Protein Inherent Structures by Different Minimization Strategies

FRANCESCO RAO

Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, Albertstr. 19, 79104 Freiburg, Germany

Received 5 August 2010; Accepted 6 September 2010
DOI 10.1002/jcc.21691
Published online 4 November 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Network-based methods provide an accurate description of the free-energy landscape of peptides and proteins sampled by molecular dynamics simulations. To that end, it is necessary to group the individual snapshots in a meaningful way. The inherent structures (ISs) provide an appropriate discretization of the trajectory into microstates, avoiding problems that can arise in clustering algorithms that have been used previously. In this work, different minimization protocols to obtain the IS of a peptide are investigated on the basis of cut-based free-energy profiles. It is found that a computationally more efficient quasi-Newtonian algorithm provides quantitative agreement to the classical conjugate gradient method in terms of the population of the peptide substates and the energy barriers separating them. That is, despite the fact that the two algorithms can occasionally quench a given peptide snapshot in different potential energy minima, the overall properties of the system are not affected. As reported by others, atom permutations affect the calculation of the IS, requiring an improved implementation of current potential energy functions.

© 2010 Wiley Periodicals, Inc.   J Comput Chem 32: 1113–1116, 2011

Key words: molecular dynamics; energy minimization; free-energy surface; complex networks; proteins

Molecular dynamics (MD) simulations provide essential information about the thermodynamics and kinetics of proteins. However, conventional methods to analyze the simulated trajectories project the free energy surface onto one or two order parameters, providing an inaccurate characterization of protein dynamics. For this reason, a completely new arsenal of tools inspired by network theory was recently proposed to overcome some of those limitations and artifacts.[1,2] A useful approach maps the protein trajectory onto a conformation-space network, whose nodes represent the different microstates and whose links correspond to direct transitions between them observed during the simulation. This method has been successfully applied to the study of peptide folding and structural transitions[1–9] as well as to interpret electron transfer experiments[10] and time-resolved IR measurements.[11,12] Other equivalent formulations that are similar in spirit have been proposed.[13–15]

Graph-theoretical methods rely on a partition of the sampled protein conformations into a discrete set of microstates. Consequently, the free-energy landscape is investigated in terms of a transition matrix formalism. As a matter of fact, finding a "good" definition of a protein microstate can be hard and system dependent. In an effort to make network-based analysis more natural and robust, the inherent structures (ISs)[16] were calculated and used as a set of microstates to describe protein dynamics.[8,9] The IS are defined as the local minima of the potential energy surface and are obtained by minimizing each snapshot of a previously generated MD trajectory. It has been found that this formalism does provide a natural and physically meaningful discretization of a protein trajectory, solving some of the problems arising in conventional methods for structural clustering.[8,17] Following earlier work in the atomic clusters community, Stillinger and Weber used the superposition approach[18] (and ref. therein) to construct global partition functions for condensed matter systems, and introduced the term "inherent structure" to refer to a local minimum.[16] This framework has now been used in global optimization studies of peptides and proteins,[19–21] master-equation descriptions,[22,23] and recently for the analysis of MD simulations.[8,9,24–26]

Several strategies exist to find the minima of the potential energy surface. The methods that require up to first derivatives of the energy with respect to the Cartesian coordinates are mainly the steepest-descent (SD) and the conjugate-gradient (CG) methods.[27] Making use of second derivatives, a significantly more efficient class of

Correspondence to: F. Rao; e-mail: francesco.rao@frias.uni-freiburg.de

Contract/grant sponsors: Excellence Initiative of the German Federal and State Governments and an European Molecular Biology Organization Fellowship.

algorithms is represented by quasi-Newton methods where gradient information from successive iterations is used to build an approximate Hessian.[27]

How the application of these different approaches to calculate the IS influences our understanding of the system thermodynamics and kinetics is still unclear. For a few atom system, it was observed that quasi-Newtonian minimization paths will not differ significantly from those of a SD approach.[28] Stillinger and coworkers[29] explored this point in further detail for a Lennard–Jones liquid and concluded that, although average properties are conserved, on rough energy landscapes the two types of algorithm quench the system in different minima.

Here, I apply different minimization strategies for the generation of the IS of a $(\mathrm{GlySer})_2$ peptide and then compare them on the basis of their free-energy landscape obtained by complex network analysis. Given that the IS representation is relevant for understanding protein dynamics, it is useful to address this issue.

