Software News and Updates

tCONCOORD-GUI: Visually Supported Conformational Sampling of Bioactive Molecules

DANIEL SEELIGER, BERT L. DE GROOT
Computational Biomolecular Dynamics Group, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Received 20 May 2008; Revised 15 August 2008; Accepted 22 August 2008
DOI 10.1002/jcc.21127
Published online 22 October 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Conformational flexibility of bioactive molecules poses a major challenge to computational biology. tCONCOORD generates structure ensembles based on geometrical considerations and has been successfully applied to predict protein conformational flexibility and essential degrees of freedom. We have now developed a graphical user interface (GUI) for tCONCOORD, which substantially facilitates the simulation setup and provides valuable insights into the structure analysis and constraint definition process in tCONCOORD. Moreover, users can influence the constraint definition process by interactively turning interactions on and off, defining completely rigid or flexible regions, or by applying artifical constraints that cause a biased sampling of the conformational space. This interface offers a versatile environment not only for the setup and analysis of tCONCOORD simulations, but also for molecular modeling and structure analysis in general. Both tCONCOORD* and the tCONCOORD-GUI† are distributed freely.

© 2008 Wiley Periodicals, Inc. J Comput Chem 30: 1160–1166, 2009

Key words: CONCOORD; tCONCOORD; protein flexibility; PyMOL; Python; sampling; molecular dynamics; docking; induced fit; SBDD

Introduction

The prediction of conformational flexibility of bioactive molecules is important for many fields of computational biology and chemistry. Protein function is often coupled to motion ranging from small side-chain movement to reorientations of entire domains. Especially, the binding process of a protein to a ligand or another protein is often associated with significant conformational changes.[1] X-ray crystallography, which is still the major source of structural information of proteins, provides mainly static pictures of a single conformation, even though a number of proteins has been resolved in different conformations, providing insights into protein flexibility directly from experimental data.[2] Structures resolved by NMR-spectroscopy are usually published as an ensemble of conformations that fulfil the experimentally determined restraints and provide more information about protein flexibility. However, the method is still restricted to proteins of limited size.

Because of the difficulties associated with derivation of information about protein flexibility from experiments, many computational approaches have been developed and successfully applied. The CONCOORD method[3] uses a geometry-based approach to predict conformational flexibility from a given input structure, and the recently developed tCONCOORD[4] (t stands for transition) has been successfully applied to predict large conformational transitions of proteins. In this work, we present a newly developed graphical user interface (GUI) for tCONCOORD that substantially facilitates the setup of complex simulations and provides graphical support for tCONCOORD's protein structure analysis. In particular, it allows an interactive, user-defined selection of structure regions that should be treated as fixed or flexible during simulation (for e.g., loop (re)modeling) and manual constraint definition or manipulation (in e.g., a binding pocket or for the structure generation toward a known target). The GUI has been developed in the Python programming language and can be easily embedded as a plugin into the PyMOL Molecular Graphics System.[5] Currently, it is particularly developed as a PyMOL extension and, hence, cannot be straightforwardly integrated into another molecular viewer. However, a Python module is available that provides parsers, manipulating and writing functionalities for all file formats used by tCONCOORD and thus, may serve as an interface to other applications. In the rest of the article, we describe how the GUI is useful for different molecular modeling tasks. We show how various simulation parameters can be changed and a sampling of a small organic compound can be carried out. We

Correspondence to: B. L. de Groot; e-mail: bgroot@gwdg.de

Contract/grant sponsor: Deutsche Forschungsgemeinschaft (DFG); contract/grant number: GR 207914

*http://www.mpibpc.mpg.de/groups/de_groot/dseelig/tconcoord.html

†http://www.mpibpc.mpg.de/groups/de_groot/dseelig/cncplugin.html

© 2008 Wiley Periodicals, Inc.

furthermore show how the tCONCOORD-GUI assists understanding of interactions in protein structures and how interactions can be interactively switched on and off. We finally present two examples of more complex modeling tasks. In one example, we focus on the enhanced sampling of a substructure, in that case the generation of alternate loop conformation in a protein binding site while the rest of the protein structure remains unchanged. In the second example, we show how artificial constraints can be used to force a simulation to sample closed conformations when starting from an open (apo) conformation.

