A Natural Coarse Graining for Simulating Large Biomolecular Motion

Holger Gohlke* and M. F. Thorpe†
*Department of Biological Sciences, J. W. Goethe-University, Frankfurt, Germany; and †Center for Biological Physics, Bateman Physical Sciences, Arizona State University, Tempe, Arizona

ABSTRACT Various coarse graining schemes have been proposed to speed up computer simulations of the motion within large biomolecules, which can contain hundreds of thousands of atoms. We point out here that there is a very natural way of doing this, using the rigid regions identified within a biomolecule as the coarse grain elements. Subsequently, computer resources can be concentrated on the flexible connections between the rigid units. Examples of the use of such techniques are given for the protein barnase and the maltodextrin binding protein, using the geometric simulation technique FRODA and the rigidity enhanced elastic network model RCNMA to compute mobilities and atomic displacements.

INTRODUCTION

The first articles applying the numerical technique of molecular dynamics to a protein were in the mid 1970s, beginning with articles such as those by Levitt and Warshel (1) and by Karplus and McCammon (2). In this technique, the classical equations of motion $F=ma$ are integrated forward in time, with the force F being determined from the gradient of a phenomenologically determined potential. Much effort has been devoted to determine potentials suitable for studying proteins, with AMBER (3) and CHARMM (4) being two of the most widely used today, which grew out of the early work on the consistent force field (CFF) (5). In the last ~30 years, molecular dynamics has become the standard technique for studying the motion of proteins, with over 10,000 articles published containing the words ''molecular dynamics simulations'' and ''proteins''. In Fig. 1, we show how the number of articles published, embracing this technique, has continued to increase rapidly, with nearly 1400 articles appearing in 2004.

In recent years, the structures of some very large biomolecular assemblies, like viral capsids (6), the ribosome (7), and membrane protein complexes (8) have been determined by x-ray crystallography. These involve hundreds of thousands of atoms, and are currently presenting a challenge to find simulation techniques to better understand the motion of these large complexes. We can expect many more such structures to become available in the future, using x-ray crystallographic techniques, and probably even larger structures when cryo-EM techniques plus molecular mechanics refinement (9,10) are able to produce structures at atomic resolution.

It is likely that molecular dynamics will continue to produce important insights in the possible local motions of proteins, but there is an urgent need for new techniques so that larger number of atoms can be handled giving motions at 10 Å and greater, corresponding to biological times of up to a second and longer. Current molecular dynamics simulations are limited to ~100 ns for proteins with a few tens of thousands of atoms, which is seven orders of magnitude less than simulations on the scale of up to seconds of biological time that are desirable to explore the diffusive motions of biomolecules. Assuming that Moore's law holds, this would require a wait of nearly 50 years ($10^7 \approx 2^{23}$; because computer power doubles only every 2 years, this results in 46 years in total), which is clearly unacceptable.

A great deal of effort in recent years has been put into accelerating molecular dynamics techniques using, e.g., parallel tempering (11) or larger time steps (12). These enhancements to molecular dynamics techniques are proving useful but are unlikely to be able to produce the orders of magnitude improvements that are now needed. More promising are methods that use spatial coarse graining.

Spatial coarse graining uses larger units than single atoms, in the expectation that such a fine level of detail is not required to describe the motion of very large complexes. (Analogously, motions of electrons need not be considered if one is only interested in the motions of nuclei within the molecular mechanics framework.) This of course must always be justified and great care taken. For example, although coarse graining may work well away from an active site in a protein, it would not be appropriate around ligand binding sites. There are a number of schemes currently in use and under development and we discuss two in detail in this paper. Of course there are many other coarse-grained models like Go models (13,14) that are widely used. Subunits were fixed by using coarse grained protein models as long ago as 1976 (15).

Another coarse grained model that has been used is Rosetta (16), that replaces a short sequence segment (with up to nine residues) by a single body with six degrees of freedom—three translational and three rotational. This model

Submitted February 16, 2006, and accepted for publication June 14, 2006.

Address reprint requests to M. F. Thorpe, Center for Biological Physics, Bateman Physical Sciences, Arizona State University, Tempe, AZ 85287-1504. Tel.: 480-965-3085; Fax: 480-965-4669; E-mail: mft@asu.edu.