The simplest minimization algorithm of practical use in classical macromolecular simulations is SD. In this method, the coordinates are adjusted in the negative direction of the gradient by using an iterative procedure. SD ensures the relaxation to the closest potential energy minimum in conformation space. However, it is known to have convergence problems in many practical cases. The CG algorithm is an improved SD approach. In this case, the displacement is computed from the gradient at the current point plus the scaled previous displacement, providing convergence to the process. It is known that SD can actually be superior to CG when the starting molecular structure is some way from the minimum. However, CG is much better once the initial strain is removed.[27]

Second derivatives of the energy, indicating the curvature of the function, can be used to efficiently locate the position of the minimum. The Newton–Raphson (NR) method,[27] diagonalizing the second-derivative matrix (i.e., the Hessian), finds the optimum step size along each eigenvector to reach the minimum. This algorithm is computationally expensive and, in practice, cannot be efficiently applied to systems larger than a few tens of atoms. In this sense, quasi-Newtonian methods, like the adopted basis Newton–Raphson (ABNR,[30]) or the limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS)[31][*] methods are more interesting. The ABNR method performs energy minimization using a NR algorithm applied to a subspace of the coordinate vector spanned by the displacement coordinates of the last positions. Consequently, this method has small memory requirements and can be applied to large molecules.

Second-derivatives methods need a Hessian positive definite to obtain a stable minimization process. Far from the minimum this is not generally true and this class of algorithms may fail even though they work very well close to a minimum. Serial combination of the SD and ABNR algorithms (SDABNR) can be useful (as done in refs. 8 and 32), motivated by the consideration that the former quenches the system to the closest energy minimum and the latter assures convergence. This is particularly relevant when doing normal-mode analysis starting from an experimentally determined structure that might be away from the closest potential energy minimum of the force field.[32]

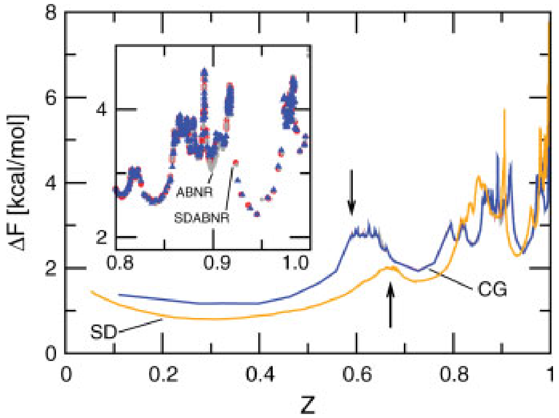

Figure 1. Cut-based free-energy profiles using the inherent structures calculated by a CG, ABNR, SDABNR, and SD algorithm are shown in blue, gray, red, and orange, respectively. In the inset, a detail of the profile is shown.

A MD simulation of the $(\mathrm{GlySer})_2$ peptide using the Langevin algorithm and the implicit solvation FACTS[33] is performed[†] and the IS are calculated by single derivatives as well as quasi-Newtonian methods (ABNR). GlySer peptides have been used for quite some time as flexible linkers for polypeptide dynamics and represent a good system to study conformational changes.[34,35] A trajectory of 280 ns at 300 K is obtained for a total of 10^6 conformations. A quantitative analysis of the peptide free-energy surface at a larger temperature was already presented by the author in ref. 9.

A set of calculated IS define an ensemble of microstates. The latter are then used in conjunction to graph-theoretical methods to characterize the peptide free-energy landscape,[8,9] in terms of valleys and the barriers separating them. To this aim, a cut-based free-energy profile analysis is performed, which consists in a flux analysis of the transition network with respect to the most populated peptide microstate.[36] In Figure 1, the cut-based free-energy profile of the $(\mathrm{GlySer})_2$ peptide is shown. The profiles obtained by calculating the IS with a CG, ABNR or SDABNR approach are very similar, indicating that the three methods are equivalent in terms of the stability of the substates and the barriers between them (blue, gray, and red symbols/lines in Fig. 1, respectively). This profile represents the free-energy surface projected on the cumulative partition function-based reaction coordinate Z. Cut-based profiles have proven to be a better approach compared with traditional free-energy profiles (where the landscape is projected onto one or more arbitrarily chosen order parameters[37,38]) because they provide a correct estimate of the height of the free-energy barriers, as well as their positions on the landscape.[17,36,39] The population of a valley is extracted from these profiles by looking at the span of the given valley along Z, whereas the kinetics of interconversion is provided

[*] The LBFGS method[31] is very similar in spirit with respect to ABNR but it is not yet implemented in CHARMM.