Simulation Setup

tCONCOORD uses a two-step approach to generate structure ensembles. In a first step, a single input structure is rigorously analyzed and translated into a set of geometrical constraints. This set consists of topological constraints (e.g., bonds, angles, planarities) and noncovalent constraints such as hydrogen bonds, salt bridges, and hydrophobic clusters. For the constraint definition process, tCONCOORD relies on various parameters such as upper and lower bounds for distance constraints and an optimized set of atomic radii.[6] Depending on the particular purpose, simulation parameters can be adjusted and parts of the constraint definition process interactively turned on and off (see Fig. 1A). "Interactively" here means that the influence of changing certain parameters, e.g., the hydrogen bond desolvation threshold, is immediately visualized such that hydrogen bonds that are affected are displayed in a different color and listed in a different table (see Fig. 2). Because the constraint definition process usually takes only a few seconds, it can also be run several times with different parameters and the results be displayed with the GUI.

If all interaction flags are switched off, only bonds, angles, planarities, and chirality constraints are defined. Generating structures from a protein chain with these settings yields an ensemble of random coil structures. However, the utility of this flexibility in the constraint definition process becomes more evident when applied to small organic molecules.

Conformational ensembles of organic compounds (Fig. 1B) are needed to determine whether an accessible conformation matches an existing pharmacophore model and, hence, complies with the minimum prerequisite of a potential inhibitor. Because the bioactive conformations of ligands often lie several kcal/mol higher than their respective global minima[7] and tend to be more extended,[8] the definition of noncovalent intramolecular interactions for organic compounds is usually not recommended and serves as an example where the user might want to switch off constraints in the tCONCOORD-GUI to prevent the exclusion of parts of the conformational space.

For the conformational sampling of macromolecules, e.g., proteins, tCONCOORD provides multiple options that determine the sampling properties of the generated ensemble, including the definition of position constraints, flexible parts, exclusions and manually added constraints, which can be easily controlled by the GUI. The rest of the article describes selected application examples.

Structure Analysis

The three-dimensional structure of proteins is determined by many interactions, such as covalent bonds, hydrogen bonds, and the hydrophobic effect. Although during most conformational transitions the local structure remains intact, it was found that conformational changes of proteins are often associated with opening of one or more hydrogen bonds. Because detection of these unstable hydrogen bonds is essential for protein flexibility prediction, we have developed a method to identify such unstable hydrogen bonds based on their local environment.[4] The method rests on estimating the likelihood for the hydrogen bond to be attacked by water molecules. Irrespective of the geometry in the input structure, a hydrogen bond that can be attacked by water molecules is more likely to open than a hydrogen bond that is shielded by hydrophobic residues.[9–12] In the constraint definition process in tCONCOORD, we calculate a solvation score for each hydrogen bond. If this score exceeds a predefined threshold, the hydrogen bond is labeled as unstable and not included as a geometrical constraint. Thus, in the subsequently generated ensemble these hydrogen bonds are not necessarily conserved.

The tCONCOORD-GUI allows to load the results of the constraint definition process and to visualize the defined constraints (see Fig. 2). The structure is displayed in a PyMOL window controlled by the tCONCOORD-GUI. A protein chain is thereby represented as a simplified C_α model for clarity. Different kinds of interactions are represented as colored arrows. Moreover, each interaction is listed in a table and detailed information, e.g., hydrogen bond geometries or energies, is displayed in a text field. Hydrogen bonds that are not taken into account because of high solvation probabilities are also displayed. The threshold can be changed interactively to enable the user to thoroughly study all hydrogen bonds and to set up simulations with different thresholds.

Depending on the particular question, it might be useful to switch off interactions, e.g., to study the influence of a certain hydrogen bond on the conformational flexibility. The tCONCOORD-GUI therefore allows to interactively influence the constraint definition process by selectively enabling and disabling interactions.

