Computational Molecular Dynamics in Emerging Biological Fields

Joseph Middleton (a), Joan Cordiner (a)

(a) Chemical and Biological Engineering, University of Sheffield, S10 2TN

j.cordiner@sheffield.ac.uk

Abstract

Due to the pandemic, mRNA-based vaccines and therapeutics are now a widely tested technology rather than just emerging. It is an efficient way of delivering genetic information to the target, making use of human cellular transcription enzymes to convert the mRNA strands into proteins that can then perform their intended functions in vivo. Traditional methods of process discovery, which require exploring the process conditions experimentally, would be prohibitively expensive meaning instead we must use modelling to construct a digital twin of the process from first principles, or *Ab Initio*. This can be achieved using Molecular Dynamics, which sits at the intersection between the fields of Quantum Chemistry and Classical Mechanics; with quantum operations being expressed directly as combinations of the quantum mechanical operator and classical (Hamiltonian) Mechanics.

Similar to classical mechanics there is a limited set of analytically solvable problems, meaning that the multi-dimensional, many-body problems must be solved numerically. Even using approximations, numerical solutions take a significant amount of time to converge. This paper reviews the current state of molecular modelling, through the practice lens of modelling the mRNA transcription process. The paper will discuss the theory behind, the mathematics and the physical computation of the process and will make suggestions as to what the next stage of computational molecular dynamics might look like and what might be required in order to make large scale biological simulation more accessible.

**Keywords:** mRNA, Quantum Dynamics, Computing

* 1. Quantum Chemistry

Quantum Dynamics, which was reformulated by Schrödinger in 1926, is based around Equation 1. A quantum evolution operator for the wave equation, Ψ, which takes the form of a standard eigen-problem (Galler et al., 2021). H, is the Hamiltonian operator, which represents the total sum of energies in the system. Similar to the classical interpretation, when describing a quantum system, T is the kinetic energy, dependent on the momentum of the particle, and V is the potential energy of the system, granted by an external field. In order to accurately represent a particle system in 3D space H, becomes Equation 2 (Arnolʹd et al., 1989). In the equation, the Hamiltonian is time dependent though it does not have to be. In addition, if the time element of the Schrödinger equation is separable, then the wave function becomes stationary as it represents the probability of a particle’s position in space.

|  |  |
| --- | --- |
|  | (1) |

|  |  |
| --- | --- |
|  | (2) |

The Schrödinger formulation is not the only interpretation of how a wave function can propagate, by using the Heisenberg or Interaction formulation as seen in Eugen Merzbacher (1998), the view of the wave system can change. In the interaction formulation the wave equation propagation can be broken down into discrete time-steps, allowing for easier understanding and analysis.

Because the set of analytically solvable solutions to the Schrödinger equation is limited, numerical solutions must be used to calculate the results to non-trivial molecules. All of the particles, both electrons and nuclei, interacted within the simulation scaling the computational complexity required at least quadratically. The Time-Independent or Time-Dependent Schrödinger can be used depending on the type of simulation required. Time-Dependent models allow for bonding and other dynamic processes to be simulated, whereas using the Time-Independent equation produces a stationary wave-function that varies only with position, Dick (2020, pp. 265–310).

In order to make simulations more accessible and allow for larger, more complex molecules, the computational complexity must be reduced. The most well-known method is the Born-Oppenheimer approximation (Born & Oppenheimer, 1927), which uses the momentum basis of the Schrödinger Equation to separate the nuclei and electrons into two wave-functions, which can then be propagated whilst the other remains fixed. Since the electrons contribute most significantly to the simulation complexity, solving the electronic wave-function through approximation will decrease computation times and increase the maximum simulation size possible.

The nature of Equation 1 is that both the wave-function and constants are undetermined; before attempting to solve a problem the wave-function for the quantum system must be found. Generating an approximate wave-function that correlates well with the analytical solution can be found through perturbation or variational methods (Nayfeh, 2011). Achieved by either perturbing one of the analytically solvable problems, in the case of an electronic wave-function this will be a single electron Schrödinger, or by combinations of basis sets as set out by Roothaan (1951). This and the use of a Self-Consistent Field, Hartree-Fock, approach with a Slater determinant (Slater, 1929) allows for iterative convergence for better approximate answers. Now, in the age of fast computing, the possibility of utilising Post Hartree-Fock methods such as Density Field Theory (DFT) and valence band theory can be fulfilled, these methods can be tuned to a variety of tasks, from full-scale ab-initio quantum dynamics to stationary bond energy predictions.