[†] MD simulations, using the Langevin algorithm with a friction coefficient equal to 0.6 ps^{-1}, were calculated with the CHARMM program,[30] the polar hydrogen energy function (PARAM19) was used. The effects of water have been included using the generalized born FACTS implicit solvation model.[33] SHAKE was employed so that an integration step of 2 fs could be used.

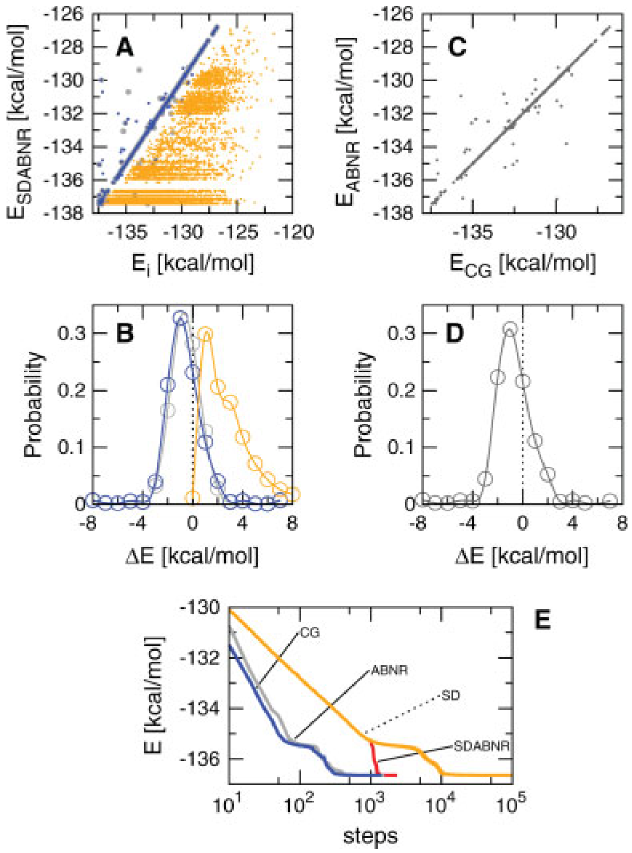

Figure 2. Inherent structure energies resulting from different methods. (A) Comparison between SDABNR (*y*-axis) and CG (*x*-axis, blue dots), ABNR (gray), and a partially converged SD (orange) approach. (B) Energy difference distribution between SDABNR and the above three methods. (C) Comparison between ABNR (*y*-axis) and CG (*x*-axis). (D) Energy difference distribution between ABNR and CG. (E) Example of energy minimization of a peptide snapshot as a function of steps for the aforementioned algorithms.

by the highest barrier between two valleys. For the $(GlySer)_2$ peptide, the largest populated valley has a population close to 60%, which interconverts to the second most populated valley crossing a barrier of less than 2 kcal/mol (these two valleys are defined by values of the cumulative partition function of $0 < Z < 0.58$ and $0.58 < Z < 0.77$, respectively). For larger values of Z at least two other valleys are detected for a cumulative population of roughly 20%. In this study, removing the initial strain by 1000 steps of SD using the SDABNR approach does not produce any substantial improvement with respect to ABNR alone, indicating that the system is, in average, close enough to the minimum. On the other hand, the free-energy profile obtained by running only 1000 steps of SD to obtain an approximate estimation of the IS provides a different picture (orange symbols/lines in the figure). The population of the largest valley is larger with respect to the other methods and the barrier to go from the first valley ($0 < Z < 0.68$) to the second one ($0.68 < Z < 0.87$) is smaller by more than 1 kcal/mol. Overall, these first results indicate that a second-derivative method is in quantitative agreement with CG, providing the same partitioning of the landscape into substates, with the correct populations and barrier heights.

Looking into more details the values of the energy obtained by the different minimization methods, it is found that they are generally very similar with few outliers. In Figure 2A, a comparison between the IS energies resulting from the different methods is shown. Energies obtained by SDABNR versus ABNR or CG are shown as gray and blue dots, respectively. In few cases, a given peptide snapshot is quenched in different minima but, in general, this is not the case (points on the diagonal). As expected, a SD calculation carried out for 1000 steps (orange points) shows poor convergence and larger energies. A distribution of the energy differences with respect to SDABNR is presented in Figure 2B. Both the distributions relative to the ABNR and CG algorithms have a peak at values smaller than zero, indicating that they, in average, quench to lower energy structures with respect to SDABNR. Moreover, when the CG and ABNR energies are compared one against the other (see Figs. 2C and 2D), ABNR energies are slightly lower. For a typical quenching of a peptide snapshot, the ABNR and CG methods converge with a similar number of steps as shown in Figure 2E. However, the calculation of an ABNR step is computationally more efficient with respect to all the other methods, providing an overall speedup of at least 1.4 (see Table 1 for details).