Enhanced Sampling of Substructures

Structure-based drug design (SBDD) has been shown to be more effective when protein flexibility is taken into account.[13–17] However, in most cases not the flexibility of the whole protein, but rather, a part of the structure, e.g., the binding site, is of interest. tCONCOORD therefore provides the possibility to define position constraints for a group of atoms the coordinates of which remain unchanged in all generated structures. Moreover, a group of atoms can be defined for which only primary structure constraints are defined which yields extensive sampling of the conformational space. Both the definition of position constraints and of flexible parts are useful features for enhanced sampling of protein parts, e.g., a loop in a binding site.

tCONCOORD uses Gromacs[18,19] index files for defining position constraints and flexible residues. Because PyMOL provides a much more elaborate way of selecting atoms than the *make_ndx* program from the Gromacs suite, we developed an interface that allows to translate PyMOL selections into Gromacs index files and vice versa. This interface—which might also be of interest for the Gromacs community—for instance, provides an easy way of defining a

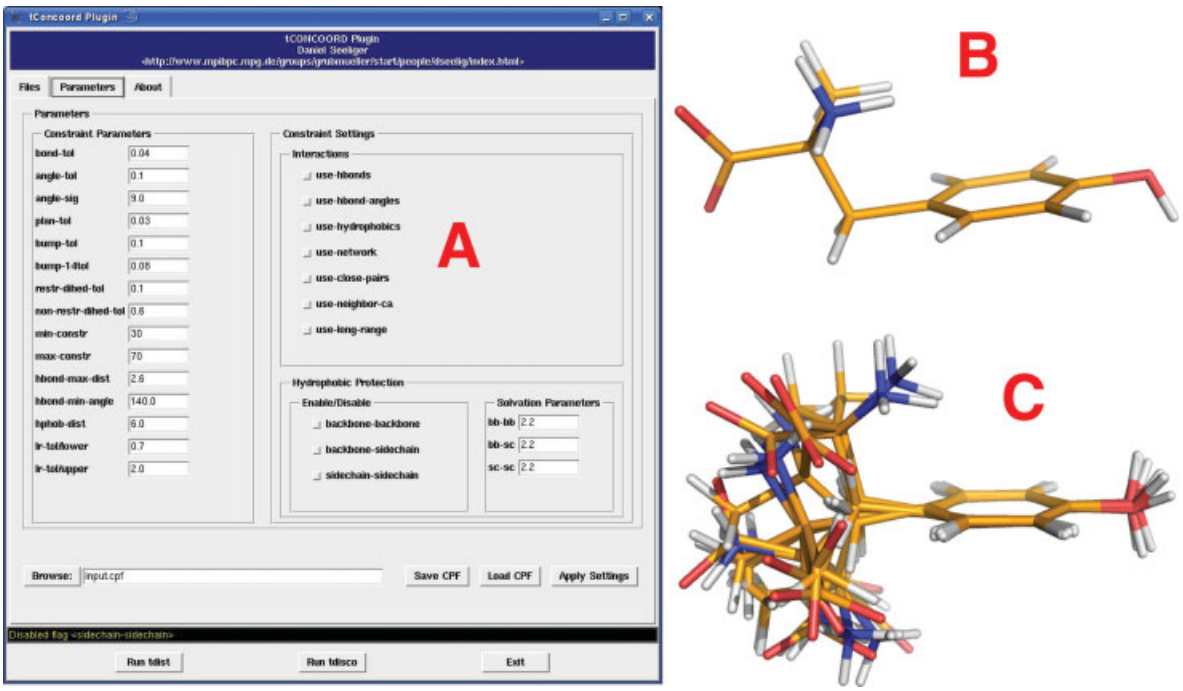

Figure 1. (A) Graphical user interface for defining input parameters and constraint definition options. (B) Organic compound used as input structure. (C) Structure ensemble generated by tCONCOORD.