* 1. Molecular Dynamics & Mechanics

Full molecular modelling can be used when attempting to study the makeup of molecules and their behaviour in different environments. This takes a molecular potential field generated and uses it to propagate the atoms whilst being constrained by the molecule’s bonds and other interactive forces.

Methods to generate molecular potential fields generally take three forms: Molecular Dynamics, being Quantum Chemistry extended over molecules and their interactions; Molecular Mechanics using a macro-mechanical interpretation of bond mechanics, energies and inter-molecular forces to generate potentials; and Semi-Empirical methods, which only utilise full quantum chemistry in certain parts of the molecule, including just the outer valence electrons or only certain physical locations of interest.

Though computationally cheaper molecular mechanics methods have significant drawbacks, including requiring extensive parametrisation and are restricted to ground-state simulations Levine (2014, pp. 634–654). Generally, all the parameters of the models need to be found empirically, requiring models to interpolate or extrapolate to the conditions of the simulation. Quantum chemistry, or *full ab-initio* methods, which take a variety of forms, are more computationally complex. Which becomes problematic when the potential of the molecule must be calculated each timestep. For this cost, they generate accurate approximate solutions, including information on the electronic properties of the material, meaning that these methods are capable of modelling covalent bond reactions. Semi-empirical simulations present a good balance of accuracy and speed. In these parts of the molecular structure which require complex reaction modelling use Quantum Chemistry methods while the rest can be treated with less complex methods, such as Molecular Mechanics (Warshel & Levitt, 1976).

All of these methods can be improved by implementing adaptive resolution, in which parts of the molecular structure that influence the molecule as a group can be summed into a single large pseudo-atom to reduce the number of interactions, possibly improving calculation speed (Kmiecik et al., 2016). During long simulations the size of the time-step contributes significantly to the computation required with larger time-steps reducing the number of costly potential calculations. In order to ensure this does not cause simulation decoherence, fast vibrations of light nuclei can be damped (Tuckerman et al., 1991) whilst still potentially allowing them to quantum tunnel.

As these simulations do not take place in a vacuum, the effects of the solvent must be taken into consideration in order to accurately study how large molecules behave. In continuum solvent models, the solvent potential fields are averaged and applied to the molecule of interest through the external potential field, Equation 2. Like the molecule being studied, the method of deriving the external potential can be of a quantum or classical nature, with classical solutions being tuned for specific solvents. In contrast, the quantum chemical methods are more generally applicable (Levine, 2014, pp. 634–654). Alternative models are able to predict how the physical properties of the solvent affect the molecule directly.

* 1. mRNA and Biological Applications

In eukaryotic cells, mRNA, or messenger RNA, is the encoded form of DNA transcribed by a polymerase in the nucleus of a cell where it has its on-coding regions spliced, is capped and has a poly(A) tail added before being transported into the cells’ cytoplasm ’for reading by a ribosome (Cooper & Adams, 2023, pp. 263–278). mRNA-based vaccines and therapeutics exploit this process in order to confer immunity or immune responses to various threats, either by delivering mRNA directly into the cells cytoplasm or by using viral vectors directly into the nucleus of the cell. Beyond both vectors utilising human cellular transcription and translation to provoke an immune response, they are not similar. Viral vectors may cause host-genome integration and have provoked adverse reactions in recipients; it is by this process that the conditions of an active infection can be simulated, provoking a strong immune response (Ura et al., 2014). In order to facilitate direct mRNA delivery, lipid nanoparticles can be used, which are taken up most readily by dendritic cells, which then translate the mRNA stored within into antigens and communicate them with T-cells, generating an immune response (Kliesch et al., 2022).

