Estimating computational limits on theoretical descriptions of biological cells

Roland R. Netz[a,1] and William A. Eaton[b,1]

[a]Fachbereich Physik, Freie Universität Berlin, 14195 Berlin, Germany; and [b]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, NIH, Bethesda, MD 20892

Contributed by William A. Eaton, December 29, 2020 (sent for review November 3, 2020; reviewed by Dmitrii E. Makarov and Stefano Piana-Agostinetti)

There has been much success recently in theoretically simulating parts of complex biological systems on the molecular level, with the goal of first-principles modeling of whole cells. However, there is the question of whether such simulations can be performed because of the enormous complexity of cells. We establish approximate equations to estimate computation times required to simulate highly simplified models of cells by either molecular dynamics calculations or by solving molecular kinetic equations. Our equations place limits on the complexity of cells that can be theoretically understood with these two methods and provide a first step in developing what can be considered biological uncertainty relations for molecular models of cells. While a molecular kinetics description of the genetically simplest bacterial cell may indeed soon be possible, neither theoretical description for a multicellular system, such as the human brain, will be possible for many decades and may never be possible even with quantum computing.

biological uncertainty principle | biological complexity | computational limits

In 1966 Crick asserted: "The ultimate aim of the modern movement in biology is to explain all biology in terms of physics and chemistry (1)." Our interpretation of Crick's comment is that explanation at a deep level will come from physics and chemistry theory. With the advent of computers and their use as a major tool in scientific research, theorists have come to rely more and more on simulations to understand physical, chemical, and biological systems, rather than on analytical models. Simulations are no longer just used to check the range of validity of theoretical equations, but to understand experimental results for systems that are too complex to be amenable to treatment by analytical models. The challenge, of course, for biologically oriented theorists is not to simply run simulations, such as molecular dynamics, but to gain an improved understanding of how a particular biological system works from an insightful analysis of the trajectories.

By concentrating on a set of global variables that are assumed or known to be important for the functioning of a cell, such as mean concentrations of metabolites, RNA, and proteins, coarse-grained cell dynamic descriptions are currently becoming available from developments in the field of system biology and compare very favorably with experimental data (2). The computational burden of such coarse-grained simulations is straightforwardly manageable by modern computers, but the selection of the global variable, which is crucial, and the neglect of spatial concentration variations may influence the results obtained. In contrast, simulations that resolve molecular and spatial detail do not rely on choosing relevant coarse-grained degrees of freedom beforehand, but pose fundamental problems in terms of the computational effort. Motivated in part by the recent success of simulating subcellular biological organelles and the goal of simulating whole cells with molecular resolution (3, 4), the question arises of how complex can a system be before it is no longer possible to be simulated on a computer. To make a quantitative assessment of the level of complexity that can be simulated at the molecular level, we consider here the two main molecular methods currently being employed (3, 4), namely the physics and chemistry methods of molecular dynamics simulations and the probabilistic methods of solving molecular kinetics equations. We consider the simplest cell capable of reproducing itself, the bacterium *Mycoplasma genitalium*, and the most important and most interesting multicellular system, the human brain.

Results

A force-field–based molecular dynamics simulation consists of solving Newton's equations of motion for every atom of every molecule using simplified, empirical interatomic energy functions to determine the position and velocity of every atom in the system as a function of time. These molecular mechanics energy functions do not consider the electronic degrees of freedom, which are essential for describing the making and breaking of covalent bonds or electron transfer, as occur for the many chemical reactions in a biological cell. Chemical reactions require the use of quantum mechanics, which is much more time-consuming because the interatomic forces must be obtained by solving Schrödinger's equation rather than from empirical functions. Fortunately, the quantum-mechanical calculations are only necessary for the atoms of the active sites of the enzymes and their bound substrates. The remainder of the protein, the unbound substrates, and the aqueous solvent, can be simulated with acceptable accuracy using force fields. This combination of quantum mechanics and molecular mechanics (QM/MM) for proteins originated with the work of Warshel and Levitt (5) and is now extensively used (6).

