcome development.

The development of Simulaid started in the mid 90s when the proliferation of name conventions and file formats made setting up a simulation quite frustrating. At first, the development of Simulaid focused on the name and file format conversions as well as on various cleanups. Once the framework of dealing with molecular structures had been established, Simulaid also became a vehicle for the development of various analysis tools (e.g., circular variance analysis[1] or the TRAJELIX helix analysis[2]). Other features have been implemented to avoid both having to use several programs, and having to write the interface to them (e.g., DSSP,[3] dial plots,[4] or 2-D RMSD maps). However, in most cases, the analysis has been enhanced from the original version. As the program's use increased, users requested interfaces with other programs, resulting in interfaces to the PB solvers Delphi[5] and UHBD,[6] the PB/SA module of Amber,[7] the Oniom module of Gaussian,[8] and the semiempirical program Amsol.[9] As the program evolved, it became progressively easier to accommodate new requests. Features have also been implemented just for "completeness sake" or in response to new developments (e.g., the "long" structure file format for Charmm) or for increasing the ease of use and robustness of the program (e.g., keyboard logging and online help).

Since 2001, Simulaid is available from the author's web site and has been requested hundreds of times—in many cases by users requesting updated versions. It is now listed on several software collection websites:

http://www.randomfactory.com/lfc/credits.html
http://www.ccp5.ac.uk/librar.shtml
http://cheminformatics.org/phplinks/index.php?PID = 8&PHPSESSID = 4d188c21c3213b9913f806dfd5fd76e7&sr = 10&pp = 10&cp = 2
http://www.the-science-lab.com/Biology/Biophysics.html
http://chemport.ipe.ac.cn/cgi-bin/chemport/getfiler.cgi?ID=v2m5sjh7hemp2NWR1WHFAuJsnYV8X5JZoYts0NKJl9bKrniwG-bof7bb7lnBTNd5Q&VER = C

Correspondence to: M. Mezei; e-mail: mihaly.mezei@mssm.edu

The development of Simulaid is an ongoing process as new ideas come up or request for new feature(s) reach the author. The distribution website has a list of recently implemented features that is kept up-to-date.

Computational Methods

Most features of Simulaid require the input of a single structure file to start with. This file is used to establish the connectivity matrix (based on bond-length thresholds for different chemical bonds) and the set of solvent molecules (assumed to follow the rest, collectively referred to as the solute). Several features of the program rely on the topology thus prepared. Simulaid prints several descriptors of the structure read: atom number and residue number ranges of the different solute molecules (i.e., chains/segments), the number of L and D protein residues (if any), the estimated volume of the protein and nucleic acid residues, and the range of x, y, and z coordinates in the structure read. The connectivity matrix is analyzed for likely anomalies (e.g., hydrogens with more than one bond or atoms with no bond).

Several features of Simulaid treat periodic boundary conditions. The following cells types are implemented: (a) cube, (b) rectangle, (c) Wigner-Seitz cell of face-centered cubic close packing, (d) hexagonal prism, (e) truncated octahedron, coordinate axes going through the square faces (Charmm[10] convention), (f) truncated octahedron, coordinate axes going through the hexagon faces (Amber[7] convention), and (g) Wigner-Seitz cell of hexagonal close packing. Although the hexagonal close packing cell is one of the two optimal cell shapes (smallest volume/inside sphere volume), it has not been used so far in simulations—perhaps because finding an image involves a rotation added to the usual translation.

There is a hyperlinked documentation of the features of Simulaid included in the distribution specifying the formats of files defined specifically for Simulaid. Furthermore, the majority of interactive quizzes (where the prompt may not be fully informative) provide a help option: typing "?" as the answer will result in an explanation and a repeat of the quiz.

In the following description, the various features of Simulaid are grouped together under the menu structures used to access these features.

Open Log File

When a log file is opened, all keyboard input is written to this file. The run can be repeated by redirecting the keyboard input ("standard input") to a previously generated log file or a version that was modified to reflect some changes in the run parameters. However, there can be problems with this approach: if the first run created any files, the repeat run may request an additional input (to allow overwriting the existing file); or if the input file is different, an additional input request may be triggered. To avoid such conflict, the user is given the option to make the input predictable, i.e., request all possible input, even when the answer is unequivocal from the data read so far.

Cleanup of Structure Files

Structure files from various sources can have a number of defects; e.g., missing or nonconsecutive atom and or residue numbers, missing chain ID, etc. The cleanup option will enter chain ID when it is missing and make sure that both the atom numbers and the residue numbers are consecutive. The user has the option of specifying the starting atom and residue numbers, as well as the option of restarting the residue numbers at the start of each chain.

Structure File Conversions

Depending on the amount of information in the input structure file, the file can be converted into the following formats: (a) Charmm CRD; (b) Extended Charmm CRD; Brookhaven PDB; Charmm PDB; Macromodel; MMC Monte Carlo; Gromos/Gromacs; InsightII car; Amsol 3.5; Gaussian[8] ONIOM[11]; or a simple file of the coordinates prefixed with atomic number or atom symbol. Also, atoms can be reordered based on a Charmm residue-topology file (RTF).

Trajectory File Conversions

Trajectories can be converted to (a) Charmm, (b) Amber, (c) Macromodel, and (d) Xcluster formats. Input trajectory can be in Charmm, Amber, or Macromodel format, as well as in the format used by the Monte Carlo program MMC.

The trajectory conversion allows some filtering. The filtering can either (a) select a reduced set of frames from the trajectory; and/or (b) recenter the solute molecule(s) into a periodic cell (useful when a trajectory generated by some of the MD programs contains solute molecules that moved away from the simulation cell), as well as (c) use only a subset of each structure (e.g., drop the solvents). In addition, (d) each structure can be rotated. Besides changing the format and/or or the content of a trajectory, asking for a reduced set of frames can be also used to fix the header of a damaged Charmm trajectory.

An additional unique feature of the trajectory conversions is the option to change the atom order in the converted trajectory. The new order is established either by a file with the new sequence number of each atom, or by a second structure file where the atoms are in the new (target) order. If the sequence numbers in this file refer to the original order of the atoms, Simulaid can use them to establish the new order. If the sequence numbers are consecutive in the second structure, then Simulaid can establish the new order by matching the atom and residue names of two user-provided structure files. In the latter case, residues are assumed to be in the same order and the match is limited to rearranging atoms within a residue. In the Appendix, a protocol is presented to convert a Charmm/NAMD-generated trajectory for analysis in Amber that includes such a rearrangement.

