Molecular Dynamics Simulations: Methods and Value in the Folding Problem

Devon Chandler-Brown

15 March 2013

Introduction

The protein folding has been an outstanding problem in molecular biology for a long period of time. Stated simply, the question of protein folding is that of how the primary amino acid sequence of a polypeptide determines its final, three-dimensional structure. The ultimate goal of this endeavor is to be able to predict the folding pathway as well as the final structure based upon sequence information as input.[1] At face value, this problem seems rather simple given that the chemical principles that govern protein folding are thought to be well established. Forces driven by ionic, Van der Waals, hydrophobic and hydrogen bonding interactions are thought to be the primary energetic sources that direct protein folding. The constraints on these forces are the rotational limitations around back-bone bonds and the entropic cost of folding a free, unstructured polymer into a relatively organized structure.[1] Given clear understanding of these forces, it is theoretically possible to make good predictions as to the energy landscape that determines folding paths and stable conformations.

This physics-based approach was regarded as the optimal method for structural predictions for some time; however it ran up against several obstacles that led to alternative approaches. Chief among these challenges were sufficient computing power. Until recent advances in computing, running folding models over even short periods of time (on the order of nanoseconds to 1 microsecond, much less than the folding time of most proteins) was extraordinarily difficult due to the magnitude of allowed conformations in initially unstructured proteins.[2] This led to the rise of a different methodology for performing structural predictions. By ignoring the explicit folding pathway and exploiting the rapidly increasing number of structures, homology-based structural prediction was developed. Chief among these systems is ROSETTA, developed by the Baker group. This algorithm proceeds by mapping homology between 3-9 residue fragments the query and known structures in the protein database (PDB). These probable substructures are assembled and coarsely shaped by Monte Carlo methods and then the resultant structure is atomistically modeled to minimize energy.[3] To date, this (and similar methods) is the most consistently accurate algorithm to predict protein structure.

Despite the success of this method, it has several shortcomings – some, due to data limitation but others due to intrinsic properties of the method. The former can continue to be addressed with careful experimental probing of the shape of molecular force fields and the accumulation of more structures. The two key shortcomings associated with the methodology, however, require the consideration of alternative approaches. One of these issues is that it relies on homology to generate the initial structural predictions which necessarily limits the search space to known conformations associated with these primary amino acid sequences. This limits the ability to predict novel folds without experimentation. The second key limitation to the homology-based approach is that it discards much of the information concerning the folding pathway. Knowledge of this pathway is interesting not only for its intrinsic information, but also informs cases where native structure is constrained kinetically in addition to thermodynamically. These shortcomings even appear in more recent CASP competitions, wherein some structures were unable to be accurately predicted (Rhiju Das, presentation).

These shortcomings of homology-based approaches argue that ways of incorporating molecular dynamics (MD) into these systems or developing more comprehensive MD simulations that can be utilized with modern computing systems. Here, we will discuss the basic methodologies associated with MD simulation, including the choice of force fields, methods of solvent simulation and techniques that may improve future MD models. We conclude by briefly discussing novel ways these methods might be utilized to improve upon structure prediction.

Building a Molecular Dynamics Model

Molecular dynamics is a method of building protein folding models based primarily on Newtonian mechanics. The forces, based on the spatial organization of the atoms, are allowed to guide the relative position of the atoms over time in a stepwise fashion according to a series of time and space dependent differential equations.[4] In order to make this type of approach feasible, general assumptions are made about the time scales on which molecular movements occur. MD traditionally ignores quantum effects on electron movement (with respect to the nucleus) and therefore treats atoms as rigid spheres (known as the Born-Oppenheimer approximation). Additionally, this assumes that vibrational motions occur on much more rapid timescales than the displacements required for molecular folding and therefore can be dismissed from primary consideration. These initial assumptions limit the number of terms needed to describe the motion of the atoms.[2]