0006-3495/06/09/2115/06 $2.00

doi: 10.1529/biophysj.106.083568

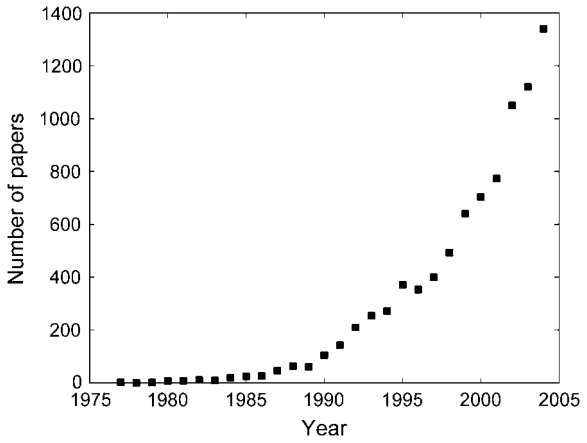

FIGURE 1 Showing how the number of papers applying the molecular dynamics technique to proteins has increased. These figures were found by searching on the words ''molecular dynamics simulations'' and ''protein'' occurring in any field as indexed by Google Scholar. The increase in the number of articles has been rapid but subexponential.

has been widely used in studies of protein folding, for example.

The elastic network model (ENM) (17,18), uses only the C_α atoms as markers for each residue, which are treated as point objects and hence have three degrees of freedom. We will discuss this approach in more detail later.

In this extended comment, we ask the question ''Is there a natural way of choosing groups of atoms for coarse graining'' rather than an arbitrary procedure that selects, for example, every tenth atom. We show that the rigid units of a biomolecular complex can be predetermined using geometrical and topological techniques, and that these do form a natural basis for coarse graining. We give two examples of the current use of such techniques in a geometrical simulation approach (FRODA) and the elastic network model where this approach has recently been incorporated (RCNMA). This approach to coarse graining is straightforward to implement and can be incorporated into almost any numerical simulation technique.

Rigid region decomposition

To use the rigid regions of the biomolecule for coarse graining, we must first review what is meant by this concept. This approach, which is summarized here, has been developed by Thorpe and co-workers in a series of articles (19–23) and is available in the software package FIRST (Floppy Inclusions and Rigid Substructure Topography). A protein can be viewed as being held together by forces of varying strengths. We identify the most important and strongest forces and describe them by constraints. The most important constraints are along the polypeptide chain; the covalent bond lengths and angles, as well as the locked dihedral angle associated with the peptide bond. When the protein undergoes a hydrophobic collapse and folds into the native state, additional constraints come into play. The hydrophobic interactions are described by tethers, and the hydrogen bonds are identified and assigned appropriate constraints. This produces a network of constraints, which is then analyzed to identify the rigid regions and the flexible joints between them. The rigid regions identified in this way can vary in size from three atoms up to a few hundred atoms. Examples of such rigid region decompositions are shown for the protein barnase and the maltodextrin binding protein in Fig. 2.

What do we mean when we say a region is rigid? The point here is that such a region has a well-defined equilibrium structure about which harmonic vibrations are thermally driven and take place about the fixed atomic equilibrium positions. Thus, such rigid regions have vibrational properties similar to those of an amorphous solid (24). However, the biologically important diffusive motion is expected to be associated with the motions of the flexible