The approximate computational time (T_{CPU}) for a QM/MM simulation of a system of N atoms, which contains a total of p regions with n atoms each that are treated quantum mechanically, can be written as

Significance

This work addresses a new and important question concerning biological complexity. What is the level of biological complexity that can be theoretically described in molecular detail? Simple algebraic equations are presented that determine the computational time required to simulate simplified molecular models for a single cell, the genetically simplest bacterium, and the most important and interesting multicellular system, the human brain. The results place limits on when, if ever, such theoretical descriptions will be possible.

Author contributions: R.R.N. and W.A.E. designed research, performed research, and wrote the paper.

Reviewers: D.E.M., The University of Texas at Austin; and S.P.-A., D. E. Shaw Research.

The authors declare no competing interest.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: rnetz@physik.fu-berlin.de or eaton@nih.gov.

Published January 25, 2021.

$$T_{CPU} \approx T_{CPU}^{MM} + T_{CPU}^{QM} \approx \left(\alpha N \ln N + \beta p n^3\right) \frac{T_{real}}{\Delta \nu}, \qquad [1]$$

where T_{real} is the time over which the process is modeled, Δ is the time discretization step, and ν is the computer speed in floating-point operations per second (flops). The classical contribution (first term) is dominated by calculating the electrostatic interactions between the partial charges on all the atoms of the protein, which scales as $\alpha\ N\ \ln\ N$, where the numerical prefactor that counts the number of floating-point operations, α, is on the order of 10 (7). The quantum-mechanical part (second term) is on the density-functional level and is dominated by the effort required to diagonalize the Hamiltonian, which scales as the cube of the number of atoms n in the quantum regions of the molecules. The numerical prefactor, β, is on the order of $\sim 10^4$.* Importantly, both terms in Eq. **1** essentially scale linearly in system size, i.e., linearly in N or p.

A *M. genitalium* cell contains a total of $\sim 3 \times 10^9$ atoms [spherical cell volume of radius of 0.2 μm (12) times the atom number density of water of 10^{11} atoms per μm^3]. Of the ~77,000 protein molecules in the cell (12), ~26,000 ($= p$) (12) are enzymes with active sites. Assuming $n = 100$ atoms, $\Delta = 10^{-15}$ s, and $\nu = 10^{17}$ flops (the speed of the currently fastest supercomputer: the "Summit" at Oak Ridge), T_{CPU} for the 2-h doubling time (12) of the bacterium is $\sim 10^9$ y, where >99% of the computation time is for the QM part. Although such a QM/MM calculation cannot be performed now, will it be possible in the future? Over the past 25 y, the speed of supercomputers has increased roughly 10-fold every 5 y, as predicted by Moore's law (13). However, it is not clear whether computer speed will continue to increase exponentially at this rate for the next ~50 y, which would be needed to shrink the computational time down to a month.

The most important and most interesting multicellular system is, of course, the human brain, where an ultimate goal of science is to understand thinking, memory, and behavior. Given a particular stimulus, for example, an accurate simulation may be able to explain or predict a response. This multicellular system contains $\sim 10^{11}$ neurons (14), $\sim 10^{11}$ proteins per neuron (15), and an estimated $\sim 10^{26}$ atoms for the average human brain of 1,200 cm^3 calculated as above (1.2×10^{15} $\mu m^3 \times 10^{11}$ atoms per μm^3), so the situation is quite different. Using a conservative guess that the active site complexes of only 10^9 of the 10^{11} proteins in the average neuron must be treated quantum mechanically, the calculation of the quantum-mechanical part for 1 h would take $\sim 10^{24}$ y and $\sim 10^{23}$ y for the Newtonian part. It seems unlikely that computer speed will continue to increase at the same rate for the next 125 y (after which a brain QM/MM simulation could be done in a month). We are, therefore, forced to conclude that, while an atomistic molecular dynamics simulation including quantum effects of a single bacterial cell may be possible in this century, such simulations of a human brain for even 1 h will not be possible until much later and may never be possible. Even if quantum computing could be adapted for molecular dynamics calculations, an enormous speed-up would be needed in order for such simulations to be performed in a reasonable time (16).