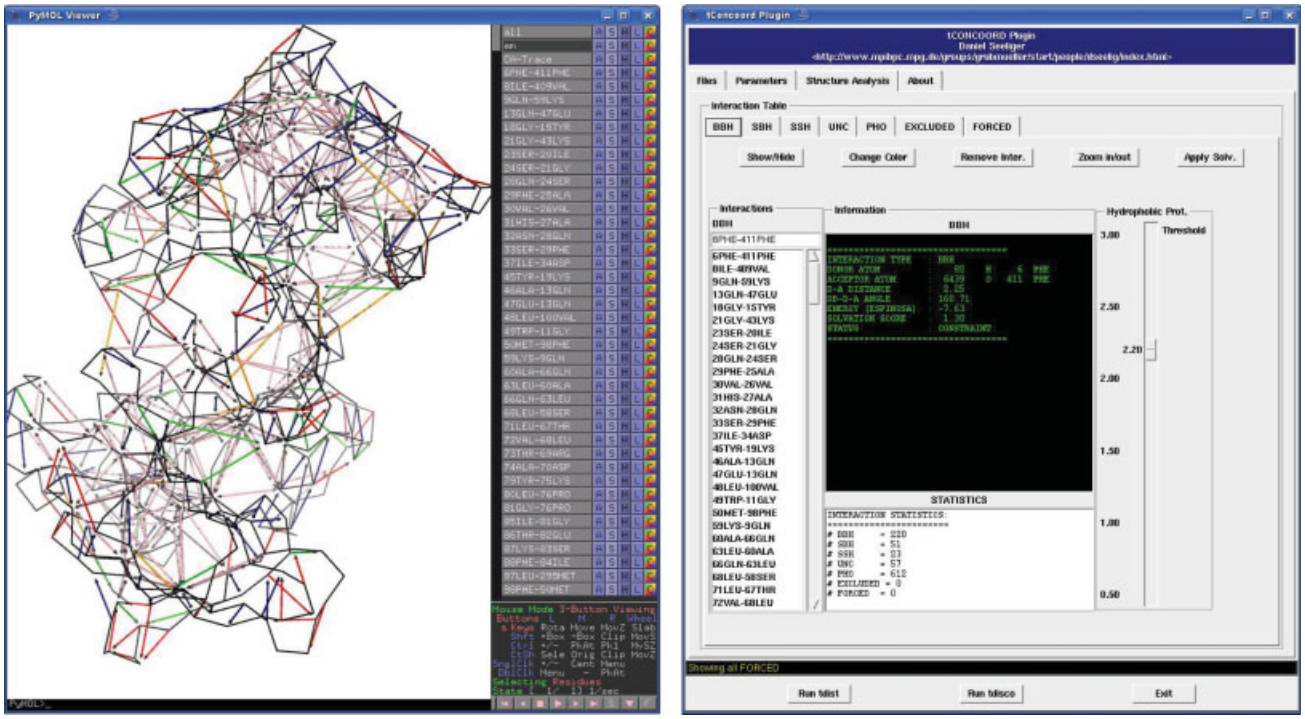

Figure 2. Structure analysis in tCONCOORD: Left: PyMOL window controlled by the tCONCOORD GUI. Interactions are represented by colored arrows connecting C_α-atoms of the protein structure. Right: tCONCOORD-GUI. Interactions are listed in different tables. Information about each interaction is displayed in a text field and each interaction can be removed by clicking on a button.

