Drug Discovery Today • Volume 00, Number 00 • April 2012 PERSPECTIVE

ELSEVIER

feature

High-throughput molecular dynamics: the powerful new tool for drug discovery

Matthew J. Harvey[1] **and Gianni De Fabritiis**[2], gianni.defabritiis@upf.edu

Molecular dynamics simulations are capable of resolving molecular recognition processes with chemical accuracy, but their practical application is popularly considered limited to the timescale accessible to a single simulation, which is far below biological timescales. In this perspective article, we propose that the true limiting factor for molecular dynamics is rather the high hardware and electrical power costs, which constrain not only the length of runs but also the number that can be performed concurrently. As a result of innovation in accelerator processors and high-throughput protocols, the cost of molecular dynamics sampling has been dramatically reduced and we argue that molecular dynamics simulation is now placed to become a key technology for in silico drug discovery in terms of binding pathways, poses, kinetics and affinities.

Computational techniques are widespread in the initial stages of drug discovery: high-throughput virtual screening of large libraries of molecules, hit-to-lead development supported by computational structure–activity relationship studies and re-docking of leads using more computationally expensive techniques [1,2] are all routinely used to guide the early stages of rational design. Despite the utility of these methods, the approximations made in the representation of targets are large. In particular, there is now widespread consensus that the use of a rigid model of a protein target is highly limiting as it fails to account for natural conformational variations, and in particular for the thermodynamics and kinetics of the binding process, which can be of great importance in drug action [3,4].

One established technique for simulating molecular systems is molecular dynamics (MD) simulation, in which protein and drug are modelled classically using Newtonian mechanics. As a computational method able to quantify the process of molecular recognition between a protein and small molecules, going beyond a qualitative description to provide, for example, estimates of affinities and kinetic rates, MD would seem of broad relevance to the drug discovery field. Nevertheless, routine MD of biological systems has long been perceived to be intractable as the timescale accessible by a single simulation is insufficient to resolve the timescales of biologically-relevant processes. We propose that this limitation is owing to the high resource cost in terms of hardware and power, which limits not only long runs but also the ability to perform high-throughput ensemble simulations. Recently, modern graphical processing units (GPUs) have shown to be highly capable at MD simulations, to the extent that a GPU-equipped machine might now perform microsecond-scale simulations comparable to that of a large computer cluster [1]. By dramatically reducing the cost of generating MD trajectories and new high-throughput MD protocols, we argue that high-throughput MD will lead molecular simulation to become a standard tool in biological research and biomedicine.

Technology of MD simulation

MD modelling of biomolecules typically treats each atom of the solvent and solute as separate point particle. A force-field, parametrised to capture the chemical properties of the environment of each type of particle governs the evolution of the system, which proceeds according to Newtonian dynamics in a stepwise

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2012.03.017

Please cite this article in press as: Harvey, M.J. High-throughput molecular dynamics: the powerful new tool for drug discovery, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.03.017

TABLE 1

Representative timescales for significant processes in protein motion and associated computation times on 1 GPU and a reasonable GPU cluster. The computational time is estimated for a protein system in explicit solvent (24k atoms), simulated on an Nvidia Geforce 580 GPU with ACEMD [1]. Timescales derived from Fig. 1 of [5]

Timescale	Process	1 GPU compute time	100 GPU compute time
1 fs–1 ps	Bond vibrations	1–1000 ms	
1 ps–1 ns	Side-chain rotations Hinge bending	1–1000 s	Seconds
1 ns–1 us	Loop motions Helix coil transitions	10 min–10 days	Seconds to hours
1 us–1 ms	Allosteric modulation Molecular recognition	10 days–30 years	Hours to months
1 ms–1 s	Protein folding	30 years	Months to years

manner. At each step of the simulation, the net force on each particle is calculated and the particles' positions updated.

To correctly capture the dynamics of the system the timestep of each iteration must lie close to the timescale of the fastest modes of vibration present and so is usually less than 5 fs. Because the timescales of biological interest lie many orders of magnitude higher, in the micro- to millisecond range (Table 1), it is a significant engineering challenge to produce computer system able to access this range in an acceptable amount of time.

The first MD simulations on a protein were reported in 1977 for BPTI7 in vacuum for just a few picoseconds [30], while contemporary simulations are able to access microseconds [31] for all-atom systems. This dramatic increase in speed surpasses Moore's Law [6] and is accounted for in part by increasing algorithmic sophistication but mostly by the parallelisation of code to run on multiprocessor supercomputers [7–9]. Although these parallel MD codes have been spectacularly successful, they must be run on expensive, dedicated, high performance computing (HPC) resources. Given this cost, MD studies have focused on obtaining and analysing a small number of trajectories that are long enough to completely sample the process of interest.