Atom and Residue Name Conversions

Although in principle protein and nucleic acid atom names are standardized by the PDB format, in practice several varieties have been developed. In particular, the PDB convention that the first 2 characters of the atom name is the chemical element symbol, and thus that one-character element symbols have to be the second character, has not been followed by several programs, giving rise to undesirable outcomes: e.g., alpha carbons that are usually called CA are interpreted as calcium atoms, resulting in

a mixed up bond-topology. Simulaid has several options that help to navigate the name-convention minefield: (a) atom names can be "regularized" (i.e., made to conform to the PBD convention discussed above), (b) the regularization may be undone (i.e., names are left-adjusted and if a leading digit is found, moved to the end of the name), or (c) atom and residue names can be converted from one specific convention to an other. This latter function is governed by a file "**pdb_nam.dat**" (included in the distribution package) that currently has the information of protein and nucleic acid residues in many of the popular conventions. This file can be extended by the user to deal with residues or conventions not yet included.

Trajectory/Structure File Conversions

These conversions can create single structure files from a trajectory or from a file that has a number of consecutive structure files; or reverse the process: i.e., create a trajectory from single structure files. When creating single structure files, the user has the option of selecting specific frames to extract by providing a list of frame numbers (entered either from the keyboard or read from a file), or by selecting the first and last frames, as well as an increment for the extraction. The extracted file names will be of the form <name>.<number.<ext>, where <number> is either the original frame number or the new sequence number. When combining structure files into a trajectory, the structures to be combined could be either single files or files containing a set number of structures. When the structures to combine are in more than one file, the name of each should be of the form <name>.<number>.<ext>. Here <number> should start at one and be incremented by one; for the first file, however, ".1" can be omitted.

Animate Trajectory

When compiled with the optional Iris-GL code, Simulaid can also animate a trajectory. Although only line representation is implemented and the atom selection options are quite limited, Simulaid is the only program that can animate a trajectory with variable number of solvents, such as generated by a grand-canonical ensemble simulation

Editing Conformations

There are various ways that structures can be modified with Simulaid.

Delete Selected Atoms

The atoms to be deleted can be specified in a number of ways: (a) specify a chain (segment) to retain (i.e., delete all chains but the one specified) or to delete, (b) delete all but the protein backbone atoms, (c) delete all but the alpha carbons, (e) delete all aliphatic hydrogens, (f) specify a range of atom numbers to delete or to keep, and (g) specify a range of residues (chain/segment ID and residue number) to delete or to keep, or (h) delete all solvents. The different types of selections can be applied consecutively; e.g., first specify a chain to keep and then specify several residue ranges to delete.

Translate and Rotate the Conformation

This feature allows the user to perform consecutive translations and rotations of a structure. The translation can be specified either by entering the displacement vector coordinates or by asking to center the system. The rotation can be performed either by specifying one of the three coordinate axes and the angle to rotate around it, or by entering a 3×3 matrix of rotation (that will multiply the coordinates used as column vectors). Several such operations can be executed in consecutive steps.

This feature also allows the user to ask to keep all molecules within a periodic cell. In that case, once the requested translations and rotations are performed, all molecules (i.e., solvents and different chains/segments) will be checked to ascertain whether they are still within the periodic cell and will be translated back if they are not.

Mutate or Extend the Structure

This feature allows the user to either: change the identity of an atom (name and charge), add new atoms (one at a time), or add new bonds (one at a time). When adding an atom, the user has to specify, beside the name and charge of the new atom, an existing atom to which the new atom is bonded, the bond length, plus the bond angle and torsion angle formed by the new atom and a neighbor and a second neighbor of the selected atom. When the neighbor choice is unequivocal, the program will make the choice; when there are choices, the program will list the possible neighbor candidates. Adding bonds may be employed to "fix" the Simulaid-generated topology when the structure read has some bonded atoms farther from each other than Simulaid's bond threshold.

Replace the Coordinates of the Structure

This feature reads a second structure (that may be in a different format), and transfers its coordinates to the original structure.

Fix Water Coordinates

This feature replaces all water molecules with waters having the experimental geometry, but keeping the original position and orientation of each water molecule.

Extract a Periodic Cell

This feature deletes from a conformation any solvent that falls outside a periodic cell whose shape and dimensions are specified by the user.

Generating Structure Derived Files

From a given structure file, Simulaid can generate several different files for use by other programs.

Extract the Sequence

The extracted sequence can be written in the format of (a) PDB SEQRES records, (b) Charmm, (c) Wisconsin GCG, (d) PIR, and (e) title followed by 1-character residue names.

*Create a **.dat** File for UHBD*[6]

The program UHBD requires a file that contains the specification of atom types, partial charges, Lennard-Jones parameters and charge group information for each atom in a residue. This information is in the so-called topology and parameter files for the programs Charmm, Amber, and Gromacs (each using its own format). Simulaid can convert the information in these files into the format expected by UHBD.

Create Torsion Description Input for Macromodel or MMC

The Monte Carlo algorithms in Macromodel and in MMC provide for sampling a selected list of torsion angles. This function prepares the list of possible torsions in either of these formats, annotated by the names of the atoms involved. For Macromodel, torsions over peptide bonds can be omitted. For MMC, torsions over peptide bonds or over all backbone angles can be omitted. In addition, torsions moving only hydrogens can be omitted. The list thus generated will, however, include torsions over atoms that are in rings or loops—these may be filtered out by MMC or Macromodel. Also, torsions that are to be kept fixed must be removed manually by the user.

Create Charmm RTF Residue Record

For selected residue names, Simulaid will create an RTF file entry of that residue, based on the connectivity deduced from the coordinates of the input structure. Internal coordinate (IC) records will not be generated.

Print the Characteristics of a Periodic Cell

This feature writes the coordinates of the neighboring periodic cell centers on a **.pdb** file and of the vertices of the periodic cell to another **.pdb** file.