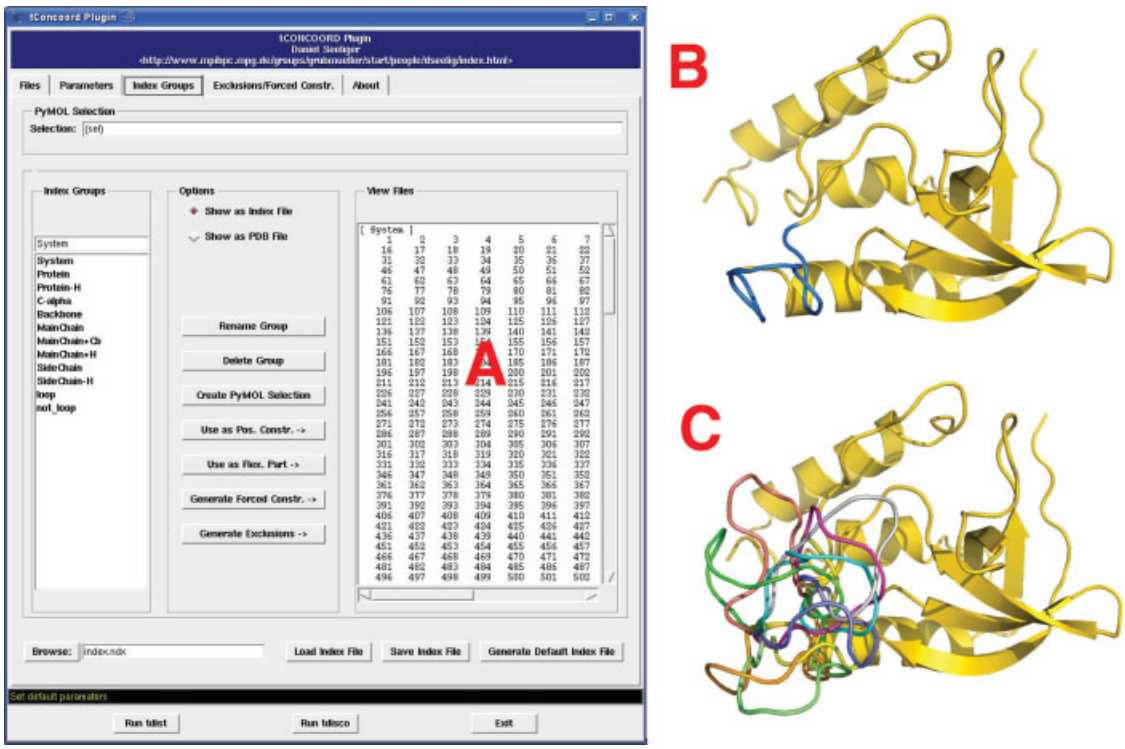

Figure 3. Group definition using PyMOL selections. (A) tCONCOORD-GUI with a loaded Gromacs index file. (B) Structure of staphylococcal nuclease (PDB 1ey4). (C) Loop ensemble generated with tCONCOORD.

binding site by simply selecting all residues around a certain distance to the ligand or by mouse-clicking.

As an example for a local sampling, we chose staphylococcal nuclease that has a flexible loop as part of its binding site (blue in Fig. 3B). We use the GUI to generate two index groups (Fig. 3A), one containing the loop atoms, and the other one containing the rest of the system. For the second group, which should be kept fixed, we define position constraints. The first group containing the loop atoms, we define as flexible to ensure exhaustive sampling. When all definitions are finished, the GUI writes all necessary input files and runs the constraint definition program with the appropriate command line options. Hence, the user does not need to have any detailed knowledge about running tCONCOORD from the shell; all calls to tCONCOORD programs are carried out by the GUI. The sampling itself, which can take several hours depending the system size on the number of structures to be generated, is started as a thread which means that the process continues when the GUI or PyMOL are closed. When the sampling is finished, the simulated ensemble can be directly loaded into the PyMOL viewer.

Together with the GUI, tCONCOORD can also be used to insert missing residues. Thereby, the missing sequence is translated to 3D coordinates and inserted at the appropriate position in the protein chain. Subsequently, an ensemble of structures is generated with sterically correct conformations of the missing sequence with respect to the resolved part of the protein. Thus, aided by the GUI, tCONCOORD can be used for both remodeling of structure fragments as well as for the de-novo modeling of missing residues as, e.g., loops.

Biased Sampling

Conformational sampling can be a time-consuming task, depending on the dimensionality of the conformational space and the conformational flexibility of the system. In tCONCOORD, we have therefore implemented the possibility to perform biased sampling simulations by imposing artificial constraints—either from knowldege or intuition or otherwise—that can be defined with graphical support. Figure 4 shows an application of a biased tCONCOORD sampling of archeoglobus fulgidus osmoprotection protein (PROX). Upon ligand binding PROX undergoes a large conformational change, which has been classified as hinge-bending [D-h-2] motion.[2] The interdomain angle changes as much as 55.1°[1] from the unbound state (Fig. 4A) to the ligand-bound state (Fig. 4B). The backbone RMSD between the two conformations is 5.0 Å.