The presence and general understanding of mRNA-based therapies has increased dramatically due to the pandemic in which the technology developed vaccines at unprecedented speed, and the widespread deployment of those vaccines has seen the technology tested on a scale not often seen (Wherry et al., 2021). The enormous interest and large quantity of funding available for mRNA-based products has further spurred development in wider applications such as disease mitigation or cancer treatments.

Unlike viral based products, which can be produced in cell-based reactors (Ura et al., 2014), products that are transcribed raw within the reactor setup require external control to ensure the purity, and therefore safety, of the product. In order to model the transcription process, the rate kinetics, side products and promoters of the reaction must be understood. Given the significant cost of experimenting with mRNA in a research environment, the challenge of observing products, and the large number of experiments required to construct an accurate digital twin of the process, finding an alternative way of understanding and predicting process properties is desirable. In understanding the initial rate kinetics of the polymerase, typically T7rna, the development of any reactor would be significantly sped up. Future processing steps, such as lipid encapsulation, could also be modelled depending on the modelling methods. Correctly understanding and modelling these processes will increase production efficiency and decrease potential time to market.

* 1. Hardware and Software Implementations

When assessing the calculation speed for a problem there exists no standard metric. For example, for programs that do not rely on parallelisation, the maximum operating frequency matters considerably more than the number of simultaneous cores present in the processor, with the opposite also being true. The instruction set architecture, generally register-memory in modern processors (Flynn, 1998), largely determines calculation speed by the efficiency of its physical instructions and how long it takes to access and return the data to registers, the best case, or memory required in the calculation, the worst case. In-between these memory layers exist processor caches, which have fast access times due to their physical architecture and proximity to the processor, reducing the amount of power and time wasted waiting for memory.

Ensuring as much required program data is in the cache can lead to performance increases by reducing memory fetch time. Generally, this can be achieved by storing data that is likely to be used together sequentially in arrays the size of the target cache. The desired precision of the calculations can have a significant impact on performance, as performing more precise floating-point logic takes longer. This can be addressed by choosing an acceptable level of accuracy for the entire program or specifically for time-consuming or repetitious calculations.

Though the individual time savings from each optimisation are very small, on the scale of single nanoseconds, the overall impact after calculating each particle-particle interaction over a representative large molecule is significant. For example, within T7rna a simple lab transcription polymerase, there are 6802 non-hydrogen atoms (Yin & Steitz, 2002), that must be accounted for in at most in each timestep. Running calculations on a single processor, even with optimisations, is still limiting; calculation throughput can be dramatically increased by parallelising the workflow through partitioning, though at the cost of complexity. This can be partially achieved within the processor or within processor groups, such as graphics processing units, by using SIMD instructions in which single instructions act on multiple pieces of data. Whereas fully horizontal parallelisation requires inter-processor networking and careful partitioning to avoid simulation decoherence.

Moving away from general purpose computing hardware towards specialised machines such ANTON, an application-specific integrated circuit (ASIC) based MD machine built by Shaw et al. (2008) capable of simulating 17 μs of time per day presents a way forward. Instead of utilising ASICs, field programmable gate arrays (FPGA) present a cheaper and more dynamic approach to solving MD simulations by supporting field reprogramming whilst allowing simulation logic to be integrated directly into the hardware. Because of this flexibility, FPGAs generally perform logic operations slower than ASICs, though the gap can be closed by utilising embedded hardware units. However, both are significantly more power efficient than performing the same calculations on general-purpose hardware (Betkaoui et al., 2010). Making this dedicated hardware approach more accessible would be a novel step to creating more performant molecular dynamics simulations.

* 1. Conclusions

Developing new methods of modelling large biological molecules accurately, benefits not only the manufacturing of mRNA therapies, but also the modelling of new designs and understanding the why and how of potential adverse reactions. Rather than performing calculations on more general-purpose compute hardware, time and money can be saved by using a more specialised hardware. This is by no means a new approach; existing MD systems already exist, albeit not in a widely available manner, and were not developed recently. Dedicated hardware is already in use in processes like digital signal processing, which are required to keep up with the high speed networking that our computational infrastructure requires to operate. The potential computational and cost benefits are real and can be used to enable reaction modelling and deepen our understanding of large pharmaceutical molecules which are becoming important in our global health.