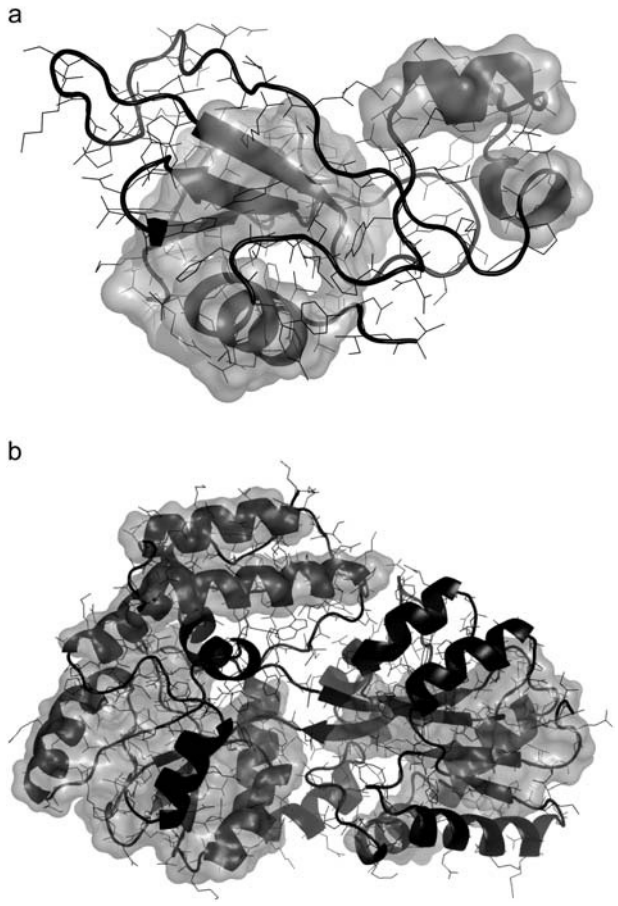

FIGURE 2 Showing (*a*) the three largest rigid regions in the protein barnase and (*b*) the five largest rigid regions in the maltodextrin binding protein determined by the program FIRST (available for download or interactive use via http://flexweb.asu.edu). The largest rigid regions or cores of the proteins are shown in the bottom left-hand corners in both cases. Note that the rigid regions can move as such as they are surrounded by flexible regions.

regions, and this is the part of the structure where numerical methods can most profitably concentrate their attention. Note that no relative motion is allowed within rigid regions. Such regions can only move as a rigid body with six degrees of freedom.

Flexibility is a static property and determines the possibility of motion, where nothing actually moves. It involves only the virtual motion of the network. Finding the rigid and flexible regions is rather like examining a building and identifying parts that are likely to move (doors, windows, etc.). Resources can then be concentrated on those parts of the building in looking for motion (mobility), rather than wasting efforts trying to move fixed walls, etc. Yet, to determine the actual motion and its amplitude requires introducing a kinematics that produces real movements and hence mobility. From a study of rigidity and flexibility alone, no information is available about the direction and amplitude of the possible motions.

Examples

In this section we give two examples showing how the natural coarse graining in terms of the rigid regions as determined by FIRST can be used to study dynamics and hence mobility.

FRODA

In a recent article, a new algorithm (FRODA, which stands for Framework Rigidity Optimized Dynamic Algorithm) was introduced that has been designed to move the flexible parts of the protein, producing motion. The motion of the protein is guided by ghost templates that are specially tailored to ''cover'' each rigid region and then used to efficiently guide the motion through allowed regions of conformational space. In addition to the constraints used in determining the rigid regions, the inequality constraints associated with hard sphere van der Waals overlap are added. This makes the pathway through conformational space tortuous, as the protein can be regarded as a dense packed assembly of spheres, which can roll around each other while maintaining the covalent, hydrophobic, and hydrogen bond constraints between them. Details of this technique can be found elsewhere (25).

After applying FIRST to determine the rigid and flexible regions, FRODA can be used to explore the mobility using random Brownian type (Monte Carlo) dynamics. This procedure emphasizes the geometry of the motion, while including sufficient local chemistry to be realistic. Such an approach can be expected to be particularly appropriate for very large biomolecular assemblies, where the geometry will largely determine the large scale motions.

FRODA suppresses the high frequency motions and focuses on the low frequency diffusive motions and as such can be compared with NMR mobilities as shown in Fig. 3 *a*