If we take the apo (unbound, open) conformation as input for a tCONCOORD sampling we obtain, due to the pronounced conformational flexibility, an ensemble that covers a large conformational space, containing both open and closed conformations and a large variety of twisted conformations. The probability of finding a conformation with a backbone RMSD smaller than 2 Å to the holo (ligand-bound, closed) conformation is only 0.2%.

Therefore, if we are interested in in predicting holo-conformations based on the apo structure and if prior knowledge is available that the structure closes upon ligand binding, it is of interest to carry out a simulation with a directed bias toward sampling closed conformations with higher probability. To set up such a biased sampling, we can use the tCONCOORD-GUI to impose

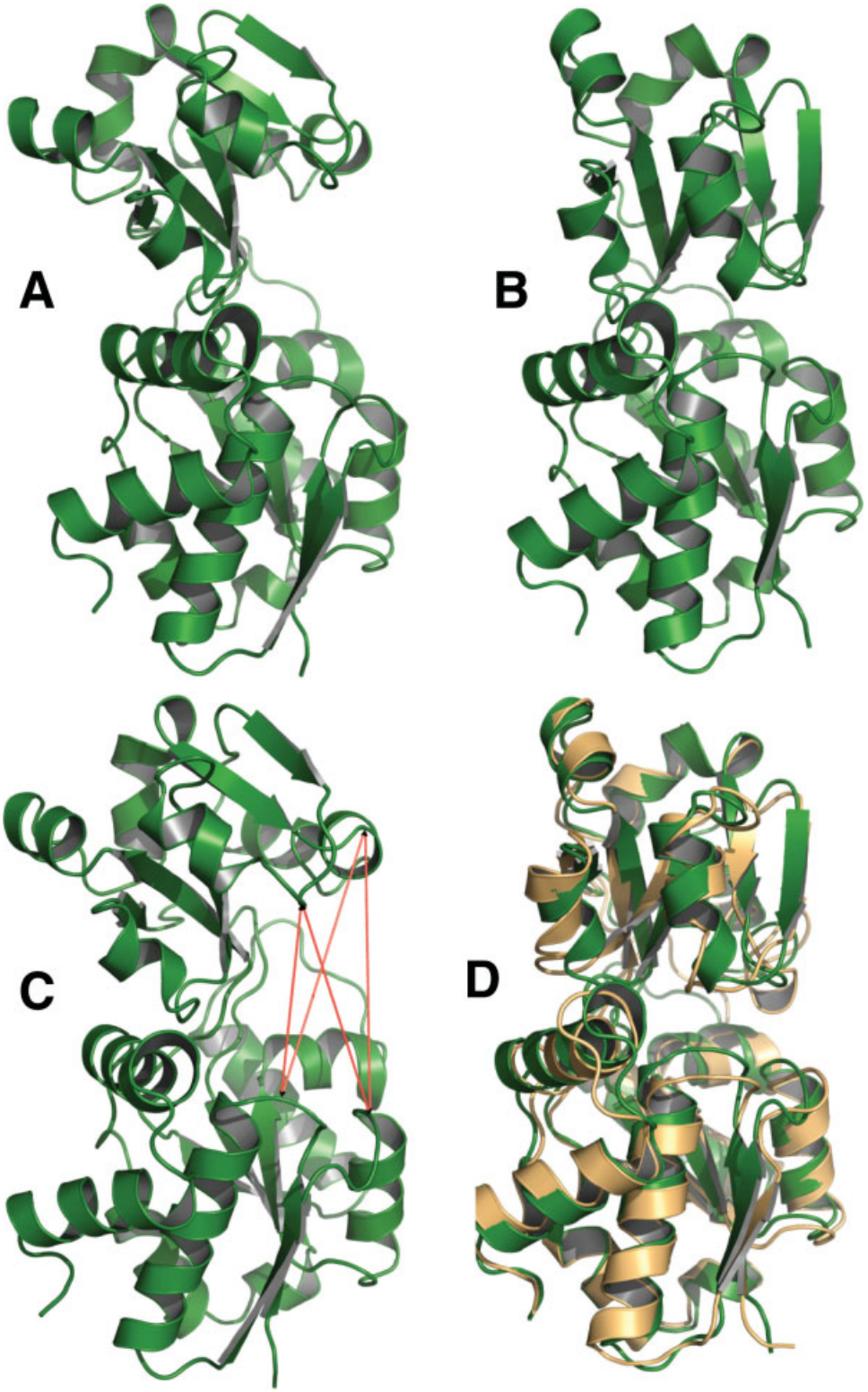

Figure 4. Biased sampling of osmoprotection protein (PROX). (A) apo structure (PDB 1sw5). (B) holo structure (PDB 1sw2). Upon ligand binding PROX undergoes a large conformation change (RMSD holo/apo: 5.0 Å). (C) Apo conformation used as input for biased tCONCOORD sampling with artificial constraints (red arrows). (D) Superposition of the holo conformation (green) and a tCONCOORD structure (orange) with 1.6 Å RMSD. Note that tCONCOORD generates atomistic structures; a cartoon representation is shown here for clarity.