Orientational Optimizations

Determine the Optimal Orientation in a Periodic Cell

The orientation of a solute in a periodic cell affects the smallest image–image distance.[12] Finding the orientation of the solute where the smallest image–image distance is the largest under a given type and dimensions of periodic boundary conditions (PBC) can do either of the following: (1) improve the quality of the simulation; or (2) allow the reduction of the PBC cell size without reducing the quality of the simulation. Depending on the size and shape of the molecule and the smallest image–image distance allowed, optimization results in a significant reduction of the number of solvent (water) molecules. Note also, that there is another method for orientational optimization developed by Qian et al.[13]

Determine the Optimal Orientation in a Box

For calculations that involve a grid of a fixed number of grid-points laid over the solute (e.g., Delphi,[14] cavity-biased grand-canonical ensemble simulations) with the requirement that the shortest distance from the solute to the nearest edge exceed a given minimum value, the orientation of the solute in the box influences the grid spacing. Finding the orientation that can be enclosed in the smallest cube or rectangular box results in the finest grid spacing because, due to the fixed number of grids, the grid size is proportional to the edge of the enclosing box.[15]

Determination of the Smallest Enclosing Sphere

When simulating a solute in a droplet, use of nonspherical shape will result in significant distortions due to surface tension.[16] Thus, it is important to use a spherical droplet in simulations. The smallest droplet with a given minimum solvation layer thickness will be centered at the smallest enclosing sphere of the solute. However, determining this sphere is a nontrivial task.[12]

Clustering Atoms

This option calculates the distances between the atoms of the structure read, and uses one of the methods implemented to cluster them (see Analyses Section below).

Analyses

Simulaid is capable of performing a large variety of analyses. They are grouped together (both in the analysis menu and in this paper) by analysis type. The following analyses will generate graphical output in Postscript format: plots of functions, color-coded matrices, traces of various bonds, and dial plots[4] that describe the evolution of angles during a simulation. For each graphical output, the values visualized are also printed on a separate file with annotations.

Some of the analyses involve the time evolution of angles. These are visualized with dial plots[4] (developed in the Beveridge Laboratory). The time axis is the radius of the dial and the angle value determines the position of the plot at that time (radius). Changes in angle are drawn as arcs of color cyan. The initial angle is shown in blue inside a small gray disk at the center of the dial, and the final angle as a blue tick outside the dial. The average over the whole time period analyzed is shown as a red line drawn at the average angle. The user has the option of plotting the averages over successive time periods, resulting in a less crowded plot. The number of dials per row can also be specified by the user, providing control over the size of the dial plots.

Several analysis functions perform clustering. Currently three clustering algorithms are implemented in Simulaid: (a) single link clustering,[17] (b) *K*-means clustering,[17] and (c) maximum neighborhood clustering.[18] Single link clustering connects all pairs of elements that are within a preselected threshold; all elements that are connected by a path of such connections will be in the same cluster. *K*-means clustering requires a preselected number of clusters and the clustering proceeds in an iterative manner by selecting putative cluster centers. In each iteration, each elements of the set is first assigned to the center it is nearest to, followed by selection of a new center that is in the "middle" of each such cluster. When the metric of clustering is Cartesian distance between coordinates, the average of the cluster members' coordinates can serve as the new putative center. For other metrics, such averaging may be meaningless (e.g., for the RMSD between conformations). In such cases, Simulaid selects as the new center a cluster member whose

largest "distance" from the rest of the cluster members is the smallest. The iteration stops when the cluster membership does not change. Maximum neighborhood clustering selects the element that has the most neighbors within a preselected threshold, and makes this element and its neighbor a cluster. The cluster thus established is removed and the procedure is repeated until no more elements remain. For single link and maximum neighborhood clustering, the user can specify either the distance threshold or the number of clusters required. In the latter case, Simulaid will adjust the threshold until the required number of clusters is obtained.

Geometry/Topology Analyses

These analyses provide information on bonds, angles and torsions.

Bond, Angle, and Torsion List. This feature prints the name, number and residue information of atoms forming bonds, plus the bond length. If requested, the similar information is printed for all angles and torsions as well.

1–4 Neighbor List. This feature lists all atom pairs that are separated by three bonds, as well as the torsion angle formed by these three bonds.

Functional Group Analysis. This feature prepares a list of chemical groups found in the structure read. This analysis will identify atoms that are not members of any standard functional groups (if any). Unexpected presence of such atoms usually indicates some problem with the structure under consideration.

Bond, Angle, and Torsion Distribution. For all bond and angle types, Simulaid will calculate the average, range, and standard deviation. In the present context, different types mean bonds between different chemical elements.

Analyses of Nonchemical Bonds

Simulaid can analyze three different types of interactions (bonds): hydrogen bonds, salt bridges, and hydrophobic bonds. Hydrogen bond between a polar hydrogen attached to a donor heavy atom D and an acceptor atom A with negative partial charge requires that both the A-D distance and the A···H-D angle are below a user-defined threshold. Salt bridges are considered when two oppositely charged atoms (groups) are within a user-defined threshold. Hydrophobic bonds are considered when two aliphatic carbons are within user-defined threshold and are also separated by at least a user-defined number of chemical bonds (allowed values: 2, 3, or 4).

Hydrogen-Bonded Chains. Simulaid can detect hydrogen-bonded bridges formed by waters between two solute atoms. These bridges are identified and statistics are prepared for the occurrence of various bridge types. The user can restrict the solute atoms that can anchor such a bridge; the end of the bridge may either be required to be also an anchor atom, or all solute atoms can be allowed.

Bond Tracks. For hydrogen bonds or hydrophobic bonds or salt bridges, Simulaid can prepare a track of their presence or absence. The track is a plot whose horizontal axis is the simulation time, and each bond is assigned a value (evenly spaced) on the vertical axis, in the order the bond first appeared. For each bond, lines are drawn between the time of the appearance and disappearance of that bond. These tracks are also useful to assess convergence: the envelop of the tracks should level off before the simulation has ended, because an envelop of increasing shape suggests that there are still bonds that have not been formed at all during the simulation, but would appear in a longer run. The tracks thus generated can be clustered using the correlation among them as the metric, and the bond tracks rearranged by clusters may then be replotted.

The plot of bond tracks is supplemented with a plot of the number of bonds as a function of the simulation time. The graphical output is concluded with a matrix that represents the frequency with which each pair of residues is bonded during the simulation.

Atomic Property Analyses

Simulaid can attach certain atomic properties to each atom and print a structure file with this value. For PDB files, the value will be in the temperature factor column, while for Charmm CRD files in the weight column.