References

Betkaoui, B., Thomas, D. B., & Luk, W. (2010). *Comparing performance and energy efficiency of FPGAs and GPUs for high productivity computing*. 94–101. https://doi.org/10.1109/FPT.2010.5681761

Born, M., & Oppenheimer, R. (1927). Zur quantentheorie der molekeln. *Annalen Der Physik*, *389*, 457–484. https://doi.org/10.1002/andp.19273892002

C, S. J. (1929). The theory of complex spectra. *Physical Review*, *34*, 1293–1322. https://doi.org/10.1103/physrev.34.1293

Cooper, G. M., & Adams, K. W. (2023). *The Cell: A Molecular Approach* (9th ed., pp. 263–278). Oxford University Press.

Dick, R. (2020). Time-dependent perturbations in quantum mechanics. In *Graduate texts in physics* (pp. 265–310). Springer International Publishing. https://doi.org/10.1007/978-3-030-57870-1\_13

Eugen Merzbacher. (1998). *Quantum mechanics*. Wiley.

Flynn, M. J. (1998). *Computer architecture pipelined and parallel processor design*. Jones and Bartlett Publishers Inc.

Galler, A., Canfield, J., & Freericks, J. K. (2021). Schrodinger’s original quantum-mechanical solution for hydrogen. *European Journal of Physics*, *42*, 035405. https://doi.org/10.1088/1361-6404/abb9ff

I, A. V., Vogtmann, K., & Weinstein, A. (1989). *Mathematical methods of classical mechanics*. Springer.

Kliesch, L., Delandre, S., Gabelmann, A., Koch, M., Schulze, K., Guzmán, C. A., Loretz, B., & Lehr, C.-M. (2022). Lipid–Polymer Hybrid Nanoparticles for mRNA Delivery to Dendritic Cells: Impact of Lipid Composition on Performance in Different Media. *Pharmaceutics*, *14*(12), 2675–2675. https://doi.org/10.3390/pharmaceutics14122675

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, *116*(14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Levine, I. N. (2014). *Quantum chemistry* (pp. 634–654). Pearson.

Nayfeh, A. H. (2011). *Introduction to perturbation techniques*. John Wiley & Sons.

Roothaan, C. C. J. (1951). New developments in molecular orbital theory. *Reviews of Modern Physics*, *23*, 69–89. https://doi.org/10.1103/revmodphys.23.69

Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J., Chao, J. C., Eastwood, M. P., Gagliardo, J., Grossman, J. P., Ho, C. R., Ierardi, D. J., Kolossváry, I., Klepeis, J. L., Layman, T., McLeavey, C., & Moraes, M. A. (2008). Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, *51*(7), 91–97. https://doi.org/10.1145/1364782.1364802

Thakkar, A. J. (2014). *Quantum chemistry : a concise introduction for students of physics, chemistry, biochemistry, and materials science*. Morgan & Claypool Publishers.

Tuckerman, M. E., Berne, B. J., & Martyna, G. J. (1991). Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *The Journal of Chemical Physics*, *94*(10), 6811–6815. https://doi.org/10.1063/1.460259

Ura, T., Okuda, K., & Shimada, M. (2014). Developments in Viral Vector-Based Vaccines. *Vaccines*, *2*(3), 624–641. https://doi.org/10.3390/vaccines2030624

Warshel, A., & Levitt, M. (1976). Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, *103*(2), 227–249. https://doi.org/10.1016/0022-2836(76)90311-9

Wherry, E. J., Jaffee, E. M., Warren, N., D’Souza, G., & Ribas, A. (2021). How did we Get a COVID-19 Vaccine in Less Than 1 Year? *Clinical Cancer Research*, *27*(8). https://doi.org/10.1158/1078-0432.ccr-21-0079

Yin, Y. W., & Steitz, T. A. (2002). Structural Basis for the Transition from Initiation to Elongation Transcription in T7 RNA Polymerase. *Science*, *298*(5597), 1387–1395. https://doi.org/10.1126/science.1077464