An alternative, albeit much more approximate, approach to the problem is a description of cells by treating them at a probabilistic rather than explicit particle level. In such description, only the spatial coarse-grained probability distribution of each type of molecule as a function of time is considered (3). Because molecules can diffuse from one part of a cell to another to chemically react or simply bind to another molecule, it is necessary to solve a set of partial differential equations, called the reaction–diffusion Master equation. The solution to this equation yields the probabilities of finding the number of each molecular type—protein, bound complex, lipid, nucleic acid, metabolite, ion, etc.—at a given position in a cell as a function of time. A rough estimate of the time (T_{CPU}) required to solve the reaction–diffusion Master equation by simulating it as molecules jumping between subvolumes (voxels) on a lattice mesh and reacting within the voxels reads

$$T_{CPU} \approx \frac{\gamma T_{real} m M K \left(L/l\right)^3}{\Delta \nu}, \qquad [2]$$

where again T_{real} is the time over which the process is modeled, M is the number of different reactive molecular species in a cell treated as a bag of molecules, m is the typical number of specific and nonspecific possible reactions per molecular species, L is the linear cell size, l is the spatial discretization size needed to accurately describe the concentration profile of each different species, K is the maximum copy number of each species per discretization volume element (voxel) treated in the Master equation, Δ is the time step in simulating the Master equation, and v is the computer speed in flops. The numerical prefactor γ is on the order of 10^2 and accounts for the computational expense of one iteration step.†

The number of different molecular species in a *Mycobacterium genitalia* bacterial cell (M) is at least ~500 (12), which is the number of different proteins and does not include small molecules, posttranslationally modified proteins, or complexes that would have to be treated as separate species in the reaction–diffusion Master equation. The mean number of reactions per molecular species can be estimated as $m = 10$. To obtain the concentration profile for this cell with $L = 400$ nm, a discretization of $l = 10$ nm can be used with $K = 1{,}000$ copy numbers in each voxel as a safe upper bound. To account for the fastest unimolecular and

*In current ab initio molecular dynamics simulation schemes that employ state-of-the-art density-functional theory, the most time-consuming step is the orbital transformation, which scales as RQ^2 with the number of occupied orbitals Q and the number of basis functions R (8), which themselves can be assumed to scale with the number of atoms as $Q = \gamma n$ and $R = \varepsilon n$, where the numerical coefficients are on order of $\gamma \sim 10$ and $\varepsilon \sim 10$. Assuming, furthermore, that the number of iteration steps in the self-consistent field determination is on the order of 10, the numerical prefactor in Eq. **1** becomes $\beta \sim 10\varepsilon\gamma^2 \sim 10^4$, which could be multiplied by a factor that accounts for the numerical effort in each step. We note that approaches that utilize sparse matrix properties and scale linearly in n are currently becoming available (8). Also, adaptive QM/MM methods are under development and could be used to treat reactive regions quantum mechanically only during the fraction of time that a chemical reaction is occurring, which depends on diffusion times and on the lifetime of the bound substrate (9). From Eq. **1**, it transpires that such approaches are particularly needed to reduce the overall computation time if the QM part dominates the total CPU time, which is true for $\beta p n^3 > \alpha N \ln N$. Likewise, coarse-graining approaches that lump a number of atoms into effective particles or molecules and that replace water by a continuum medium, characterized by a dielectric constant and a viscosity, reduce the time spent on the classical MM part and can cope with reactions in a heuristic fashion (10), but also introduce approximations the reliability of which is less clear. While the accuracy of current force fields is at a level where protein folding and supermolecular aggregation processes can be faithfully modeled (11), it is clear that systematic errors in the resulting free energies remain and will sensitively perturb reaction kinetics. Likewise, density functionals that are currently employed in ab initio simulations entail a number of approximations and neglect electron dynamics and excited states but are subject to ongoing improvements. Lastly, in order to estimate the dependence of simulation results on initial conditions, many independent simulation runs would have to be performed, so that our formula would be multiplied by an additional factor that accounts for the number of simulation runs. All these factors constitute corrections but do not principally invalidate our formula.