This analysis is carried out using a definition of the peptide IS based on a geometrical criterion. As originally introduced in refs. 8 and 9, minimized snapshots are considered to be in the same minimum if they have an all-atom root-mean-square-displacement smaller than 0.05 Å. This is a crude trick to overcome an approximated implementation of symmetry operations in the energy function of current force fields (mainly the improper angles term) and/or singularities in the solvation term (as in SASA[8]). This problem has extensively reported by the group of Wales.[19,40,41] As a consequence, artifacts appear when it comes to the calculation of the IS. The cut-based free-energy profile based on IS defined solely by the value of the energy (as originally prescribed by Stillinger and Weber) is shown in Figure 3. A small difference in energy of about 0.001 kcal/mol when doing a permutation of the terminal oxygen atoms results in an artificially repeated profile (i.e., the two main valleys are repeated again at $Z = 0.5$, dashed vertical line in the figure). The valleys marked by a star and by a triangle are structurally identical representing a permutation of the two terminal oxygen atoms (see figure inset). The population of the single valleys is essentially half with respect to the profile shown in Figure 1 (e.g., 30% for each valley marked with a star comparing to 60% of the correct profile shown in Fig. 1). This is a further confirmation that a more rigorous symmetrization of the current potential energy function is needed when it comes to the calculation of the IS. A correction was recently proposed[41] and, hopefully, will be introduced in the next versions of the CHARMM and AMBER force fields.

Table 1. Computational Cost for the Different Protocols.

Protocol	Time (s)	Ratio
ABNR	80.7	1
CG	114.6	1.4
SDABNR	147.3	1.8

Times are calculated on 100 random minimizations along the MD trajectory.

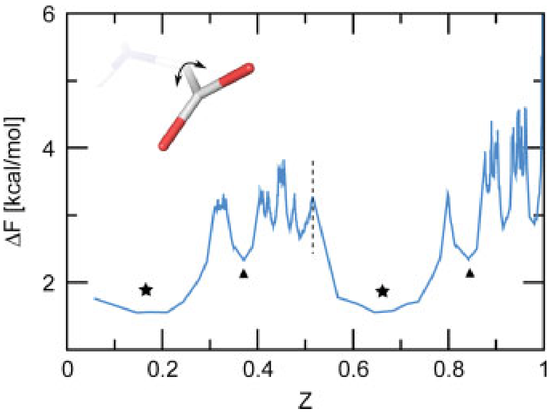

Figure 3. Cut-based free-energy profile when the inherent structures are defined by the value of the energy only. Permutation of the two terminal oxygen atoms (inset) results in an artificially doubled profile (the valleys marked by a star and by a triangle are structurally identical). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Concluding, we found that the cut-based free-energy profile of a flexible peptide is preserved when using single derivatives or quasi-Newtonian methods to obtain the IS, showing that the population of the substates as well as the height of the energy barriers is conserved (once minimization converged). In agreement with a previous analysis performed on simple liquids, it is found that first- and second-derivative methods occasionally map peptide conformations into different potential energy minima. However, this effect does not influence our understanding of neither the thermodynamics nor the kinetics of the system, and both class of algorithms give consistent results. Consequently, quasi-Newtonian methods, providing higher computational efficiency, are best suited to calculate peptide and protein IS starting from a previously generated trajectory. Finally, an improvement in the implementation of current potential energy functions taking care of the present small discrepancies upon atom permutations will allow a more elegant and accurate combination of the IS formalism with graph-theoretical approaches.

Investigation of further quasi-Newtonian algorithms (like LBFGS) or larger macromolecules is useful but goes beyond the scope of this study.

Acknowledgments

The author is grateful to M. Karplus and D. Wales for useful discussions.

