Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Molecular simulation as an aid to experimentalists

Wilfred F van Gunsteren[1], Jožica Dolenc[1,2] and Alan E Mark[3]

Computer-based molecular simulation techniques are increasingly used to interpret experimental data on biomolecular systems at an atomic level. Direct comparison between experiment and simulation is, however, seldom straightforward. The available experimental data are limited in scope and generally correspond to averages over both time and space. A critical analysis of the various factors that may influence the apparent degree of agreement between the results of simulations and experimentally measured quantities is presented and illustrated using examples from recent literature.

Addresses
[1] Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH, CH-8093 Zurich, Switzerland
[2] Faculty of Chemistry and Chemical Technology, University of Ljubljana, SI-1000 Ljubljana, Slovenia
[3] School of Molecular and Microbial Sciences and the Institute of Molecular Biosciences, University of Queensland, Brisbane QLD 4072, Australia

Corresponding author: van Gunsteren, Wilfred F (wfvgn@igc.phys.chem.ethz.ch)

Current Opinion in Structural Biology 2008, **18**:149–153

This review comes from a themed issue on
Theory and simulation
Edited by Chandra Verma and Juan Fernandez Recio

Available online 14th February 2008

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2007.12.007

Introduction

Computer-based molecular simulation techniques have become a standard tool to facilitate the interpretation of experimental data at an atomic or molecular level. If the assumptions and approximations inherent in a simulation model, such as (1) which degrees of freedom are simulated, (2) how the forces acting on the particles are approximated, (3) how the system is propagated in time and (4) the nature of the thermodynamic and spatial boundary conditions, are appropriate [1], any property $Q(\mathbf{r}, \mathbf{p})$ that depends primarily on the coordinates $\mathbf{r}$ and momenta $\mathbf{p}$ of the chosen degrees of freedom can in principle be calculated and compared to its experimental counterpart. Experimental properties commonly compared to atomistic simulations include X-ray or neutron diffraction intensities, NMR parameters (i.e. chemical shifts, J-coupling constants, relaxation times and dipolar-couplings), various quantities derived from EPR, CD, infrared, Raman or fluorescence spectroscopy, data derived from measurements of fluorescence energy transfer, and even electrical or mechanical response functions (e.g. AFM).

There are two basic reasons why molecular simulations can facilitate the interpretation of experimental data from biomolecular systems. The first is that experimentally measured properties correspond in general to averages of a quantity Q over both space and time, $\langle Q \rangle_{\text{space,time}}$ where $\langle Q \rangle$ is an average over the distribution $P(Q(\mathbf{r}, \mathbf{p}))$ of Q-values depending on single configuration and momentum values [2]. The averaging inherent in the measurement means it is usually not possible to determine the Q-distribution $P(Q(\mathbf{r}))$ or the configurational distribution $P(\mathbf{r})$ from $\langle Q(\mathbf{r}) \rangle_{\text{exp}}$ (note, the $\mathbf{p}$-dependence has been omitted for simplicity). Furthermore, representing the conformational distribution $P(\mathbf{r})$ by one conformation derived from, for example, X-ray crystallography, may lead to inconsistencies when calculating alternative experimental quantities that have a different dependence on $P(\mathbf{r})$. Examples of this situation can be found in references [3–5]. Measured NOE intensities may result from different molecular conformations [3]. Seemingly contradictory data from X-ray and NMR studies of the *cis–trans* interconversion induced by cyclophilin could be explained from the conformational distribution obtained by MD simulation [4]. Small-angle X-ray scattering profiles calculated from simulated ensemble averages agree better with experimental data than profiles calculated from single X-ray crystal structures [5]. The second reason simulations facilitate the interpretation of experimental data is that the number of quantities N_Q^{exp} that can be measured for a molecular system is generally very much smaller than the number of degrees of freedom N_{dof} of the system. This makes the problem of determining the conformational distribution $P(\mathbf{r})$ from the set of $\langle Q(\mathbf{r}) \rangle_{\text{exp}}$ values highly underdetermined [6–8] even if combining alternative sets of experimental data pertaining to one system could improve the situation [9].