†Eq.2 is derived from the fact that in one time step of the iterative solution of the reaction–diffusion Master equation, one updates the probability of each copy number (contributing a factor K) of each molecular species (factor M) in each voxel [factor $(L/l)^3$] by summing over all reactions of that species (factor m). By only tracking the local mean concentrations, the reaction–diffusion Master equation becomes equivalent to the chemical reaction–diffusion equation, which is characterized by $K = 1$. The largest shortcoming of reaction–diffusion Master equation approaches compared to particle-based methods is that molecules are treated as noninteracting pointlike particles, except mean-field interactions that are averaged over voxels or when particles react to form separate species. This precludes treatment of extended molecules with conformational degrees of freedom such as DNA or molecular aggregates such as lipid bilayers. Needless to say, hybrid approaches that treat some molecules in a cell on a particle level and the rest on a stochastic level are conceivable.

bimolecular reactions, a time step of $\Delta = 1$ μs may be sufficient. With our highly oversimplified model that considers a bacterial cell as a bag of ∼500 different molecular species, the time (T_{CPU}) required from Eq. **2** for the Oak Ridge computer to simulate a single bacterial for its 2-h doubling time and the above parameters is roughly 1 mo. Therefore, the limiting factor is not the computational time, but is determined by the time required to experimentally or theoretically determine accurate forward and reverse rate coefficients for all relevant chemical reactions and intermolecular interactions in the cell.

Simulating a human brain with $\sim 10^{11}$ neurons (14) is again a wholly different matter. Using the bacterial values for parameters other than $M = 4{,}000$ (17) and $L = 10$ μm (15), the computation time from Eq. **2** with current computing power is increased by a factor of ∼10 for the larger number of different proteins, a factor of 10^{11} for the number of neurons compared to a single bacterial cell, and by an additional factor $(25)^3 \sim 10^4$ for the difference in L to give $T_{CPU} \sim 10^{15}$ y for a 1-h simulation. Consequently, a 1-mo calculation for an enormously oversimplified treatment of the brain for 1-h real time as a collection of bags of molecules would not be practical for about 80 y (again using Moore's law, which will not necessarily hold for the next 80 y). So, as with the molecular dynamics calculations, we conclude that, with a realistic model, simulating the brain at the molecular level with a reaction–diffusion Master equation may not happen for a very long time and may never be possible because of both limits on computational capability and the determination of rates for all processes for a realistic model of the brain.[‡]

Discussion

Our estimates of the computational time required to simulate highly simplified models of cells indicate that the simplest bacterial cell may be theoretically described in the not too distant future by solving molecular kinetics equations. Simulation of this cell by molecular dynamics calculations will take much longer, and would be feasible in ∼50 y if Moore's law continues to hold. However, our estimates suggest that simulation of a multicellular system such as the human brain may never be possible, presumably even if quantum computing could be adapted for such calculations. We must emphasize that simulating parts of biological cells has (3, 4) and will continue to yield extremely valuable information. On the other hand, enormously important insights can be obtained on how any cell or multicellular system functions, including the human brain, by coarse graining or alternative descriptions that do not include molecular detail. In fact, many researchers share the view that the most important advances in theoretical understanding of complex biological systems will not come from the detailed molecular simulations we have discussed in this work, but from the discovery of collective organizing principles that may be independent of such microscopic details (18, 19). Moreover, new theoretical approaches arising from the growing field of systems biology may also provide answers to important questions that are unanswerable by the methods we have discussed (20).