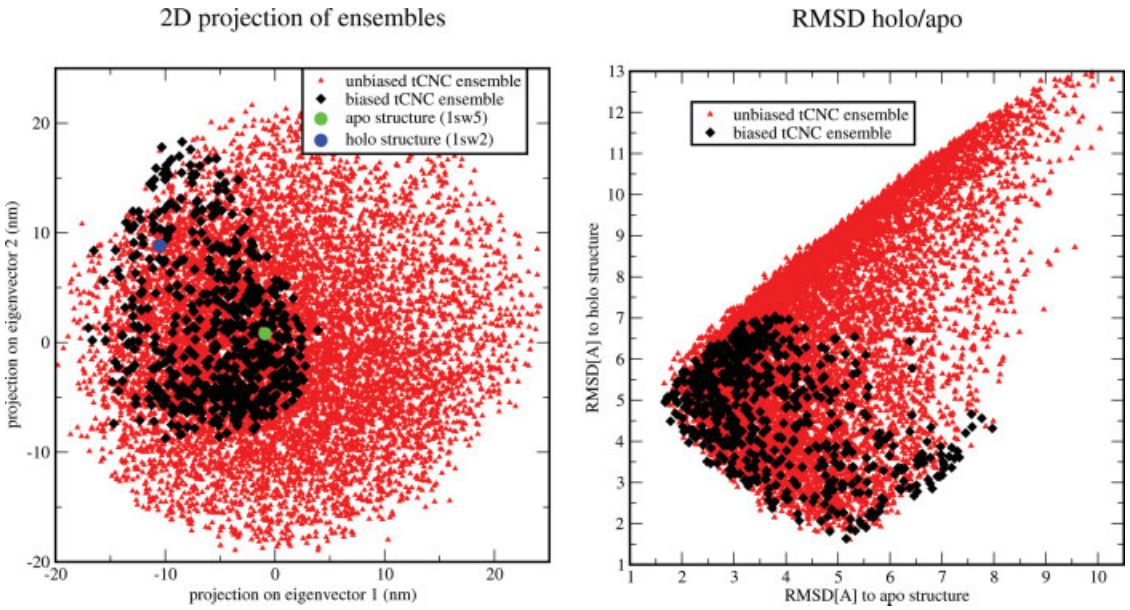

Figure 5. Comparison of sampling properties of a biased and an unbiased tCONCOORD ensemble. Left: Projection on the two eigenvectors with the largest eigenvalues of a principal components analysis (PCA) carried out on the combined ensemble. The biased ensemble (black) covers only a subset of the conformational space sampled by the unbiased ensemble (red). The projection of the X-ray structures of the apo (green) and holo conformation (blue) are also shown. Right: RMSD to the apo (x-axis) and holo (y-axis) structure. The probability of finding a conformation with an RMSD < 2 Å to the holo conformation increases 5-fold for the biased ensemble.

additional constraints on the system. In this example, we used two residues from each domain and defined constraints such that in the generated ensemble the distance between these residues is smaller than in the input structure. Here, the selection of suitable residues on which the artificial constraints are imposed was guided by prior knowledge of the holo conformation. Similarly, information about distances between certain residues could also be obtained from other experimental data, e.g., from FRET experiments, EPR, or NMR.