Hyrophobicity (Hydropathy) Calculation. Simulaid has implemented the (a) Kyte–Doolittle hydropathicity scale,[19] (b) Eisenberg normalized consensus scale,[20] and the (c) White group octanol scale.[21] An option is also provided for the user to enter a different scale. Many more such scales have been developed—an extensive collection has been presented and analyzed by Palliser and Parry.[22] Each atom in a residue is assigned the same value specific for that residue.

Circular Variance Calculation. It has been shown that the circular variance[23] of the vectors drawn from a test point to a collection of other points is diagnostic of whether the test point is inside, near the surface or outside the set of points to which the vectors are drawn.[1] This option calculates the circular variance of all atoms with respect to the atoms of the solute molecule(s).

Delphi Potential Labeling. The potential energy interpolated from a Delphi[5] map can be interpolated to obtain the electrostatic potential at the position of each atom.

Molecular Property Analyses

Simulaid can calculate several properties that are molecule related. These calculations will be performed separately for each molecule (chain/segment) on the structure read and—if requested—on each structure of a trajectory.

Solvation Shell Volume Calculation. A Monte Carlo algorithm is used to estimate the volume of the solvation shell around the solute. If the solute consists of more than one molecule, the volume of the interface between each pair will also be calculated.

Principal Axis Calculation. Simulaid can calculate the direction of the principal axes of each molecule in the structure analyzed.

Radius of Gyration and Hydrodynamic Radius Calculation. Simulaid can calculate both the radius of gyration and the hydrodynamic radius of each molecule in the structure analyzed. When partial charges are available, the dipole moments are also calculated.

Root Mean Square Deviation (RMSD) Calculations

RMSD between two structures can be calculated by Simulaid either without overlay or after first overlaying the structures to be compared. The overlay uses the Kabsch formalism.[24] The set of atoms to be used for the overlay and the set of atoms to calculate the RMSD can be each be selected by the user.

1-D RMSD Calculation. For a trajectory, this option calculates the RMSD between the input structure (or the first structure in the trajectory) and each frame in the trajectory and plots it as a function of simulation time.

2-D RMSD Calculation. This option calculates the RMSD between all pairs of (selected) structures in a trajectory, prepares a color-coded matrix of these values, a plot of the average RMSD over different run lengths, as well as a plot of the distribution of the frequency of occurrence of the various RMSD values.

Because the RMDS distribution over a trajectory generated by a random walk based on uniformly distributed probabilities will not be constant, the plot of the calculated frequency distribution is supplemented with a plot of the distribution normalized by the distribution expected from random walk. This normalized plot usually will provide a better display of second or third peaks of the distribution—such peaks are diagnostic of events occurring at different time scales.

In addition, the RMSD values can be used as the metric for clustering the conformations of the trajectory. If clustering is requested, the 2-D RMSD matrix can be replotted after sorting the trajectory frames by clusters, with lines drawn to delineate the clusters. In addition, a plot of cluster membership as a function of simulation time will be generated—another useful tool for the assessment of the convergence of the simulation.

Cross RMSD Calculation. Simulaid can also calculate the RMSD between pairs of selected structures in two different trajectories (of the same system). Furthermore, if the 2-D RMSD map of the two systems were previously calculated, Simulaid can cluster the frames of each system, and plot a cross-RMSD map using the order of frames after clustering. With this approach, the similarity (or lack thereof) of the clusters in the two trajectories can be visualized. The extent of similarity of the clusters is also quantified on the output.

Distance Calculations

Residue Distance List. For selected residues (designated reference residues), Simulaid will calculate their distance from another selected set (designated neighbor residues). The distance between two residues is calculated in two ways: either (a) based on a (user selectable) representative atom (by default, the alpha carbon of a protein residue, the phosphorus of a nucleic acid reside and the oxygen of water; for unrecognized residue type the first atom) or (b) on the closest pair of atoms (i.e., the contact distance). Residue pairs within a user-specified threshold distance (different for the two methods of residue distance calculation) are listed on the output, which also lists the atoms of contact.

Track the PBC-Adjusted Distance of Pairs of Atoms. In a simulation run under the customary periodic boundary conditions, whenever a molecule leaves the simulation cell it is usually moved back to the simulation cell at the position that is the periodic image of the outside position. Thus, if the diffusion of a molecule is to be studied on a time scale permitting it to leave the cell, simple distance calculations will fail. Simulaid can track the distance between a selected pair of atoms and keep track of these jumps, and thus recover the actual time evolution of the distance.

Calculate the Distribution of the Distances Between Selected Atom Pairs. For a list of selected atom pairs, Simulaid can calculate the distribution of their distances. In addition, when the distances of interest involve several equivalent atoms (e.g., the hydrogens of a methyl group), Simulaid can calculate the shortest distance between selected groups of atoms, and generate the distribution based on that distance. Furthermore, the distances are also averaged with r^{-6} weighting. This can be of interest for comparison of simulation results with NMR data.

Residue Adjacency Matrix Analysis. Schuyler et al.[25] has shown that the sum of entries in the columns of successive powers of the adjacency matrix between residues will have peaks at residues with special properties. Simulaid can calculate the adjacency matrix between the residues (based on the distance both between the nearest atoms and between the representative atoms), calculate successive powers of this matrix, and plot the column sums of different powers as a function of the residue number.

Ramachandran[26] φ–ψ Angle Analysis

For the residues specified by the user, Simulaid will calculate the Ramachandran φ–ψ angles. For analysis of a single structure, Simulaid will prepare a Ramachandran map, color-coded (on the rainbow scale) according to the residue number. For analysis of trajectories, Simulaid will prepare (a) a Ramachandran map of the residues specified (or of all residues), in all frames analyzed, color coded according to the frame number, (b) dial plots of all φ and ψ angles analyzed, (c) for each residue analyzed, a trace of the evolution of the residue's conformation in the φ–ψ plane, and (d) for each residue analyzed the autocorrelation function of the φ–ψ values.

The calculated φ and ψ angles are also used to obtain the PROSS[27] secondary structure classification of each residue in each structure; and for trajectory analyses, Simulaid will provide statistics on the occurrence of each category for each residue analyzed.

Torsion Angle Dial Plots

Simulaid can generate dial plots for a number of user-specified torsions. If requested, the correlations between the torsions selected will also be calculated.