In physics, the Heisenberg uncertainty principle places limits on the precision in determining pairs of values for a single particle, such as position and momentum. The notion of uncertainty in biology is quite different (21) and there is, as yet, no biological uncertainty principle, although the well-known inability to determine the exact nucleotide sequence of a cell's genome due to experimental errors has been proposed as a biological one (22). In biology, of much greater importance is understanding how cells function, which immediately raises the question we have addressed here of what level of complexity can possibly be simulated on a computer. Eqs. **1** and **2** place limits on the complexity of cells that can be understood with current theoretical molecular methods. They therefore represent a first step in developing equations for realistic molecular models that could be considered biological uncertainty relations. Hopefully, the work presented here will stimulate thinking on the larger issue of formulating a comprehensive biological uncertainty principle.

Data Availability. All study data are included in the article text.

ACKNOWLEDGMENTS. We thank Maximilian Becker, Florian Brünig, Gerhard Hummer, Martin Karplus, Philip Loche, Rob Phillips, Igor Schapiro, and Attila Szabo for helpful advice and discussion. This work was supported by the Deutsche Forschungsgemeinschaft via Grant SFB 1114, Project B3, Project 235221301; by the European Research Council via an Advanced Investigator Grant, Grant Agreement 835117; and by the intramural program of the National Institutes of Diabetes and Digestive and Kidney Diseases of the NIH.

[‡]Clever programming techniques that replace some of the copy number probability distributions by mean copy numbers (which is permissible if copy numbers are much larger than unity) and stochastic simulation algorithms based on Gillespie methods will certainly bring down our computational time estimates, but not change our conclusion significantly.

1. F. Crick, *Of Molecules and Man* (University of Washington Press, 1966).
2. J. R. Karr *et al*., A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
3. T. M. Earnest, J. A. Cole, Z. Luthey-Schulten, Simulating biological processes: Stochastic physics from whole cells to colonies. *Rep. Prog. Phys.* **81**, 052601 (2018).
4. A. Singharoy *et al*., Atoms to phenotypes: Molecular design principles of cellular energy metabolism. *Cell* **179**, 1098–1111.e23 (2019).
5. A. Warshel, M. Levitt, Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249 (1976).
6. M. Valiev *et al*., NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **181**, 1477–1489 (2010).
7. U. Essmann *et al*., A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
8. V. Weber, J. VandeVondele, J. Hutter, A. M. N. Niklasson, Direct energy functional minimization under orthogonality constraints. *J. Chem. Phys.* **128**, 084113 (2008).
9. J. M. Boereboom, R. Potestio, D. Donadio, R. E. Bulo, Toward Hamiltonian adaptive QM/MM: Accurate solvent structures using many-body potentials. *J. Chem. Theory Comput.* **12**, 3441–3448 (2016).
10. H. I. Ingólfsson *et al*., The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 225–248 (2014).
11. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
12. M. Breuer *et al*., Essential metabolism for a minimal cell. *eLife* **8**, e36842 (2019).
13. G. E. Moore, Cramming more components into integrated circuits. *Electronics (Basel)* **38**, 114–117 (1965).
14. F. A. C. Azevedo *et al*., Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* **513**, 532–541 (2009).
15. R. Milo, R. Phillips, *Cell Biology by the Numbers* (Garland Science, New York, 2016).
16. S. Lloyd, Ultimate physical limits to computation. *Nature* **406**, 1047–1054 (2000).
17. D. Polioudakis *et al*., A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785–801.e8 (2019).
18. R. B. Laughlin, D. Pines, J. Schmalian, B. P. Stojkovic, P. Wolynes, The middle way. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 32–37 (2000).
19. R. Phillips, Musings on mechanism: Quest for a quark theory of proteins? *FASEB J.* **31**, 4207–4215 (2017).
20. M. Bizzarri, A. Palombo, A. Cucina, Theoretical aspects of systems biology. *Prog. Biophys. Mol. Biol.* **112**, 33–43 (2013).
21. J. P. Zbilut, A. Giuliani, Biological uncertainty. *Theory Biosci.* **127**, 223–227 (2008).
22. P. Strippoli *et al*., Uncertainty principle of genetic information in a living cell. *Theor. Biol. Med. Mod.* **2**, 40 (2005).