The definition of artificial constraints is straightforward. We use the GUI to select the two pairs of residues and define two index groups. Afterward, constraints between these groups can be defined and displayed as arrows in the PyMOL viewer (Fig. 4C).

The GUI writes the appropriate input files for the structure generation program, which can be either started directly from the GUI or from the command line.

The sampling properties of the resulting ensemble differs significantly from the unbiased ensemble. In the left graph of Figure 5, the projection on the two eigenvectors with the largest eigenvalues of a principal components analysis (PCA) carried out on the combined ensembles is shown. Each dot in this plot represents a single structure. The projection of the two X-ray structures are shown in green (apo) and blue (holo). The unbiased tCONCOORD ensemble is represented by red triangles, whereas the ensemble from the biased tCONCOORD run is shown in black. It can be seen that the biased ensemble only samples a subset of the conformational space covered by the unbiased ensemble. The plot on the right in Figure 5 shows the RMSD of each structure to the apo (x-axis) and the holo (y-axis) structure. Again, we see that the artificial constraints imposed in the biased sampling cause a shrinking of the available conformational space. Although the probability of finding a conformation with an RMSD lower than 2 Å to the holo conformation is only 0.2% in the unbiased ensemble, it increases 5-fold to 1.0% in the biased ensemble.

Conclusions

We have described a graphical user interface for tCONCOORD that allows better understanding and control of the constraint definition process and the structure analysis. We showed how tCONCOORD-GUI enables a wide range of applications ranging from generating ensembles of organic compounds, generation of loop conformations to a biased sampling of potential induced-fit conformations when starting from an apo structure. In addition, tCONCOORD-GUI substantially facilitates the setup of complex molecular modeling tasks through straightforward selection of atoms via PyMOL, automatic generation of input files and the graphical representation of the constraint definition process.

References

1. Brylinski, M.; Skolnick, J. Proteins Struct Funct Bioinformatics 2008, 70, 363.
2. Gerstein, M.; Krebs, W. Nucleic Acids Res 1998, 15, 4280.
3. de Groot, B. L.; van Aalten, D. M. F.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. C. Proteins Struct Funct Genetics 1997, 29, 240.

4. Seeliger, D.; Haas, J.; de Groot, B. L. Structure 2007, 15, 1482.
5. DeLano, W. L. The PyMOL Molecular Graphics System, 2002. Available at http://www.pymol.org.
6. Seeliger, D.; de Groot, B. L. Proteins Struct Funct Bioinformatics 2007, 68, 595.
7. Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. J Chem Inf Model 2005, 45, 422.
8. Diller, D. J.; Merz, K. M. Jr. J Comput-Aided Mol Des 2002, 16, 105.
9. Fernandez, A.; Colubri, A.; Berry, R. S. Phys A 2002, 307, 235.
10. Fernandez, A.; Sosnick, T. R.; Colubri, A. J Mol Biol 2002, 321, 659.
11. Fernandez, A.; Berry, S. Biophys J 2002, 83, 2475.
12. Fernandez, A.; Rogale, K.; Scott, R.; Scheraga, H. A. Proc Natl Acad Sci USA 2004, 101, 11640.
13. Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. J Mol Biol 1997, 266, 424.
14. Carlson, H. A. Curr Opin Chem Biol 2002, 6, 447.
15. Meagher, K. L.; Carlson, H. A. J Am Chem Soc 2004, 126, 13276.
16. McGovern, S. L.; Shoichet, B. K. J Med Chem 2003, 46, 2895.
17. Teague, S. J. Nature Rev Drug Discovery 2003, 2, 527.
18. Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. Comput Phys Commun 1995, 91, 43.
19. Lindahl, E.; Hess, B.; Van der Spoel, D. J Mol Model 2001;7:306; Available at http://www.gromacs.org.