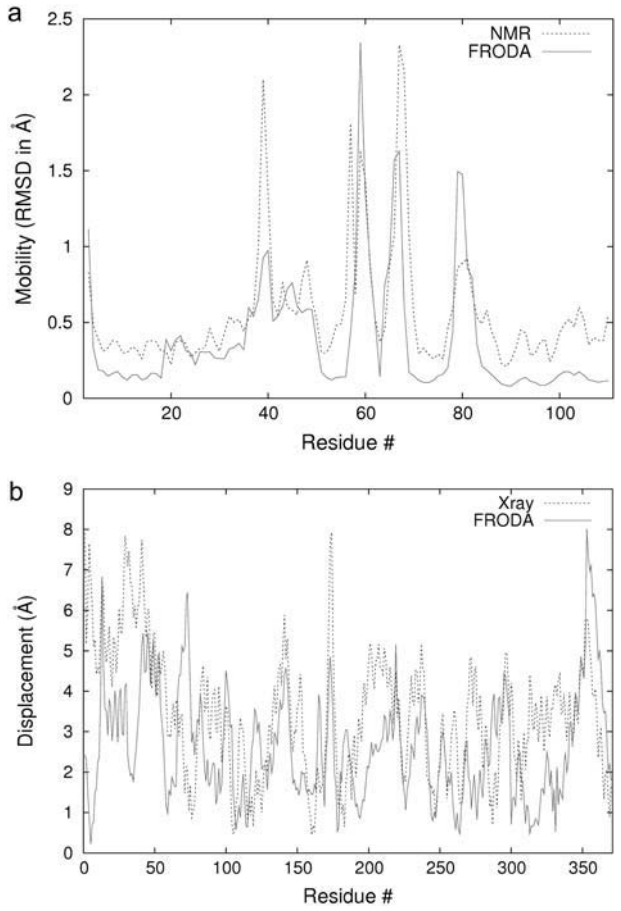

FIGURE 3 Comparing the mobility of barnase (*a*), residue by residue as measured in NMR (*blue line*) with that predicted by FRODA (*red line*). The high-frequency modes that are absent in FRODA are expected to produce a small nearly constant background, which would raise the red curve a little. Note that FRODA gives absolute amplitudes and no scaling is involved. Both sets of data involve 20 conformers that have been globally aligned. The FRODA set was chosen to be maximally separated in root mean square deviation space from the ~10,000 separate conformers generated. In panel *b*, displacements of C_α atoms of the maltodextrin binding protein between a ligand-bound and an apo crystal structure of the protein (*blue line*) as well as predicted by FRODA (*red line*) are shown. The FRODA simulation was started from the apo form, and the displacement of C_α atoms was determined with respect to the 60,000th conformation generated, where the conformation is closer to that of the ligand bound structure.

for barnase. FRODA does not do such a good job in predicting Debye-Waller or *B*-values, which measure the root mean square deviation of each atom about its average position. This is to be expected as coarse-grained methods ignore the higher frequency motions. Whereas mobility occurs in barnase mostly in three loop regions, a large ligand-induced hinge-twist motion between two domains is observed in the case of the maltodextrin binding protein. FRODA is able to qualitatively predict the observed displacements between ligand-bound and apo-forms of the protein (Fig. 3 *b*). This is a much less-defined procedure as the protein wanders around in conformational space in an

undirected way and so would not be expected to reach the ligand bound state exactly—the fact that it gets close is encouraging. With directed targeting, it would be possible to approach the ''target'' closely (25,26), but this was not the purpose here.

RCNMA

Based on an analytical solution to Newton's equations of motion, Normal Mode Analysis (NMA) is able to predict the most probable cooperative motions of molecular systems (27). The introduction of computationally much cheaper alternatives has allowed biologically relevant motions even for systems of the size of the ribosome (28) to be found. In these Elastic Network Models (ENM) (17,29), the all-atom representation used in NMA is replaced with a reduced representation by considering, e.g., only C_α atoms between which simplified potentials in terms of Hookean springs of equal strength act (Fig. 4). Further coarse graining can be achieved if one considers the macromolecule to be constructed of rigid bodies (''blocks'') (15) that are connected by flexible parts (Rotations-Translations of Blocks approach (RTB)) (30). So far, blocks were determined by including up to six protein residues consecutive in sequence into one block (30,31) or by considering whole protein subunits of a virus capsid as rigid (32). However, these routes do not distinguish rigid parts of a protein from flexible regions.

This limitation can be overcome by a recently introduced multiscale modeling approach that combines concepts from rigidity and elastic network theory RCNMA (which stands for Rigid Cluster Normal Mode Analysis) (33). Here, the protein is initially decomposed into rigid clusters by FIRST, circumventing the definition of blocks in an ad hoc manner. Furthermore, tertiary interactions within the protein are considered as flexibility determinants. In the subsequent step, information about amplitudes and directions of motions is obtained for the thus coarse-grained ENM by performing an RTB analysis. By allowing only translational and rotational degrees of freedom of the blocks in this analysis but no relative motions within a block, the system is effectively treated as if C_α atoms within a block were connected by springs of infinite strength.