These problems can be circumvented partly by using molecular simulations to generate conformational distributions $P(\mathbf{r})$ from which averages $\langle Q(P(\mathbf{r})) \rangle$ can be obtained and compared or coupled to experimental values $\langle Q \rangle_{\text{exp}}$ [10]. The use of molecular simulations to interpret experimental data is growing rapidly in popularity with progressively more indirect and adventurous comparisons being made. In light of this we feel it is appropriate to focus this contribution on the potential pitfalls of using molecular simulation in this manner. In particular, we present simple criteria against which such studies can be

judged and illustrate these with reference to recent literature.

Interpretation of experimental data using simulation

When relating the *average* of a property over a given conformational distribution $P(\mathbf{r})$, whether from a simulation ($\langle Q\rangle_{\mathrm{sim}}$) or measured experimentally ($\langle Q\rangle_{\mathrm{exp}}$), to the conformational distribution itself, three general cases can be distinguished:

(Q1) $\langle Q\rangle$ does not reflect the shape of $P(\mathbf{r})$ as $\langle Q\rangle$ is insensitive to conformation; see, for example, reference [11].
(Q2) $\langle Q\rangle$ does not reflect the shape of $P(\mathbf{r})$ as $\langle Q\rangle$ is determined by rarely sampled conformations with small (*irrelevant*) Boltzmann weights; see, for example, reference [12].
(Q3) $\langle Q\rangle$ reflects the dominant conformations of $P(\mathbf{r})$.

Only in the case Q3 can $\langle Q\rangle_{\mathrm{sim}}$ carry information relevant to the interpretation of $\langle Q\rangle_{\mathrm{exp}}$ at a molecular level. Nevertheless, agreement between $\langle Q\rangle_{\mathrm{sim}}$ with $\langle Q\rangle_{\mathrm{exp}}$ may be obtained if:

(A1) $\langle Q\rangle$ is insensitive to $P(\mathbf{r})$, that is, $\langle Q\rangle_{\mathrm{sim}}$ matches $\langle Q\rangle_{\mathrm{exp}}$ irrespective of the $P(\mathbf{r})$ being simulated [1].
(A2) There are compensating errors in the simulation model, procedure or experimental set-up [13]; see D1-5 below.
(A3) The experimental data of interest, $\langle Q\rangle_{\mathrm{exp}}$, have been used to bias the simulation.
(A4) $\langle Q(\mathbf{r})\rangle_{\mathrm{sim}}$ is sensitive to the distribution $P(\mathbf{r})$.

Again, only in the case A4 can the degree of agreement between $\langle Q\rangle_{\mathrm{sim}}$ and $\langle Q\rangle_{\mathrm{exp}}$ be used to validate the simulation and/or to interpret the experimental results.

Failure to observe a correlation between the simulation and experiment can also be due to many reasons:

(D1) The simulation is insufficiently accurate, that is, (a) relevant degrees of freedom were omitted; (b) the force field was insufficiently accurate; (c) approximations made when solving the equations of motion were too crude; (d) inappropriate thermodynamic or spatial boundary conditions were used.
(D2) The measured $\langle Q\rangle_{\mathrm{exp}}$ is inaccurate.
(D3) $\langle Q\rangle_{\mathrm{sim}}$ and $\langle Q\rangle_{\mathrm{exp}}$ are averaged differently with respect to time or spatial extent.
(D4) Related but different quantities are compared, for example, atom-positional fluctuations versus crystallographic B factors.
(D5) Different systems are compared (e.g. crystal versus solution), or systems studied under different thermodynamic conditions (e.g. temperature, pressure, pH, ionic strength, etc.).

Clearly the interpretation of experimental observables in terms of molecular conformations is not straightforward. In particular, the degree of agreement between $\langle Q\rangle_{\mathrm{sim}}$ and $\langle Q\rangle_{\mathrm{exp}}$ can easily be over interpreted and frequently leads researchers to make unrealistic claims regarding the validity of a simulation (when an apparent correlation between theory and experiment is observed) or, equally problematic, leads to the unjustified disregard of simulation results (when no correlation is evident).