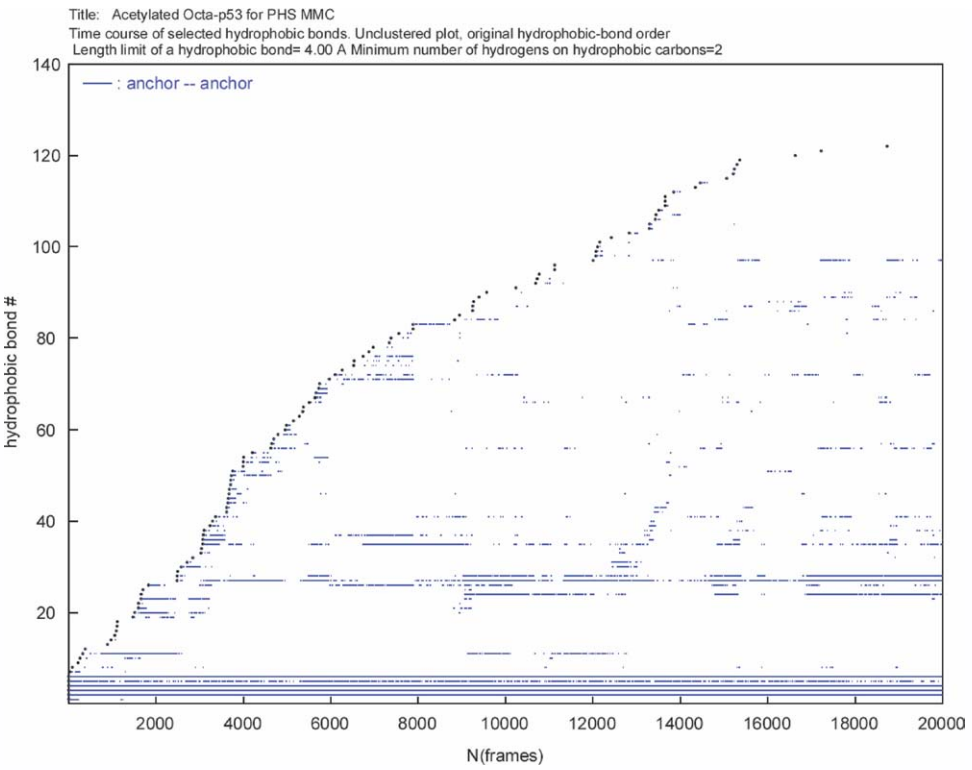

Figure 1. Track of hydrophobic bonds between all CH_2 and CH_3 groups. First occurrence of each a bond is marked by a bullet.

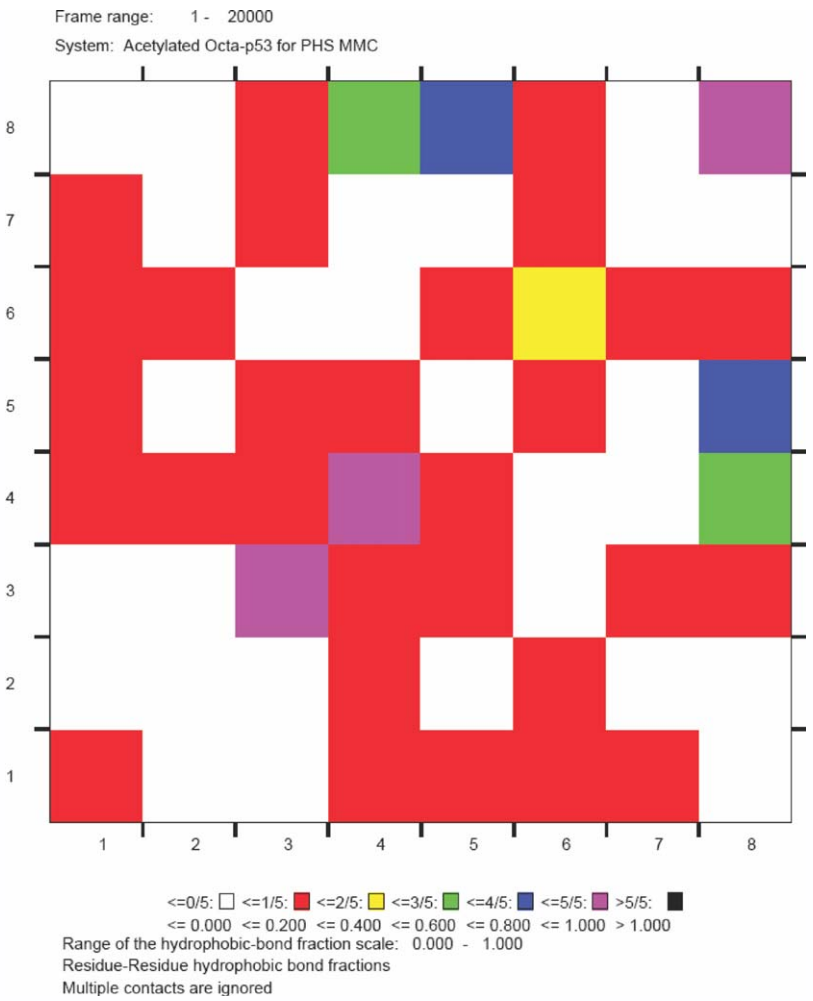

Figure 2. Residue–residue hydrophobic contact map. Both the rows and the columns are labeled by the residue number; the matrix element gives the percent of time the pair of residues represented by the matrix element are in hydrophobic contact.

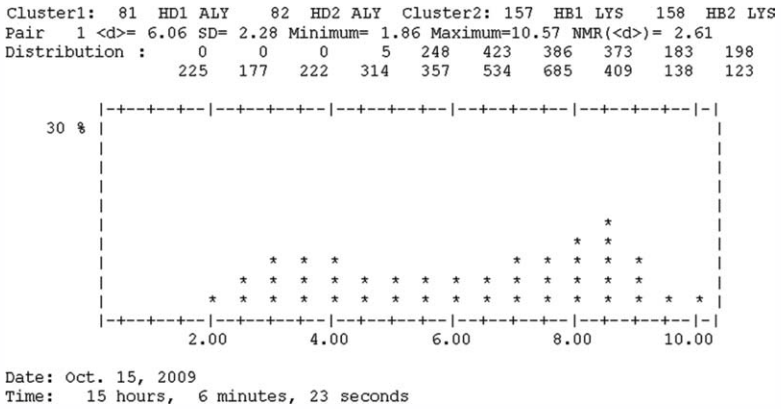

Figure 3. Histogram (distribution) of group distances between the beta hydrogens of AcLys and the delta hydrogens of the neighboring Lys. For each conformation the shortest H—H distance was used to prepare the distribution. The *x* axis is the distance scale and the *y* axis shows the percent of structures where the distance defined above falls in the range represented by each distance "bin."