In terms of efficiency, the coarse-grained ENM has on average only ~30% of the number of degrees of freedom compared to the conventional ENM, resulting in a significant reduction of memory requirements and computational times by factors of 9–27 and 25–125, respectively. In terms of accuracy, the predicted directions and magnitudes of protein motions are at least as good as if no, or a uniform, coarse graining is applied (33). As an example, the mobility of C_α atoms of barnase predicted by the coarse-grained ENM and conventional ENM is shown in Fig. 5 *a*. It can be seen that

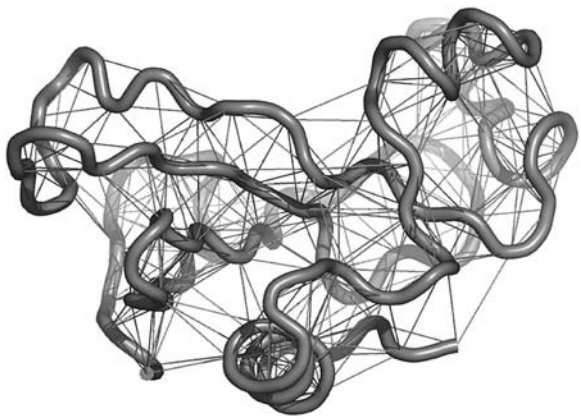

FIGURE 4 ENM representation of barnase. Between C_α atoms (connected by a tube) springs (represented as sticks) of equal strength act. The orientation of the protein is similar to that shown in Fig. 2 *a*.

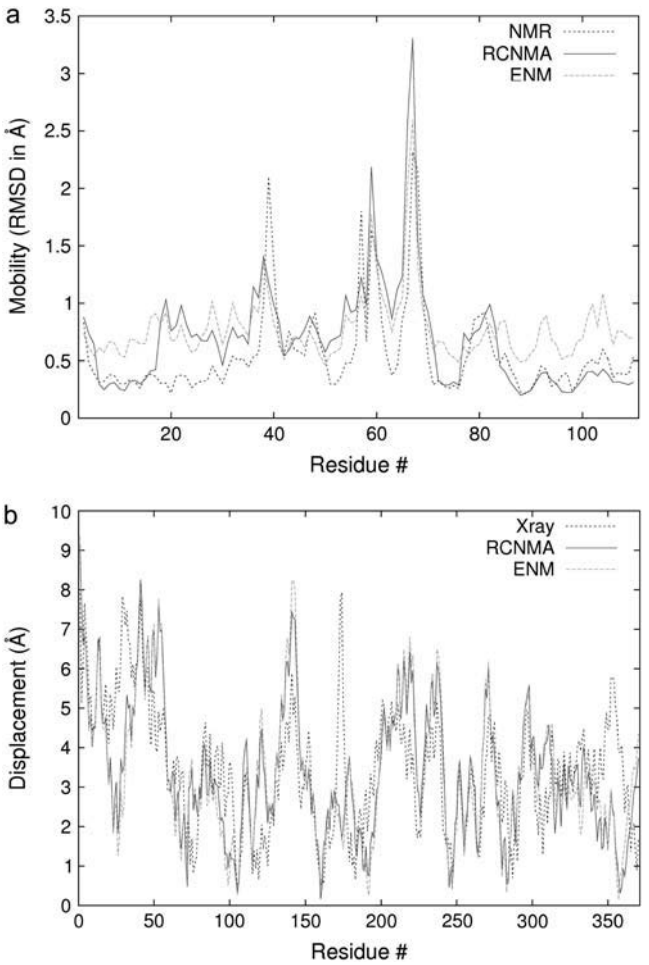

FIGURE 5 (*a*) The mobility of C_α atoms of barnase as measured in NMR (*blue line*) and (*b*) the displacement of C_α atoms of the maltodextrin binding protein between a ligand-bound and an apo crystal structure of the protein (*blue line*). In both cases, conformational changes predicted by the rigidity enhanced ENM (RCNMA) (*red line*; using the rigid cluster decomposition as shown in Fig. 2 *a*) and the conventional ENM (*green line*) are also given. The theoretical curves are scaled with respect to the experimental ones such that the area under the square of the curves is identical (17).

with the rigid regions included, the agreement with the experimentally measured mobilities is considerably improved, particularly in the N- and C-terminal protein regions. This is also demonstrated by a larger correlation coefficient of predicted versus experimental values of $r^2 = 0.56$ in the case of RCNMA compared to $r^2 = 0.50$ in the case of the standard ENM model. A similar result is also found when comparing large conformational changes between a ligand-bound and an apo form of the maltodextrin binding protein with displacements predicted by RCNMA or ENM (Fig. 5 *b*). Accordingly, the correlation coefficients of predicted versus experimental values are $r^2 = 0.62$ and 0.55 for RCNMA and ENM, respectively.