Note, in all of the above $\langle Q\rangle_{\mathrm{exp}}$ denotes an observable quantity, that is, a property that can be measured directly. Such primary experimental data should not be confused with secondary experimental data, Q', that is, data derived from $\langle Q\rangle_{\mathrm{exp}}$ by applying a given procedure, f, based on various assumptions and approximations: $Q' = f(\langle Q\rangle_{\mathrm{exp}})$. For example, whereas peak location and intensity from X-ray diffraction or NMR spectroscopic experiments represent primary data, molecular structures, NMR order parameters, and so on are secondary (derived) quantities. Comparisons of simulations to secondary, non-observed, experimental quantities reflect, at least partly, the approximations and assumptions associated with the conversion and may in reality carry little experimental information [6]. Clearly, when coupling (restraining) a simulation to a set of $\langle Q\rangle_{\mathrm{exp}}$ values in order to ensure that the conformational distribution satisfies $\langle Q(\mathbf{r})\rangle_{\mathrm{sim}} = \langle Q\rangle_{\mathrm{exp}}$, only primary experimental data should be used. For further discussion on the validation of molecular simulations see reference [14].

Experimentally observable quantities

Below, we illustrate the use of biomolecular simulations to interpret a range of experimentally observed quantities Q in terms of the considerations outlined above using examples from the recent literature. One quantity that is relatively insensitive to the particular conformational distribution is the small-angle X-ray scattering intensity, from which the radius of gyration can be derived. As shown in reference [7] very different conformational distributions can give rise to a given experimental observation. This study also illustrates the difference between comparing to primary data (X-ray intensities) as opposed to secondary data (radius of gyration). Questionable comparisons to secondary quantities are made in many studies. For example in [15] fluorescence measurements on the Trp-cage protein are used to infer changes in enthalpy, which are used to infer changes in the number of hydrogen bonds as a function of temperature, which are ultimately related to results from simulations [15]. In a similar manner the validity of simulations performed with J-coupling constants was assessed by comparing to proposed hydrogen bond energies instead of primary quantities [16]. In such cases the basic comparison is

between alternative modeling procedures, not between simulation and experiment. By contrast, where the comparison is to primary experimental quantities such as infrared spectra [17,18] or NMR spectra [19,20], a direct validation of the simulation models by experiment is possible.

The use of the so-called phi-values to restrain or validate folding simulations deserves a special comment. Phi-value analysis is a useful framework within which to analyze experimental data on folding transitions in proteins [21]. Phi-values are secondary data derived from the effect of mutations on stability and folding kinetics. An equivalent quantity cannot be extracted from simulations; instead phi-values are often defined in terms of the so-called 'native contacts'. Clearly, as the definitions of these two quantities are very different and even span different ranges, any apparent correlation can neither be used to validate the simulation nor to interpret the experiment (see D4 and D5 above) [22–24]. A comparison of simulated versus measured folding rates for proteins [25,26] based on many short simulations is also problematic owing to the different time scales accessible to simulations and experiment (D3), the definition of when the system is folded and assumptions in regard to the probability of folding at long time scales.

In contrast to X-ray or NMR studies that generally yield thousands of individual parameters (diffraction intensities or atom–atom resonances), in some types of experiment only a single quantity is measured, for example, atomic force microscopy, ionic conductance measurements or wide angle X-ray spectra. Such data are generally not suited for model validation owing to inability to relate the measurement to specific degrees of freedom within the model and thus the high probability of compensation of errors (A2 above) [27–30]. This problem is further aggravated when the quantity in question represents secondary experimental data such as in the calculation of the area per lipid, NMR order parameters along the lipid tails in membrane systems or even density profiles across membranes [31,32].

A constant difficulty in the study of biomolecular systems including proteins has been the limited availability of experimental data with high information content. Fortunately, the range of experimental quantities that are being compared to simulations continues to expand. For example, the response of groups of atoms within proteins to local electric fields has been measured experimentally and analyzed using simulations [33]. Comparisons between measured and calculated Raman spectra have provided new insight into the dynamics of water surrounding sugars [34]. Measurements of fluorescence anisotropy [35] and fluorescence lifetimes [36,37] have also been directly compared to the results of simulations providing novel insights into local dynamics as well as helping to validate the simulation models. MD simulations have shown mobile water molecules inside cavities in proteins that are responsible for detectable NOEs [38] or electron density [39], which could not be traced to fixed water molecules.