References

1. Rao, F.; Caflisch, A. J Mol Biol 2004, 342, 299.
2. Krivov, S.; Karplus, M. Proc Natl Acad Sci USA 2004, 101, 14766.
3. Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. Proc Natl Acad Sci USA 2007, 104, 1817.
4. Gfeller, D.; de Lachapelle, D. M.; De Los Rios, P.; Caldarelli, G.; Rao, F. Phys Rev E 2007, 76, 026113.
5. Settanni, G.; Fersht, A. R. Biophys J 2008, 94, 4444.
6. Yang, S.; Roux, B. PLoS Comput Biol 2008, 4, e1000047.
7. Prada-Gracia, D.; Gómez-Gardeñes, J.; Echenique, P.; Falo, F. PLoS Comput Biol 2009, 5, e1000415.
8. Rao, F.; Karplus, M. Proc Natl Acad Sci USA 2010, 107, 9152.
9. Rao, F. J Phys Chem Lett 2010, 1, 1580.
10. Li, C. B.; Yang, H.; Komatsuzaki, T. Proc Natl Acad Sci USA 2008, 105, 536.
11. Ihalainen, J. A.; Bredenbeck, J.; Pfister, R.; Helbing, J.; Chi, L.; van Stokkum, I. H.; Woolley, G. A.; Hamm, P. Proc Natl Acad Sci USA 2007, 104, 5383.
12. Ihalainen, J. A.; Paoli, B.; Muff, S.; Backus, E. H.; Bredenbeck, J.; Woolley, G. A.; Caflisch, A.; Hamm, P. Proc Natl Acad Sci USA 2008, 105, 9588.
13. Swope, W.; Pitera, J.; Suits, F. J Phys Chem B 2004, 108, 6571.
14. Noé, F.; Horenko, I.; Schütte, C.; Smith, J. J Chem Phys 2007, 126, 155102.
15. Chodera, J.; Singhal, N.; Pande, V.; Dill, K.; Swope, W. J Chem Phys 2007, 125, 155101.
16. Stillinger, F.; Weber, T. Phys Rev A 1982, 25, 978.
17. Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. J Phys Chem B 2008, 112, 8701.
18. Strodel, B.; Wales, D. Chem Phys Lett 2008, 466, 105.
19. Evans, D.; Wales, D. J Chem Phys 2003, 119, 9947.
20. Evans, D.; Wales, D. J Chem Phys 2004, 121, 1080.
21. Khalili, M.; Wales, D. J Phys Chem B 2008, 112, 2456.
22. Czerminski, R.; Elber, R. J Chem Phys 1990, 92, 5580.
23. Berry, R. S.; Elmaci, N.; Rose, J. P.; Vekhter, B. Proc Natl Acad Sci USA 1997, 94, 9520.
24. Baumketner, A.; Shea, J.-E.; Hiwatari, Y. Phys Rev E 2003, 67, 011912.
25. Nakagawa, N.; Peyrard, M. Proc Natl Acad Sci USA 2006, 103, 5279.
26. Kim, J.; Keyes, T. J Phys Chem B 2007, 111, 2647.
27. Leach, A. Molecular Modelling: Principles and Applications; Addison-Wesley Longman Ltd, 2001.
28. Walsh, T. R.; Wales, D. Z Phys D 1997, 40, 229.
29. Chakravarty, C.; Debenedetti, P.; Stillinger, F. J Chem Phys 2005, 123, 206101.
30. Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
31. Liu, D.; Nocedal, J. Math Program 1989, 45, 503.
32. Cecchini, M.; Houdusse, A.; Karplus, M. PLoS Comput Biol 2008, 4, e1000129.
33. Haberthur, U.; Caflisch, A. J Comput Chem 2008, 29, 701.
34. Bieri, O.; Wirz, J.; Hellrung, B.; Schutkowski, M.; Drewello, M.; Kiefhaber, T. Proc Natl Acad Sci USA 1999, 96, 9597.
35. Möglich, A.; Joder, K.; Kiefhaber, T. Proc Natl Acad Sci USA 2006, 103, 12394.
36. Krivov, S. V.; Karplus, M. J Phys Chem B 2006, 110, 12689.
37. Ferrara, P.; Caflisch, A. Proc Natl Acad Sci USA 2000, 97, 10780.
38. Rao, F.; Caflisch, A. J Chem Phys 2003, 119, 4035.
39. Muff, S.; Caflisch, A. J Chem Phys 2009, 130, 125104.
40. Wales, D.; Doye, J.; Miller, M.; Mortenson, P.; Walsh, T. Adv Chem Phys 2000, 115, 1.
41. Małolepsza, E.; Strodel, B.; Khalili, M.; Trygubenko, S.; Fejer, S. N.; Wales, D. J. J Comput Chem 2010, 31, 1402.