With these assumptions in place, what other factors need to be considered when generating the model? Several key considerations must be addressed in order to properly model the protein folding system. The first of these is the consideration of force field that is appropriate for the system in question, that is, what are the systems of equations that describe the energetic landscape that governs protein folding. The most widely used force fields for protein folding algorithms are the classic AMBER, CHARMM, OPLS and GROMOS force fields.[4] Furthermore, each of these fields can utilize explicit or implicit solvent descriptions.[4,5]

Force Fields

These force fields were initially developed in the 1980s and originally tested on gas phase systems. Subsequent modifications have adapted them for use in condensed phases, thereby improving their utility for predicting protein structure. All of these systems employ a basic energetic equation that sums over terms that describe the critical classical forces thought to govern protein folding (see equation 1). The first term in the series represents the potential associated with along axis vibrations. Usually the spring constant associated with these (k_d) is very high and therefore keeps the overall contribution of vibrations to atomic positioning very small. The second term incorporates bond angle bending – that is, deformation causing angles between bonds to change. The third term is a constraint on rotation around a bond. The fourth term ('impropers') is an energetic term to conserve the planar nature of certain groups such as aromatics and amides and exacts an energetic cost for bending outside the plane. The last two terms deal with electronic contributions to potential.[4,6] The first non-bonded pairs term is the Lennard-Jones potential that accounts for van der Waal's forces: electrons are not usually distributed evenly over space, leading to an uneven distribution of charge that contributes to a balance between repulsion and weak

$$
\begin{aligned}
V(r) = {} & \sum_{\text{bonds}} \frac{k_d}{2}(d-d_0)^2 + \sum_{\text{angles}} \frac{k_\theta}{2}(\theta-\theta_0)^2 \\
& + \sum_{\text{dihedrals}} \frac{k_\phi}{2}(1+\cos(n\phi-\phi_0)) + \sum_{\text{impropers}} \frac{k_\psi}{2}(\psi-\psi_0)^2 \\
& + \sum_{\substack{\text{non-bonded}\\ \text{pairs}(i,j)}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{\substack{\text{non-bonded}\\ \text{pairs}(i,j)}} \frac{q_i \times q_j}{4\varepsilon_D r_{ij}}.
\end{aligned}
\tag{1}
$$

attractions. The last is the Colombic potential between atoms with discrete charge. These potentials are then parameterized with respect to the molecule of interest and the result is used to run the simulation.[4,6]

Each of the different force fields uses this general equation incorporate empirically-derived data such that the potential can then be applied to simulation. This data is derived from quantum mechanics experiments performed on small molecule compounds such as alanine di- and tetra-peptides.[7] As more experimental data is accumulated, these models are continually updated. One of the experimentally derived parameters of key importance is the potentials associated with Ramachandran angles φ and ψ as these strongly influence final protein structure.[6] In addition, critical terms have elaborated the original potential model. Among these are attempts to include polarizability, a term that is thought to make a significant contribution proper protein folding.[1,4,6]

Of these force fields only AMBER, CHARMM and OPLS are directly comparable, as they are all atom models – meaning they incorporate all atoms, including hydrogen atoms explicitly.[6] In contrast, GROMOS treats hydrogen atoms bound to carbons (i.e. methyl groups) as rigid bodies as part of what is called a unified atom force field.[8] The trade-off between these two types of models is computation time. By treating some of the hydrogen atoms as part of a larger group, this means that fewer bodies are being examined and consequently, fewer bodies need to be modeled.[6] This increase in speed for computation also corresponds to a coarser-grained structural model. Depending on the goal of the simulation – overall protein shape or specific atomic localization – will determine which algorithm is best suited. There has been some comparison between the three all-atom force fields and, in general, they perform approximately comparably although one study suggests that OPLS may perform slightly better.[9]

Solvation Models

In addition to the force field contributions associated with the protein, additional consideration must be given to the environment in which the protein folds, namely models to consider the contribution of water to the process. These models come in two types: implicit and explicit.[4,6] Explicit models of water treat all atoms individually and track each interaction with the solute as well as each other. This is by far the most accurate model, however it comes at a high cost in terms of computational demand.[10] Implicit water models disregard some of the individual atomic and molecular nature of water and treat it as more of a polarized continuous structure which requires significantly less computational power, but again at a loss of resolution.[11]