Finally, we must draw attention to a few studies [40–43] in which the experimental data $\langle Q \rangle_{\mathrm{exp}}$ that have been used to bias the simulation when generating a conformational distribution $P(\mathbf{r})$ is subsequently used to compare $\langle Q(P(\mathbf{r})) \rangle_{\mathrm{sim}}$ with $\langle Q \rangle_{\mathrm{exp}}$ (see A3 above). Such comparisons do not validate the simulation nor demonstrate that the conformational distribution reflects experiment. In fact, restraining $Q(\mathbf{r})$ in the simulation to match $\langle Q \rangle_{\mathrm{exp}}$ can easily introduce artifacts in $P(\mathbf{r})$ even though the correct average $\langle Q \rangle_{\mathrm{sim}} = \langle Q \rangle_{\mathrm{exp}}$ is maintained, enhancing the likelihood of inappropriate conclusions being drawn [44].

Conclusions

Without question molecular simulations play a crucial role in facilitating the interpretation of experimental data at the atomic level. Nevertheless, it must be remembered that compared to the number of degrees of freedom within biomolecular systems of interest, experimental data are in general very limited in nature. In addition, almost all measured properties are inherently averaged quantities. As a consequence, comparisons between simulations and experiment while seemingly straightforward can be easily overinterpreted or misinterpreted both in terms of the apparent quality of a simulation as well as in terms of the inability of the model to reproduce the experimental observable of interest. In particular, caution must be exercised when comparing simulations to secondary or derived experimental data. The use of secondary data as restraints in simulations often leads to artifacts and should be avoided where possible. In all cases where experimental data are compared with or incorporated into a simulation care must be taken that the data relate to the same molecular system under the same thermodynamic conditions and correspond to the same physical quantities as in the simulation. Finally, experimental data that have been used to bias or restrain a simulation cannot subsequently be used for validation. Failure to observe these simple considerations will make it more likely that excessive claims regarding the validity of simulation results are made or for simulation results to be rejected without justification. More seriously it will mean that inappropriate or even incorrect physical interpretations of experimental data will be propagated in the literature.

Acknowledgements

Financial support by the National Centre of Competence in Research (NCCR) (Structural Biology) of the Swiss National Science Foundation (SNSF), the Slovenian Research Agency (ARRS) and the Australian Research Council (ARC) is gratefully acknowledged.

References

1. van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, Gee P, Geerke DP, Glättli A, Hünenberger PH *et al.*: **Biomolecular modeling: goals, problems, perspectives**. *Angew Chem Int Ed* 2006, **45**:4064-4092.

2. Zagrovic B, Pande VS: **How does averaging affect protein structure comparison on the ensemble level?** *Biophys J* 2004, **87**:2240-2246.

3. Trzesniak D, Glättli A, Jaun B, van Gunsteren WF: **Interpreting NMR data for beta-peptides using molecular dynamics simulations**. *J Am Chem Soc* 2005, **127**:14320-14329.

4. Trzesniak D, van Gunsteren WF: **Catalytic mechanism of cyclophilin as observed in molecular dynamics simulations: pathway prediction and reconciliation of X-ray crystallographic and NMR solution data**. *Protein Sci* 2006, **15**:2544-2551.

5. Lau AY, Roux B: **The free energy landscapes governing conformational changes in a glutamate receptor ligand-binding domain**. *Structure* 2007, **15**:1203-1214.

6. Glättli A, van Gunsteren WF: **Are NMR-derived model structures for beta-peptides representative for the ensemble of structures adopted in solution?** *Angew Chem Int Ed* 2004, **43**:6312-6316.

7. Zagrovic B, Jayachandran G, Millett IS, Doniach S, Pande VS: **How large is an alpha-helix? Studies of the radii of gyration of helical peptides by small-angle X-ray scattering and molecular dynamics**. *J Mol Biol* 2005, **353**:232-241.