These findings indicate that explicitly distinguishing between flexible and rigid regions is advantageous, because i), it allows to better characterize flexible and rigid regions than with springs of equal strength and ii), it leads to a less rugged energy surface that facilitates the modeling of large-scale motions. We note that the predicted mobility values were scaled to the experimental ones (17). These scaling factors are rather independent of the structure or the sequence of the protein, however (33).

When extrapolating the small harmonic motions described by the ENM to larger amplitudes great care must be taken to avoid the problem of distortions caused by nonlinearities. An example of such a nonlinear distortion would be three equally spaced co-linear points defining a rigid rod, which rotates about the center. In the linear approximation, the outer points move in parallel straight lines in opposite directions, with the center point fixed. If these amplitudes are magnified, the three points no longer just rigidly rotate about the central point, but the length also grows. Likewise, such distortions can show up for example in α-helices by amounts up to 25%, when they should be remaining in the same conformation. This effect will occur whether the α-helices are held rigid, as in the rigidity modified ENM, or if they can flex as in the original ENM. The best way to avoid such distortions is to make a series of very small amplitude motions and then redefining and recomputing an ENM. Such a series of movements can be used to define large-scale motions without introducing distortions caused by nonlinearities.

The second more serious cause of unphysical distortion that occurs in the ENM is that associated with the stretching of the springs between the C_α atoms in the region that should be kept rigid. This occurs because the strength of the springs is the same everywhere in the standard ENM, and so rigid regions will distort as they are insufficiently constrained. This second effect is completely eliminated in the RCNMA approach. Along these lines, a modification of the ENM model has been proposed recently to ease the so-called ''tip-effect''. By increasing the stiffness of degrees of freedom of these regions that are not very densely packed compared to the rest of protein, the pathological behavior in motions of regions protruding out of the main body (such as loops) observed in the conventional ENM model can be eradicated (34).

CONCLUSION

We have shown that there is a natural way of coarse graining that can be used easily and successfully when simulating motions of biomolecules. This coarse graining uses units of variable sizes that correspond to the predetermined rigid regions found by applying FIRST, which determines rigid regions and flexible joints that separate them from a network representation of the molecule, consisting of covalent, hydrophobic, and hydrogen bonds. We have used the protein barnase and the maltodextrin binding protein as illustrative examples and applied two approaches, a geometrical simulation approach, FRODA and a rigidity enhanced elastic network model RCNMA, to compute mobilities, obtaining good agreement with experimental results in both cases. Coarse graining, using regions of variable size, as determined by finding the rigid regions, is a natural way to proceed and should be useful as a front end for many numerical simulation procedures, and not just the two discussed in this article. An example is the recent work on the kinetics of viral capsid assembly, using a FIRST coarse graining to reduce the total number of degrees of freedom (35).

We thank Brandon Hespenheide, Scott Menor, Stephen Wells, and Aqeel Ahmed for continuing conversations. Parts of this work were done during the workshop ''Dynamics under Constraints'' at Bellairs Research Institute of McGill University, Holetown, Barbados, January, 2006.

H.G. is grateful to Merck KGaA, Darmstadt, and the J. W. Goethe-University for financial support. M.F.T. acknowledges financial support by the National Science Foundation (grant No. DMR-0425970), National Institutes of Health (grant No. GM067249), and the Arizona State University Foundation.

REFERENCES

1. Levitt, M., and A. Warshel. 1975. Computer simulation of protein folding. *Nature*. 253:694–698.
2. McCammon, J. A., and M. Karplus. 1977. Internal motions of antibody molecules. *Nature*. 268:765–766.
3. Pearlman, D. A., D. A. Case, J. D. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. 1995. Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91:1–41.
4. Brooks, R. R., B. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.
5. Warshel, A., and S. Lifson. 1969. An empirical function for second neighbor interactions and its effect on vibrational modes. *Chem. Phys. Lett.* 4:255–256.
6. Prasad, B. V., M. E. Hardy, T. Dokland, J. Bella, M. G. Rossmann, and M. K. Estes. 1999. X-ray crystallographic structure of the Norwalk virus capsid. *Science*. 286:287–290.
7. Yusupov, M. M., G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller. 2001. Crystal structure of the ribosome at 5.5 A resolution. *Science*. 292:883–896.