One of the most popular implicit methods of calculating solvation is known as the generalized Born (GB) method. This method works by assuming that each molecule of water acts as a sphere of radius *r*, which represents the interaction distance for repulsive forces. The interaction is also governed by Colombic forces generated by its molecular dipole. This dipole is eventually reduced to a field according to the sum across the individual molecules, resulting in a macromolecular, net dipole.[11] This approach has been compared with the explicit model SPC in the context of peptide folding and the results indicate that these approximations cause considerable deviations in the overall free energy landscape of the protein. This appeared to be irrespective of whether AMBER or OPSL force fields were used to describe the protein. This had the consequence of leading to poorly modeled final structures compared to experimental data whereas the explicit model fared much better. Of the implicit models, the combination with AMBER96 was the best, however it still deviated much more than the explicit model.[5]

Although implicit models can be used, they have not yet arrived at a point where they provide the same structure guiding power as explicit models.

Explicit models used in biomolecular simulation are thought to yield final structures much more consistent with experimental observations. Of these, however, which is best? Some commonly used solvation models for protein folding include TIP3P, TIP4P, SPC and SPC/E. As models of pure water, the TIP3P is the worst in terms of correlation with predictions of water properties while TIP4P, SPC and SPC/E fare better. That being said, there are somewhat different recommendations based on relative comparisons in the context of protein folding where TIP4P and SPC are thought to be optimal.[12] This is believed to be dependent on the differences in Lennard-Jones potentials and the spatial organization of the molecules. Another factor to bear in mind is a recommendation from Mackerell in 2004 that because each of the protein force fields is developed in conjunction with a corresponding explicit solvation model, that it may be most appropriate to use these in conjunction as the corresponding systems may be optimal.[6]

Running Simulations

Once the appropriate models for molecular dynamics are in place, it is necessary to generate the final structure. This is a computationally challenging process for the primary reason that in atomistic models (AMBER, OPSL and CHARMM), the sum of all interactions in the potential must be computed for every atom in the protein. One can imagine that this scales very rapidly with length of the protein.

In addition, Newtonian physics applies in this case under the assumption of constant energy (E) which is inconsistent with the fact that most experiments are performed under isothermal and isobaric conditions. In order to make results comparable to experimental observations, adjustments must be made to either allow an inflow of energy or permit volume to change to keep consistency with observation conditions.[6] Included in these algorithms are methods such as the Nose-Hoover which is based on a local energy reservoir and Langevin thermostat that dissipates energy my molecular friction.[4,13,14] The former is an example of a global correction, acting on all molecules simultaneously whereas the latter is a local thermostat, acting at the smallest level.

Additionally, there are some aspects of protein folding that may not be sampled due to the coarseness of the simulation. One such aspect is bond length vibrations which occur on time intervals less than nanoseconds which can therefore be treated essentially as a constrained or rigid system.[4,15] In order to accommodate this behavior, algorithms such as SHAKE or RATTLE are utilized, allowing local rigidity to be accommodated in the context of other molecular motion (such as angle bending and rotation which still need to be permitted for proper folding).[16]

All of these additional layers of computation mean that novel ways of performing computing are required to successfully perform these models – especially in atomistic simulations. Advances in both algorithms as well as parallel computing have made it possible to run these simulations in much shorter time periods.[9] One algorithmic method is the molecular dynamics utilization of replica exchange, an algorithm originally utilized for Monte Carlo protein simulations. By running two folding regimes at different temperatures and periodically exchanging pieces of the model across temperatures, it allows a coarser sampling of the overall energetic landscape by escaping local minima. This allows for faster total computation.[17] Another novel computing technique that has been employed to allow for greater sampling in less time is to use a series of parallel processors to probe short-term local folding patterns rather than attempting to model the entire long-folding pathway at once. This process has been used to model the villin fold by assembling the local MD simulations a constructing what is known as a Markovian state model (MSM). This is not a purely molecular dynamics based approach but still allows the entire folding pathway to be reconstructed as well as the final structure from similar physics-based landscape probing.[18] Novel approaches such as MSMs as well as exploitation of new computing resources such as cloud computing will allow molecular dynamics models to continue to improve the capacity for protein prediction.