8. Zagrovic B, van Gunsteren WF: **Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us?** *Proteins: Struct Funct Bioinf* 2006, **63**:210-218.

9. Fuchs PFJ, Bonvin A, Bochicchio B, Pepe A, Alix AJP, Tamburro AM: **Kinetics and thermodynamics of type VIII beta-turn formation: a CD, NMR, and microsecond explicit molecular dynamics study of the GDNP tetrapeptide**. *Biophys J* 2006, **90**:2745-2759.

10. van Gunsteren WF, Bonvin AMJJ, Daura X, Smith LJ: **Aspects of modeling biomolecular structure on the basis of spectroscopic or diffraction data**. In *Biological Magnetic Resonance: Structure Computation and Dynamics in Protein NMR.* Edited by Rama Krishna N, Berliner LJ. Springer; 1999:3-35.

11. Daura X, Antes I, van Gunsteren WF, Thiel W, Mark AE: **The effect of motional averaging on the calculation of NMR-derived structural properties**. *Proteins: Struct Funct Genet* 1999, **36**:542-555.

12. Glättli A, Daura X, Seebach D, van Gunsteren WF: **Can one derive the conformational preference of a beta-peptide from its CD spectrum?** *J Am Chem Soc* 2002, **124**:12972-12978.

13. van Gunsteren WF: **On testing theoretical models by comparison of calculated with experimental data**. In *Studies in Physical and Theoretical Chemistry: Modelling of Molecular Structures and Properties.* Edited by Rivail J-L. Elsevier; 1990:463-478.

14. van Gunsteren WF, Mark AE: **Validation of molecular dynamics simulation**. *J Chem Phys* 1998, **108**:6109-6116.

15. Iavarone AT, Patriksson A, van der Spoel D, Parks JH: **Fluorescence probe of Trp-cage protein conformation in solution and in gas phase**. *J Am Chem Soc* 2007, **129**:6726-6735.

16. Gsponer J, Hopearuoho H, Cavalli A, Dobson CM, Vendruscolo M: **Geometry, energetics, and dynamics of hydrogen bonds in proteins: structural information derived from NMR scalar couplings**. *J Am Chem Soc* 2006, **128**: 15127-15135.

17. Mu YG, Kosov DS, Stock G: **Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments**. *J Phys Chem B* 2003, **107**:5064-5073.

18. Gnanakaran S, Hochstrasser RM, Garcia AE: **Nature of structural inhomogeneities on folding a helix and their influence on spectral measurements**. *Proc Natl Acad Sci U S A* 2004, **101**:9229-9234.

19. Peter C, Rueping M, Wörner HJ, Jaun B, Seebach D, van Gunsteren WF: **Molecular dynamics simulations of small peptides: can one derive conformational preferences from ROESY spectra?** *Chem Eur J* 2003, **9**:5838-5849.

20. Showalter SA, Brüschweiler R: **Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: application to the AMBER99SB force field**. *J Chem Theor and Comp* 2007, **3**:961-975.

21. Itzhaki LS, Otzen DE, Fersht AR: **The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods—evidence for a nucleation–condensation mechanism for protein-folding**. *J Mol Biol* 1995, **254**:260-288.

22. Settanni G, Rao F, Caflisch A: **Phi-value analysis by molecular dynamics simulations of reversible folding**. *Proc Natl Acad Sci U S A* 2005, **102**:628-633.

23. Salvatella X, Dobson CM, Fersht AR, Vendruscolo M: **Determination of the folding transition states of barnase by using Phi(l)-value-restrained simulations validated by double mutant Phi(lJ)-values**. *Proc Natl Acad Sci U S A* 2005, **102**:12389-12394.

24. Gsponer J, Hopearuoho H, Whittaker SBM, Spence GR, Moore GR, Paci E, Radford SE, Vendruscolo M: **Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7**. *Proc Natl Acad Sci U S A* 2006, **103**:99-104.

25. Snow CD, Nguyen N, Pande VS, Gruebele M: **Absolute comparison of simulated and experimental protein-folding dynamics**. *Nature* 2002, **420**:102-106.