As parallel scaling limits are approached, the future of this approach is questionable and solutions have emerged in response including the development of novel [1,10], specialised hardware [11] and new MD protocols [12–14]. In particular, modern GPUs have recently been shown to be highly capable at MD simulations, to the extent that a GPU-equipped workstation might now perform microsecond-scale simulations comparable to that of a large cluster computer [1], dramatically reducing the cost of generating MD trajectories. The peak performance of a single GPU does not yet exceed that of a Central processing unit (CPU) supercomputer except for small molecular systems, but the reduced cost and the performance trend make sampling on GPUs effective.

This is a profound qualitative change: although accessing the longest timescales remains the realm of dedicated HPC, the cost of performing 'good enough' simulations is now low. This represents a shift from MD as an expensive activity to a low cost one that can be performed on cheap commodity hardware.

In the same direction, new MD protocols based on Markov state models (MSM) are beginning to reduce the dependence on long trajectories, having been shown able to reconstruct biological processes occurring on millisecond timescales [4,12,14] from large ensembles of much shorter MD simulations. For example, Buch *et al.* [12] used an ensemble of 500 trajectories, each of 100 ns in length (50 us aggregate sampling) to estimate first-order kinetic rate constants corresponding to binding and unbinding processes occurring on timescales of approximately 440 ns and 10 us, respectively. From a statistical mechanics perspective, the shift from a single long trajectory to an ensemble is directly is equivalent to working with ensemble-, rather than time-, averages, and using biased or unbiased methods to reconstruct the ensemble information and its timescales. As a consequence, we argue that in the near future there will be more attention given to cost-effective high-throughput molecular simulations [13] where efficient aggregate sampling is favoured over long and expensive trajectories.

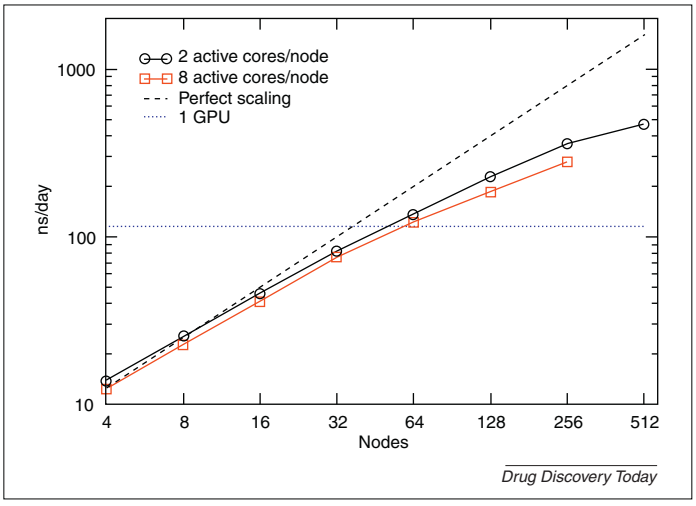

FIGURE 1

Performance of the DESMOND MD parallel code running on a high performance computing cluster [7]. The benchmark system is dihydrofolate reductase (DHFR) in explicit solvent (23k atoms). Even with a high performance network between nodes, best performance is obtained only using only two out of eight CPU cores per node, to maximise the network bandwidth/process. A more routine configuration, using all CPU cores exhibits worse scaling and is limited to using 1024 core, a limit set by the physical size of the protein. A single GPU running the fast MD code ACEMD, although not able to achieve the peak performance of DESMOND, is nevertheless able to perform comparably to 56 nodes. A two GPU-equipped compute node currently has a capital and operational cost comparable to two dual-CPU compute nodes (ca 5000 Euro, 600W) if consumer cards are used. The use of GPUs therefore represents a reduction in cost of approximately 100 times over the CPU cluster required to achieve comparable performance. *Abbreviations*: CPU, Central processing unit; DHFR, dihydrofolate reductase; GPU, graphical processing units; MD, molecular dynamics; ns, nanosecond.