Proline Kink[28] Analysis

Simulaid can calculate the proline kink parameters defined by Visiers et al.[28] for a selected proline and—for a trajectory—prepare a Postscript plot of their evolution as dial plots. As an enhancement to the original definition, Simulaid provides the option of first projecting the heavy atoms of the proline in question to the plane fitted to them. Because the calculation of the kink angles is sensitive to the position of the C_α atom of the proline, this option could reduce the noise introduced by fluctuations of the ring geometry. If requested, the correlations between the kink parameters will also be calculated.

Protein Helix Analysis (TRAJELIX)[2]

An extensive analysis of helix geometry has also been implemented. This analysis is based both on novel approaches for determining the extent of revolution, and the number of residues per turn, and on the standard analysis of helix length, position, orientation, and bend. The helix axis is calculated with the formulae of Kahn[29] using code provided by J.A. Christopher and T.O. Baldwin. If requested, the correlations between the various helix descriptors calculated will also be calculated.

Pseudorotation Angle[30,31] Calculation

Pseudorotation angles have been used previously to characterize ring puckers, using either a definition of pseudorotation angle specific for five-membered rings (sugars)[30] or a generalized version applicable to rings of arbitrary number of members.[31] The calculation gives both the phase and the amplitude as well as the mean deviation of the atoms from the plane fitted to the ring. Simulaid can perform the calculation using either definition.

DSSP[3] Secondary Structure Calculation

The Kabsch-Sander DSSP secondary structure assignment by Simulaid is based on an extended dictionary. The extensions to the directory consist of (a) differentiating sheets that form only a single pair from sheets that form two pairs, and for the latter specifying the types of both pairing; and (b) implementing the recognition of λ helix[32] where the hydrogen bonds form in the reverse direction compared with the standard α helix.

Circular Variance Map Calculation

It has been shown[1] that in the matrix of circular variances calculated for each residue pair in a protein with respect to the residues between (in the sequence) this pair, the columns corresponding to residues that link two separate domains will have low-circular variance values. Simulaid can calculate and plot this matrix and will attempt to find such columns in the matrix.

Residue Correlation Matrix Calculation

Simulaid can calculate the correlation and covariance matrices of a selected range of residues. The calculation is based on a representative atom of the residue as defined in Residue Distance List Section. Furthermore, the user can request the calculation of the eigenvalues and eigenvectors (normal modes) of the covariance matrix calculated.

Summary of the Energy Decomposition Data Generated by Amber

Holger Gohlke has written a set of Pearl scripts that runs MM-PBSA calculations of a ligand–protein complex, and generates huge tables containing various contributions to interactions between atoms of residue pairs or between atoms of residues and proteins or ligands.[33] This script is part of the Amber package. Simulaid can extract selected properties for selected residues, yielding tables of manageable size that are far easier to evaluate. Furthermore, for all energy terms within a residue the minimum, maximum, and average values are also calculated.

Results

This section describes selected examples of plots that Simulaid can produce. The analyses were based on a Monte Carlo simulation of the p53 octapeptide with acetylated lysine: Arg-His-Lys-

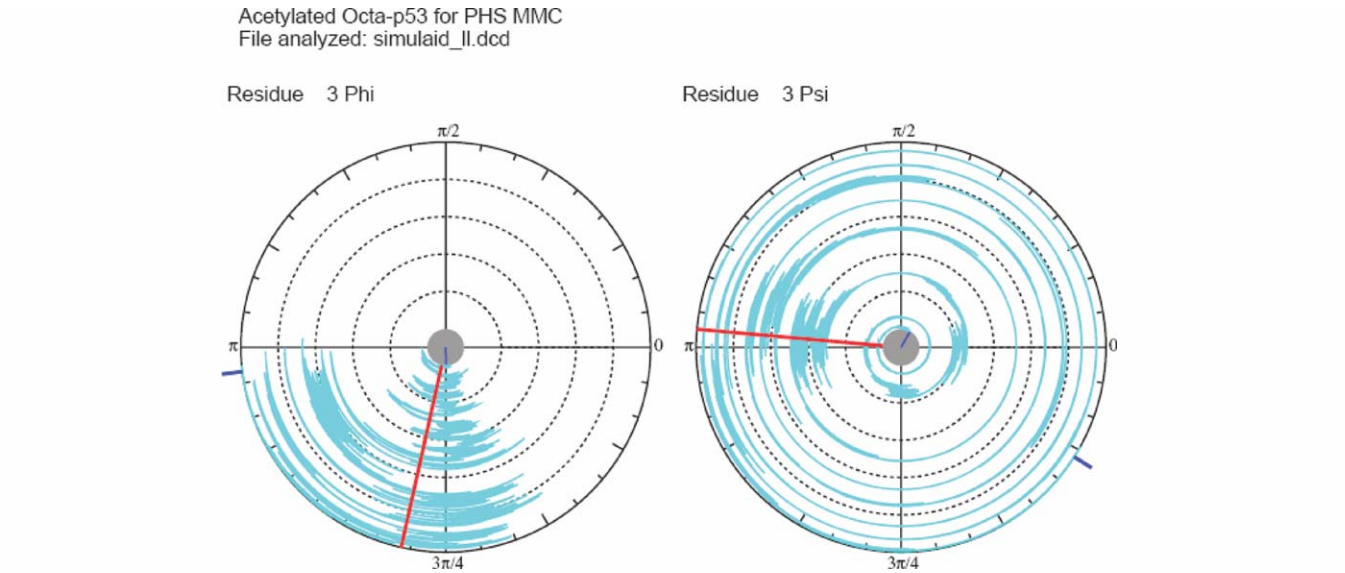

Figure 4. Dial plot of the φ and ψ angles for the acetylated lysine residue.