26. Snow CD, Sorin EJ, Rhee YM, Pande VS: **How well can simulation predict protein folding kinetics and thermodynamics?** *Annu Rev Biophys Biomol Struct* 2005, **34**:43-69.

27. Hummer G, Szabo A: **Free energy surfaces from single-molecule force spectroscopy**. *Acc Chem Res* 2005, **38**:504-513.

28. Sotomayor M, Schulten K: **Single-molecule experiments *in vitro* and *in silico***. *Science* 2007, **316**:1144-1148.

29. Sotomayor M, Vasquez V, Perozo E, Schulten K: **Ion conduction through MscS as determined by electrophysiology and simulation**. *Biophys J* 2007, **92**:886-902.

30. Sega M, Garberoglio G, Brocca P, Cantu L: **Microscopic structure of phospholipid bilayers: comparison between molecular dynamics simulations and wide-angle X-ray spectra**. *J Phys Chem B* 2007, **111**:2484-2489.

31. Klauda JB, Kučerka N, Brooks BR, Pastor RW, Nagle JF: **Simulation-based methods for interpreting X-ray data from lipid bilayers**. *Biophys J* 2006, **90**:2796-2807.

32. Cournia Z, Ullmann GM, Smith JC: **Differential effects of cholesterol, ergosterol and lanosterol on a dipalmitoyl phosphatidylcholine membrane: a molecular dynamics simulation study**. *J Phys Chem B* 2007, **111**:1786-1801.

33. Suydam IT, Snow CD, Pande VS, Boxer SG: **Electric fields at the active site of an enzyme: direct comparison of experiment with theory**. *Science* 2006, **313**:200-204.

34. Lerbret A, Bordat P, Affouard F, Guinet Y, Hedoux A, Paccou L, Prevost D, Descamps M: **Influence of homologous disaccharides on the hydrogen-bond network of water: complementary Raman scattering experiments and molecular dynamics simulations**. *Carbohydr Res* 2005, **340**:881-887.

35. Schröder GF, Alexiev U, Grubmuller H: **Simulation of fluorescence anisotropy experiments: probing protein dynamics**. *Biophys J* 2005, **89**:3757-3770.

36. Groenhof G, Bouxin-Cademartory M, Hess B, De Visser SP, Berendsen HJC, Olivucci M, Mark AE, Robb MA: **Photoactivation of the photoactive yellow protein: why photon absorption**

triggers a *trans*-to-*cis* Isomerization of the chromophore in the protein. *J Am Chem Soc* 2004, **126**:4228-4233.

37. Nilsson L, Halle B: **Molecular origin of time-dependent fluorescence shifts in proteins**. *Proc Natl Acad Sci U S A* 2005, **102**:13867-13872.

38. Somani S, Chng C-P, Verma CS: **Hydration of a hydrophobic cavity and its functional role: a simulation study of human interleukin-1β**. *Proteins: Struct Funct Bioinf* 2007, **67**:868-885.

39. Schiffer CA, van Gunsteren WF: **Accessibility and order of water sites in and around proteins: a crystallographic time-averaging study**. *Proteins: Struct Funct Genet* 1999, **36**:501-511.

40. Paci E, Lindorff-Larsen K, Dobson CM, Karplus M, Vendruscolo M: **Transition state contact orders correlate with protein folding rates**. *J Mol Biol* 2005, **352**:495-500.

41. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M: **Simultaneous determination of protein structure and dynamics**. *Nature* 2005, **433**:128-132.

42. Vendruscolo M: **Determination of conformationally heterogeneous states of proteins**. *Curr Opin Struct Biol* 2007, **17**:15-20.

43. Scott KA, Randles LG, Moran SJ, Daggett V, Clarke J: **The folding pathway of spectrin R17 from experiment and simulation: using experimentally validated MD simulations to characterize states hinted at by experiment**. *J Mol Biol* 2006, **359**:159-173.

44. Torda AE, Scheek RM, van Gunsteren WF: **Time-averaged nuclear overhauser effect distance restraints applied to tendamistat**. *J Mol Biol* 1990, **214**:223-235.