Please cite this article in press as: Harvey, M.J. High-throughput molecular dynamics: the powerful new tool for drug discovery, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.03.017

Exploiting high-throughput MD

The desire to access long timescales within a reasonable time has led to the development of a variety of enhanced sampling techniques, which typically impose constraints or biases to accelerate the evolution of a system. These methods require some *a priori* knowledge about the system so that a reaction coordinate or order parameter can be defined to guide the enhanced sampling. The evolution of these techniques has been motivated by the need to extract the maximum sampling from single long trajectories, produced with high computational cost. For example, there are several methods, such as metadynamics [15], accelerated MD [16] and conformational flooding [17] that alter the normal evolution of the system with a history-dependent biasing potential along the trajectory followed by a suitably chosen set of collective variables. A different approach is represented by coarse graining (CG) which increases the accessible timescales of MD simulation in which the effective degrees of freedom of the system are reduced by coalescing atoms into aggregate particles. Although this technique has proven useful in the study of biomolecular systems [18], it comes at the expense of reduced resolution and detail, and might fail to capture subtle but important properties (e.g. H-bonding networks in solvents). As fully atomistic MD reduces in cost, relative benefit of coarse graining is distinctly lessened.

With the increasing reach of MD simulation, force field quality has become an increasing consideration. Although even the current force fields seem to work rather well in predicting affinities, kinetics and binding poses [12,19], recent work promises to improve quality even further [20,21].

In light of the new accessibility of MD, techniques that aim to sample rare events and recover kinetic and thermodynamic properties from ensembles of much shorter, unbiased, trajectories have come to prominence, perhaps the most promising of which is Markov state modelling (MSM). Markov modelling proceeds from the discretisation of the conformational space of the system of interest and the construction of a probability matrix P of the transitions between all discrete states. Transition probabilities are determined at a characteristic lag time δt, such that *Pij* is the probability of being in state *i* at time *t* and *j* at $t + \delta t$. δt is chosen to ensure that the model is memory-less (i.e. the probability of transition $i \rightarrow j$ does not depend on how the system arrived in *i*).

A sufficiently finely discretised MSM will show a separation of timescales, with high probabilities of transition among states within the same kinetic basin, and lower probabilities of transitions between kinetically distinct groups. From this model, the pathways and kinetic rates between distinct conformations may be derived (Fig. 1).

The theoretical framework for the construction and validation of MSM of molecular kinetics has been advanced by Noé *et al.* [22]. The