AcLys-Leu-Met-Phe-Lys. This peptide was also the subject of an extensive molecular dynamics study[34] that relied heavily on Simulaid. The Monte Carlo simulation used the Primary Hydration Shell[35] method, the peptide conformational sampling was done in the torsion space, with the stepsizes tuned to 0.3 acceptance rate using a recently developed technique.[36] The examples

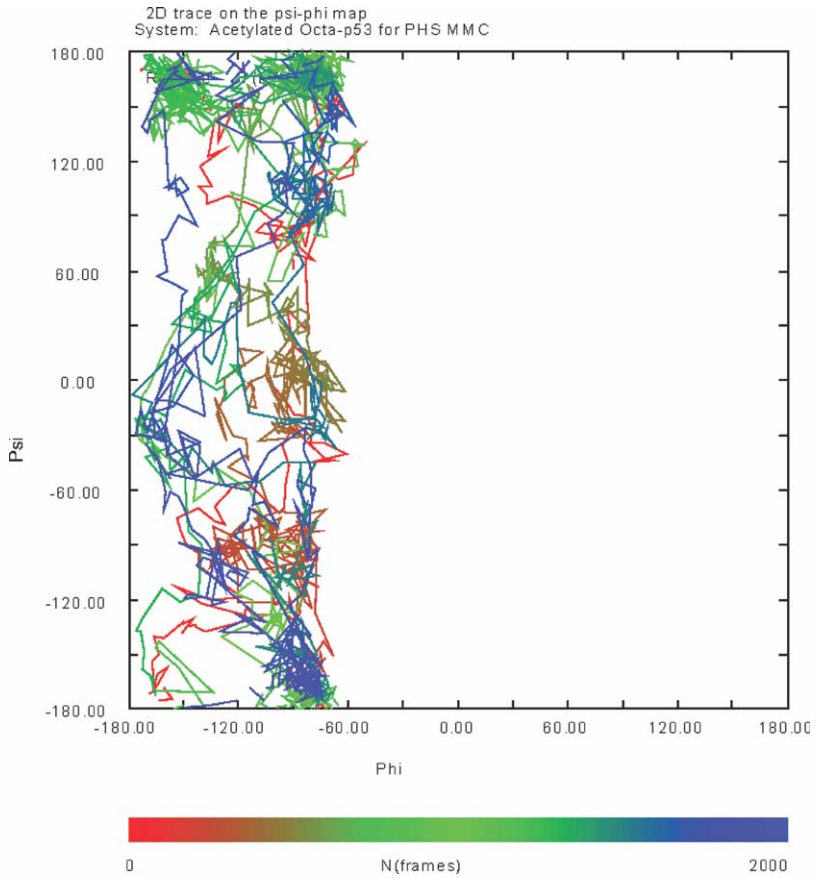

Figure 5. Conformation track of the acetylated lysine residue in the φ–ψ plane, corresponding to the dial plots of Figure 4. The track is color-coded on the rainbow scale according to the simulation time.

were chosen to illustrate the many ways Simulaid can be used to characterize conformational changes.

Figure 1 shows the track of hydrophobic bonds (between all CH_2 and CH_3 groups, with C—C distance below 4 Å) during the simulation of the octapeptide. Figure 2 shows the map of the residue–residue contacts based on these hydrophobic bonds. Figure 3 shows details of one particular bond: the distribution of the distance between the hydrogens of two $>CH_2$ groups, one on the delta carbon of the acetylated lysine and the other on the beta carbon of the neighboring lysine. The changes in the backbone conformation of residue 3 (a lysine) are represented both with dial plots of the φ and ψ angles (see Fig. 4) and the track of the conformation in the φ–ψ plane (see Fig. 5).

There are several examples of plots prepared by Simulaid that have been published and thus are not shown here. Our study of p53-CBP interactions[34] presents examples of 2D- and cross-RMSD plots, sorted by clusters, as well as cluster membership plots (see Protein Helix Analysis Section). Our paper presenting the helix analysis module TRAJELIX (see Protein Helix Analysis (TRAJELIX) Section) contains a complete set of plots generated from the analysis of rhodopsin simulations.[2] Our paper introducing circular variance as a tool for macromolecular structure analysis[1] shows a slice of a solvated protein, color coded with the atoms' circular variance with respect to the protein (see Circular Variance Calculation Section) demonstrating how useful the circular variance is for characterizing the position of the atoms with respect to the rest of the molecule the atom is in. The paper also shows circular variance maps (see Atomic Property Analyses Section) prepared by Simulaid, demonstrating the ability of such maps to detect domain separation.

Conclusions

Simulaid is a versatile program, with several unique features, that facilitates many of the tasks involved in setting up a molecular simulation and analyzing its outcome. The program is available at http://inka.mssm.edu/~mezei/simulaid (backup site: http://atlas.physbio.mssm.edu/~mezei/simulaid). The distribution includes the source code, documentation, data files, and executables for selected platforms. It is free for academic users; commercial users are charged a nominal fee for a perpetual license, including updates.

Acknowledgments

Although the development of Simulaid benefited from extensive feedback from its many users who are too numerous to list, the author is particularly grateful for general feedback, bug reports, and suggestions for additional features from members of the laboratories of Drs. R. Osman (Mount Sinai School of Medicine, NY), N. Pastor (Universidad Autonoma del Estado Morelos, Cuernavaca), P. Reggio (University of North Carolina at Greensboro), and H. Weinstein (Weil Medical College of Cornell University, NY). The author thanks Dr. M. McCallum for bringing to his attention the existence of λ-helix and Dr. C Bancroft for a careful reading of the manuscript.

Appendix: Analyzing a Charmm/NAMD Generated Trajectory in Amber

To convert a Charmm DCD trajectory into an Amber trajectory that can be interpreted fully with the Amber suite, several obstacles have to be overcome. The difficulty arises because the information in the topology and parameter files used in Charmm is contained in differently organized and formatted files in Amber, and not every feature in the Charmm force field has an obvious counterpart in Amber.

The solution for this problem calls for defining the system independently in Amber (e.g., using Leap), and then establishing the correspondence between the atom names in the original system and in the newly defined systems. Then the file format conversion can be executed in such a way that the atoms are rearranged in the converted trajectory to the newly established order. Simulaid can establish the correspondence and make the conversion, including the rearrangement, but there are several steps on the way that the user has to perform 'manually'.