Conclusions

An enormous amount of information must be included in order to use non-homology-based methods to arrive at protein predictions however modern computing methods and creative algorithms may pave the way toward future discoveries based on molecular dynamics simulations. One recent paper identified specific conformational changes associated with histone H4 acetylation using a molecular dynamics approach. In this instance, an intrinsically disordered domain (the histone tail) was included as part of the model which revealed alterations in propensity towards certain conformations upon addition of the acetyl group.[19] This is an instance where homology-based methods could not be predictive as the process at hand is fundamentally dynamic and cannot be measured using traditional crystallography (although it can be seen by NMR). Although these types of discoveries are exciting, molecular dynamics will face significant difficulties. For instance, there are some force field biases that may still need to be addressed as some may have slight biases towards certain structures such as alpha helices.[20] This, however, has been a diminishing problem with each new iteration of model. Additionally, challenges remain in the process of scaling up. Most proteins used in molecular dynamics models are relatively small in size and a combination of MD and homology-based approaches may have to be utilized to begin the assembly and structure of larger proteins and protein complexes.[1,2] Despite these challenges, continued improvements in algorithms and computer processing may make these goals possible.

References

(1) Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on.*science, 338*(6110), 1042-1046.

(2) Best, R. B. (2012). Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology*.

(3) Das, R., & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu. Rev. Biochem., 77,* 363-382.

(4) Hug, S. (2013). Classical Molecular Dynamics in a Nutshell. In *Biomolecular Simulations* (pp. 127-152). Humana Press.

(5) Zhou, R. (2003). Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins: Structure, Function, and Bioinformatics, 53*(2), 148-161.

(6) Mackerell, A. D. (2004). Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry*, *25*(13), 1584-1604.

(7) van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X.,... & Yu, H. B. (2006). Biomolecular modeling: goals, problems, perspectives. *Angewandte Chemie International Edition*, *45*(25), 4064-4092.

(8) Daura, X., Mark, A. E., & Van Gunsteren, W. F. (1998). Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *Journal of Computational Chemistry*, *19*(5), 535-547.

(9) Shirts, M. R., Pitera, J. W., Swope, W. C., & Pande, V. S. (2003). Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *The Journal of chemical physics*, *119,* 5740.

(10) Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., ... & Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, *30*(10), 1545-1614.

(11) Bashford, D., & Case, D. A. (2000). Generalized Born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, *51*(1), 129-152.

(12) van der Spoel, D., van Maaren, P. J., & Berendsen, H. J. (1998). A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *The Journal of chemical physics*,*108*(24), 10220-10230.

(13) Hünenberger, P. H. (2005). Thermostat algorithms for molecular dynamics simulations. *Advanced Computer Simulation,* 130-130.

(14) Bussi, G., & Parrinello, M. (2008). Stochastic thermostats: comparison of local and global schemes. *Computer Physics Communications*, *179*(1), 26-29.

(15) Dantus, M., Bowman, R. M., & Zewail, A. H. (1990). Femtosecond laser observations of molecular vibration and rotation.

(16) Forester, T. R., & Smith, W. (1998). SHAKE, rattle, and roll: efficient constraint algorithms for linked rigid bodies. *Journal of computational chemistry*, *19*(1), 102-111.

(17) Rosta, E., & Hummer, G. (2009). Error and efficiency of replica exchange molecular dynamics simulations. *The Journal of chemical physics*, *131*(16).

(18) Jayachandran, G., Vishal, V., & Pande, V. S. (2006). Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *The Journal of chemical physics*, *124*, 164902.

(19) Potoyan, D. A., & Papoian, G. A. (2012). Regulation of the H4 tail binding and folding landscapes via Lys-16 acetylation. *Proceedings of the National Academy of Sciences*.

(20) Best, R. B., Buchete, N. V., & Hummer, G. (2008). Are current molecular dynamics force fields too helical? *Biophysical journal*, *95*(1), L07-L09.