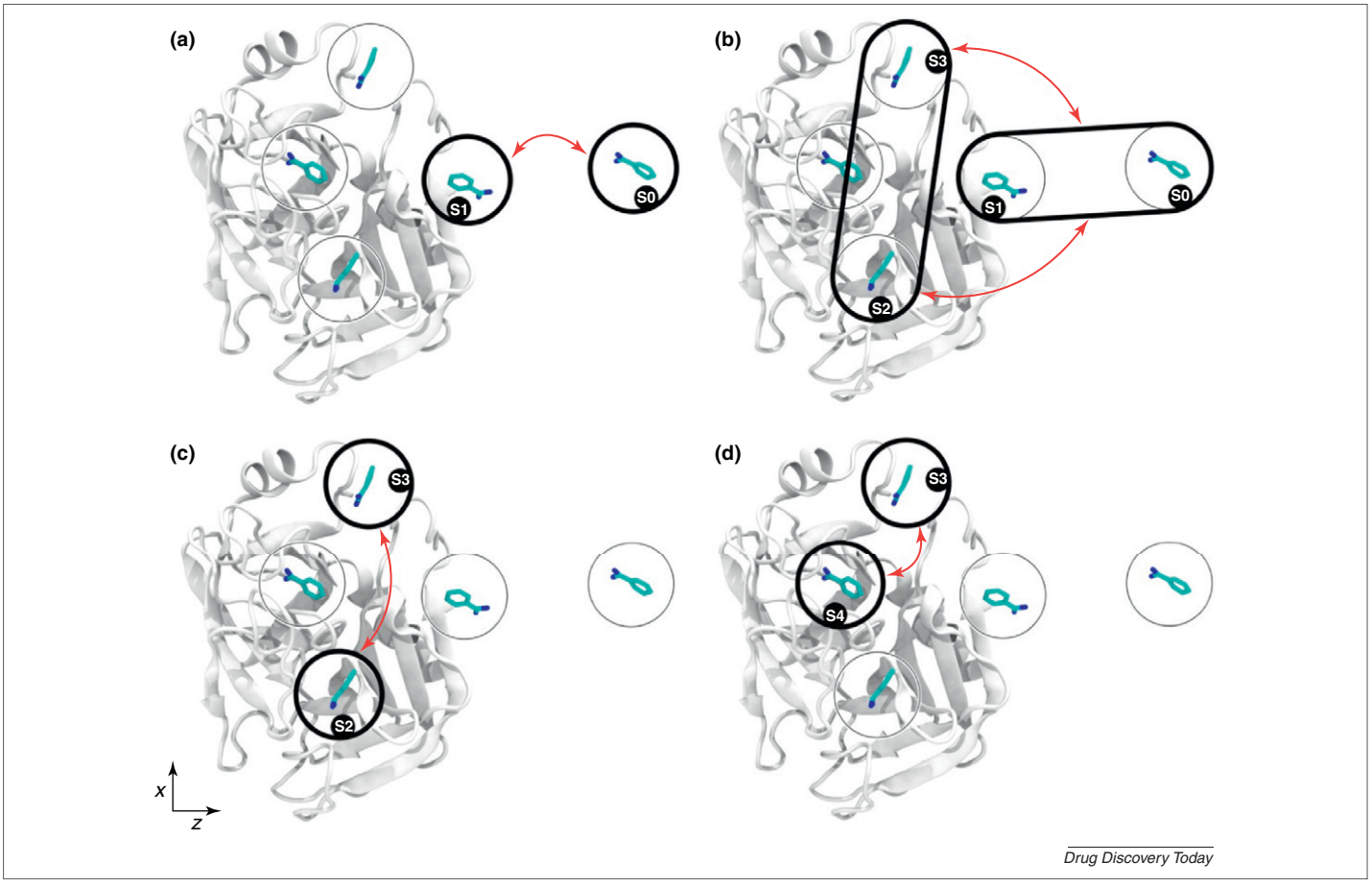

FIGURE 2

The global equilibrium of states of a system and the interconnecting kinetic pathways can be reconstructed from an ensemble of short molecular dynamics trajectories. Here, the metastable states of the binding of benzamidine to trypsin and the characteristic transition modes between them were derived from a five-state coarse-grained Markov state model built from ≈500 independent free-binding MD simulations. The ability to elucidate the modes and kinetics of full binding pathway of a ligand is a powerful tool for drug design.
Adapted from [13].

Please cite this article in press as: Harvey, M.J. High-throughput molecular dynamics: the powerful new tool for drug discovery, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.03.017

applicability of Markov modelling to re-constructing long-timescale processes was first discussed by Chodera *et al.* [23], with Pande *et al.* subsequently employing MSM to analyse the folding dynamics of a variety of small peptides [14,24,25]. In the latter work, ensembles of MD simulations were performed in a high-throughput mode on the Folding@Home distributed computing infrastructure [26]. This approach to high-throughput MD (HT-MD) has been further developed by De Fabritiis with the GPUGRID project [13]. Recently, Pronk *et al.* [27] demonstrated an integrated software framework for adaptive MSM on supercomputing resources.

Buch *et al.* [12] have demonstrated the complete reconstruction of the complete binding process of benzamidine to trypsin, a prototypical enzyme-inhibitor system, from an ensemble of unbiased MD simulations in which many complete binding events were sampled. The MSM analysis of this ensemble revealed unsuspected metastable states on the binding pathway and the characteristic transition modes between them (Fig. 2). Also quantitative estimates for k_{on}, k_{off} and ΔG_0 that compared well with experimental values were obtained. This new approach required reasonably sized computational resources equivalent to a 100 GPU cluster.

In the past year alone, Markov state modelling has been shown able to accurately and efficiently reconstruct binding processes [4,12,28] and determine pathways, kinetics and affinity for simple and more complex molecular recognition processes. As also shown in Table 1, timescales of molecular recognition are actually amenable in terms of high-throughput sampling on a medium GPU cluster, while only folding remains substantially outside of current capabilities.

Although MSM is being rapidly developed, we need to improve thorough understanding of its limitations and best practices for use have yet to be established [23,29], for example in determining the balance between length of simulation and size of ensemble and acceptable minimum length for individual simulations.

Concluding remarks

In this perspective we argue that the reduction in cost of GPUs, together with high-throughput MD protocols can provide a way to resolve the physical chemistry of molecular recognition, for instance using reasonably sized GPU clusters and high-throughput protocols. MD simulation should move from high performance computing, with its focus on maximising length of a small number of individual simulations, to high-throughput, where large ensembles of independent, 'long enough' simulations may be run. Each run is performed at maximum efficiency on available hardware (no scaling losses) with a net increase in aggregate sampling for equipment cost. In conjunction with analytical techniques and protocols, such as MSM, a robust analysis of ensemble trajectories is possible; enabling the reconstruction of global equilibrium properties and access to events that occur on timescales much longer than any one single high-performance simulation may practically reach.

Conflict of interest

The authors declare a financial interest in Acellera Ltd.

Acknowledgement

The authors thank David Soriano for a critical reading of the manuscript.