The first step is to generate the **.top** and **.pdb** files in Leap. Simulaid can provide a list of residues to help in crating these files. If there are nonstandard residues in the system, the Amber-formatted residue file must be generated for that residue. This can also be accomplished with Leap. Also, the program Intocham (URL: http://inka.mssm.edu/~mezei/intocham) can generate this file from an InsightII **.car** file.

The **.pdb** file written by Leap will be a 'regular' PDB file (i.e., containing atom names of the form 1HA2 instead of HA12, and atom names like CA will have a space before them). This file must undergo a conversion by Simulaid to 'undo regularization' - an option of atom and residue name conversions.

In addition, the Charmm.**pdb** or **.CRD** file has to be modified to reflect the fact that the terminal groups are considered separate residues in Amber. This involves manually changing the residue numbers in the Charmm coordinate file, which can be followed by a cleanup with Simulaid to keep the residue numbers consecutive.

Probably the hardest part is making sure that all atoms are successfully matched. The matching of atom and residue names is controlled by the file **pdb_nam.dat**, that is part of the Simulaid distribution. When Simulaid runs it looks for this file in the current directory, and only if it does not find it there does it go to the distribution directory. Since the above file is likely to be in need of modification/extension, it should be copied into the working directory.

The conversion itself can be invoked as a trajectory file and type conversion: convert to Amber trajectory. During the quiz Simulaid will ask the user if rearrangement to a target order is required. If the answer is "yes", Simulaid will ask for a structure file with the target order - the user should then specify the Amber **.pdb** file that resulted after the 'undo regularization' operation. Simulaid will then ask for the name conventions for the input and target system; specify Charmm and Amber, respectively. Simulaid will then proceed with establishing the atom matches. It is likely that some atoms - especially those around the terminal groups- will not be matched. The list of matches is written on a separate file where the user can see the atoms that Simulaid could not match. Since this file is in the for-

mat that Simulaid needs to establish the match from input (an alternative to establishing the match based on two structure files), the user can edit this file to directly establish the full match. Alternatively, the user can modify the file **pdb_nam.dat** based on the list of missing matches and repeat the run. If all atoms are matched successfully the conversion will proceed and the converted trajectory should correspond to the **.top** file generated by Leap.

References

1. Mezei, M. J Mol Graphics Model 2003, 21, 463.
2. Mezei, M.; Filizola, M. J Comput-Aided Mol Des 2006, 20, 97.
3. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577.
4. Ravishanker, G.; Swaminathan, S.; Beveridge, D. L.; Lavery, R.; Sklenar, H. J Biomol Struct Dyn 1989, 6, 669.
5. Gilson, M. K.; Sharp, K.; Honig, B. J Comput Chem 1988, 9, 327.
6. Briggs, J. M.; Madura, J. D.; Davis, M. E.; Gilson, M. K.; Antosewicz, J.; Luty, B. A.; Wade, R. C.; Bagheri, B.; Ilin, A.; Tan, R. C.; McCammon, J. A. UHBD: University of Houston Brownian dynamics program. http://adrik.bchs.uh.edu/uhbd.html (last accessed: 3/29/2010).
7. Case, D. A.; Pearlman, D. A.; Caldwell, J. W., Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Cheng, A.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A. AMBER: Assisted model building with energy refinement. http://amber.scripps.edu (last accessed: 3/29/2010).
8. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian-03, 2004.
9. Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Li, J.; Zhu, T.; Thompson, J. D.; Winget, P.; Lynch, B. J. AMSOL-Version 7.1 2004.
10. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
11. Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. J Phys Chem 1996, 100, 19357.
12. Mezei, M. J Comput Chem 1997, 18, 812.
13. Qian, X.; Strahs, D.; Schlick, T. J Comput Chem 2001, 22, 1843.
14. Gilson, M.; Sharp, K.; Honig, B. J Comput Chem 1988, 9, 327.
15. Mezei, M.Information Newsletter for Computer Simulation of Condensed Phases, CCP5, Daresbury Lab No. 47, 2000.
16. Mezei, M.Information Quarterly, CCP5, Daresbury Lab No. 37, 1993; pp. 38–40.
17. Downs, G. M.; Barnard, J. M. Rev Comput Chem 2002, 18, 1.
18. Cui, M.; Mezei, M.; Osman, R. J Comput-Aided Mol Des 2008, 22, 553.
19. Kyte, J.; Doolittle, R. F. J Mol Biol 1988, 157, 105.
20. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.
21. Wimley, W. C.; Creamer, T. P.; White, S. H. Biochemistry 1996, 35, 5109.
22. Palliser, C. C.; Parry, D. A. D. Proteins: Struct Funct Genet 2000, 42, 243.
23. Mardia, K. V.; Jupp, P. E. Directional Statistics; Wiley, Chichester, 2000.
24. Kabsch, W. Acta Crystallogr 1976, A32, 922.
25. Schuyler, A. D.; Carlson, H. A.; Feldman, E. L. J Phys Chem B 2009, 113, 6613.
26. Ramachandran, G. N.; Sasiskharan, V. Adv Protein Chem 1968, 23, 283.
27. Gong, H.; Isom, D. G.; Srinivasan, R.; Rose, G. D. J Mol Biol 2002, 327, 1149.
28. Visiers, I.; Braunheim, B. B.; Weinstein, H. Protein Eng 2000, 13, 603.
29. Kahn, P. C. Comput Chem 1989, 13, 185.
30. Altona, C.; Sundaralingam, M. J Am Chem Soc 1972, 94, 8205.
31. Cremer, D.; Pople, J. A. J Am Chem Soc 1975, 97, 1354.
32. Son, H. S.; Hong, B. H.; Lee, C.-W.; Yun, S.; Kim, K. S. J Am Chem Soc 2001, 123, 514.
33. Gohlke, H.; Case, D. A. J Comput Chem 2004, 25, 238.
34. Eichenbaum, K. D.; Rodriguez, Y.; Mezei, M.; Osman, R. Proteins: Struct Funct Genet 2010 78, 447.
35. Kentsis, A.; Mezei, M.; Osman, R. Biophys J 2003, 84, 805.
36. Banfelder, J. R.; Speidel, J. A.; Mezei, M. Algorithms 2009, 2, 215.