8. Ferreira, K. N., T. M. Iverson, K. Maghlaoui, J. Barber, and S. Iwata. 2004. Architecture of the photosynthetic oxygen-evolving center. *Science*. 303:1831–1838.

9. Ma, J. 2004. New advances in normal mode analysis of supermolecular complexes and applications to structural refinement. *Curr. Protein Pept. Sci.* 5:119–123.

10. Tama, F., O. Miyashita, and C. L. Brooks 3rd. 2004. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.* 147:315–326.

11. Mitsutake, A., Y. Sugita, and Y. Okamoto. 2001. Generalized-ensemble algorithms for molecular simulations of biomolecules. *Biopolymers.* 60:96–123.

12. Tuckerman, M. E., B. J. Berne, and G. J. Martyna. 1992. Reversible multiple timescale molecular dynamics. *J. Chem. Phys.* 97:1990–2001.

13. Go, N., and H. Abe. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers.* 20: 991–1011.

14. Abe, H., and N. Go. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers.* 20:1013–1031.

15. Warshel, A., and M. Levitt. 1976. Folding and stability of helical proteins: carp myogen. *J. Mol. Biol.* 106:421–437.

16. Rohl, C. A., C. E. Strauss, K. M. Misura, and D. Baker. 2004. Protein structure prediction using Rosetta. *Methods Enzymol.* 383:66–93.

17. Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2:173–181.

18. Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–515.

19. Jacobs, D. J., and M. F. Thorpe. 1995. Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.* 75:4051–4054.

20. Jacobs, D. J., L. A. Kuhn, and M. F. Thorpe. 1999. Flexible and rigid regions in proteins. *In* Rigidity Theory and Applications. M. F. Thorpe and P. M. Duxbury, editors. Kluwer Academic/Plenum Publishers, New York. 357–384.

21. Thorpe, M. F., M. Lei, A. J. Rader, D. J. Jacobs, and L. A. Kuhn. 2001. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model.* 19:60–69.

22. Rader, A. J., B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe. 2002. Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. USA.* 99: 3540–3545.

23. Hespenheide, B. M., D. J. Jacobs, and M. F. Thorpe. 2004. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys. Condens. Matter.* 16:S5055–S5064.

24. Thorpe, M. F., D. J. Jacobs, N. V. Chubynsky, and A. J. Rader. 1999. Generic rigidity of network glasses. *In* Rigidity Theory and Applications. M. F. Thorpe and P. M. Duxbury, editors. Kluwer Academic/Plenum Publishers, New York. 239–277.

25. Wells, S., S. Menor, B. M. Hespenheide, and M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motions in proteins. *Phys. Biol.* 2:1–10.

26. Samuel Flores, N. E., D. Milburn, B. Hespenheide, S. Wells, K. Keating, J. Lu, M. Thorpe, and M. Gerstein. 2006. New features in the database of macromolecular motions. *Nucleic Acids Res.* 34:D296–D301.

27. Ma, J. 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure.* 13: 373–380.

28. Tama, F., M. Valle, J. Frank, and C. L. Brooks III. 2003. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl. Acad. Sci. USA.* 100:9319–9323.

29. Tirion, M. M. 1996. Large amplitude elastic motions in proteins from single-parameter atomic analysis. *Phys. Rev. Lett.* 77:1905–1908.

30. Tama, F., F. X. Gadea, O. Marques, and Y. H. Sanejouand. 2000. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins.* 41:1–7.

31. Li, G., and Q. Cui. 2002. A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca^{2+}-ATPase. *Biophys. J.* 83:2457–2474.

32. Tama, F., and C. L. Brooks 3rd. 2002. The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.* 318:733–747.

33. Ahmed, A., and H. Gohlke. 2006. Multi-scale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins.* 63:1038–1051.

34. Lu, M., B. Poon, and J. Ma. 2006. A new method for coarse-grained elastic normal-mode analysis. *Journal of Chemical Theory and Computation.* 2:464–471.

35. Hemberg, M., S. N. Yaliraki, and M. Barahona. 2006. Stochastic kinetics of viral capsid assembly based on detailed protein structures. *Biophys. J.* 90:3029–3042.