References

1 Harvey, M.J. *et al.* (2009) ACEMD: accelerated molecular dynamics simulations in the microseconds timescale. *J. Chem. Theory Comput.* 5, 1632
2 Rastelli, G. *et al.* (2009) Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chem. Biol. Drug. Dis.* 73, 283
3 Schneider, G. (2010) Virtual screening: an endless staircase? *Nat. Rev. Drug. Dis.* 9, 273
4 Silva, D.-A. *et al.* (2010) A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Comp. Biol.* 7, E1002054http://dx.doi.org/10.1371/journal.pcbi.1002054
5 Zwier, M.C. and Chong, L.T. (2010) Reaching biological timescales with all-atom molecular dynamics simulations. *Curr. Opin. Pharmacol.* 10, 745–752
6 Moore, G.E. (1965) Cramming more components on to integrated circuits. *Electronics* 38, 8
7 Chow, E. *et al.* (2008) Desmond Performance on a Cluster of Multicore Processors, DES-RES Technical Report, DE Shaw Research
8 Hess, B. *et al.* (2008) Gromacs 4: algorithms for highly efficient load-balanced and scalable molecular simulation. *J. Chem. Theor. Comput.* 4, 435
9 Phillips, J.C. *et al.* (2005) Scalable molecular dynamics with NAMD. *J. Comp. Chem.* 26, 1781
10 Giupponi, G. *et al.* (2008) The impact of accelerator processors for high-throughput molecular modeling and simulation. *Drug Discov. Today* 13, 1052
11 Shaw, D.E. *et al.* (2007) Anton, a special-purpose machine for molecular dynamics simulation. In *Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA07)* http://dx.doi.org/10.1145/1250662.1250664
12 Buch, I. *et al.* (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl Acad. Sci. U. S. A.* 108, 10184–10189
13 Buch, I. *et al.* (2010) High-throughput all-atom molecular dynamics simulations using distributed computing. *J. Chem. Inf. Model.* 50, 397–403
14 Voelz, V.A. *et al.* (2001) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* 132, 1526–1528
15 Laio, A. and Gervasio, F.L. (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* 71, 12
16 Voter, A.F. (1997) Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* 78, 3908–3911
17 Hamelberg, D. *et al.* (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120, 11919–11929
18 Yao, X.Q. *et al.* (2010) Drug export and allosteric coupling in a multidrug transporter revealed by molecular simulations. *Nat. Commun.* 16, 117
19 Piana, S. *et al.* (2011) How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100, L47–L49
20 Buch, I. *et al.* (2011) Optimised potential of mean force calculations of standard binding free energy. *J. Chem. Theor. Comp.* 7, 1765–1772
21 Lindorff-Larssen, K. *et al.* (2001) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78, 1950–1958
22 Prinz, J.H. *et al.* (2011) Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134, 174105
23 Chodera, J.D. *et al.* (2006) Long-time protein folding dynamics from short-time molecular dynamics. *Multiscale Model. Sim.* 5, 1214
24 Elmer, S.P. *et al.* (2005) Foldamer dynamics expressed via Markov state models {I}. Explicit solvent molecular-dynamics simulation in acetonitrile, chloroform, methanol and water. *J. Chem. Phys.* 123, 114902
25 Bowman, G.R. *et al.* (2009) Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* 49, 197–201
26 Beberg, A. *et al.* (2009) Folding@Home: lessons from eight years of volunteer distributed computing. IPDPS'09. In *Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing, IEEE Computer Society 1-8* http://dx.doi.org/10.1109/IPDPS.2009.5160922
27 Pronk, S. *et al.* (2011) Copernicus: a new paradigm for parallel adaptive molecular dynamics. *SC'11 Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis* http://dx.doi.org/10.1145/2063384.2063465
28 Held, M. *et al.* (2011) Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys. J.* 100, 701–710
29 Noé, F. *et al.* (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19011
30 McCammon, J.A. *et al.* (1977) Dynamics of folded proteins. *Nature* 267, 585–590
31 Roy, J. and Laughton, C. (2009) Long timescale molecular dynamics simulations of the major urinary protein provide atomistic interpretations of the unusual thermodynamics of ligand binding. *Biophys. J.* 99, 218–226

Matthew J. Harvey[1]
Gianni De Fabritiis[2]
[1]*High Performance Computing Service, Imperial College London, South Kensington, London SW7 2AZ, UK*
[2]*Computational Biochemistry and Biophysics Lab (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Carrer Doctor Aiguader 88, 08003 Barcelona, Spain*

Please cite this article in press as: Harvey, M.J. High-throughput molecular dynamics: the powerful new tool for drug discovery, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.03.017