Methods in Pharmacology and Toxicology (2016): 1–30
DOI 10.1007/7653_2015_41

Published online: 20 March 2015

Molecular Dynamics Simulations and Computer-Aided Drug Discovery

Ryan C. Godwin*, Ryan Melvin*, and Freddie R. Salsbury Jr.

Abstract

Molecular dynamics simulations of biomolecules, proteins especially, have emerged as an important tool in the study of the conformational change, flexibility, and dynamics. These simulations, especially when combined with virtual screening, have been a tool in drug discovery. Herein, we cover the basics of molecular dynamics simulation, in the hopes that a reader would be able to intelligently conduct a simulation of their favorite protein(s), analyze the results in order to make hypotheses about the links between protein dynamics and conformation. We also discuss the integration between molecular dynamics and virtual screening, so that a reader could use the results of simulations to perform virtual screening for lead identification. Finally, we review several case studies to show what sort of information can be gained by simulation of biomedically interesting proteins, and how that may impact drug discovery, as well as a discussion of some areas in which simulation may prove more useful in the near future.

Key words Molecular dynamics, Simulations, Drug discovery, Markov analysis, Protein dynamics, Acmed

1 Introduction

Molecular dynamics simulations of biomolecules have been developed since the late 1970s and early 1980s (1) in order to harness the emerging power of computers to study the motions of proteins and other biopolymers, as well as to study the interactions of these biomolecules with small molecules, such as potential drugs. These computational techniques often complement experimental techniques such as Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography. Observing dynamics or obtaining ensembles of conformations using these methods can be difficult. However, these experimental techniques often provide highly accurate structural information that computational methods can use as starting points to study biologically important molecules such as small molecule ligands, DNA, RNA and proteins. In particular, Molecular Dynamics (MD) simulations provide a

*Author contributed equally with all other contributors.

method to examine, in atomic detail if necessary, the kinetics and thermodynamics of important biomedical systems.

Since all-atom molecular dynamics simulations require the integration of Newton's equations of motion of each atom, usually including solvent and solvent ions, over short time-steps, typically on the femto-second timescale, these simulations can be rather computationally demanding. However, the growth of computer power especially in the late 1990s and early 2000s enabled these methods to be particularly predictive in studying protein dynamics, such as in investigating the impact of protein motions on catalysis and ligand binding (2–4). The latter studies have been especially influential as they have required considerable discussion of the interplay of conformational change, such as changes in active site geometries in DHFR (2) or metallo-beta-lactamases (3) and coupled protein fluctuations (4), which show that within a single protein conformation, long-range coupling networks exist and are sensitive to interactions with different ligands.

Even more recently, molecular dynamics simulations have proven useful in studying larger biological systems and in aiding in the drug discovery process by providing a predictive complement to experimental methods, contributing predictions for dynamics and structures not easily observed in vitro or in vivo. Such predictions are useful in pharmacology for understanding the interactions of drug candidates with biological systems on an atomic scale.

Molecular dynamics simulations also prove useful when considering proteins as ensembles of conformational states (5–10), as simulations explore ensembles and output large collections of structures, which sample the conformations that occur.

In part, the notion of generalized allostery comes out of the conceptualization of proteins as ensembles of states and the understanding of conformational changes occurring due to long-range coupling networks (9). If under certain conditions all proteins are indeed allosteric, it is possible to design drugs that will bind to allosteric sites. Such binding would force the protein into a certain conformation—or specific ensemble of conformations—thereby regulating the dynamics and interactions of the protein (9, 11). However, for any given protein, an appropriate ligand and corresponding binding site to induce the desired structural change must be found. This type of search is one for which molecular dynamics simulations are well suited. Much of the relevant scientific work in the 2000s was reviewed a few years ago (12). This chapter serves a few purposes. First, it expands upon and update that previous review, especially in light of the tremendous improvements in computational algorithms and hardware, such as GPU-enabled computing. Second, we describe the minimum theoretical and technical details necessary for setting up, executing, and analyzing MD simulations so that any who are interested in participating in computer-aided drug discovery may have the tools necessary for doing so.

2 Basics of Molecular Dynamics

2.1 Structures

The minimum structural information required to start a simulation is:

1. A list of all atoms involved in the simulation
2. Initial coordinates of these atoms

For a given system, with fixed protonation states, there is only one possible list of atoms; however, there are infinitely many possible initial coordinates. Of course, most of these combinations would have enormous energy and would be negligible members of the real, physical ensemble. To achieve realistic results, a physiological initial state needs to be considered. Folding a biopolymer from an unfolded state can rarely be achieved straightforwardly—the time scales are still too long—except for the smallest systems. Therefore, simulations usually start in a folded state; the set of coordinates that likely correspond to a minimum free energy state. Online databases–e.g., the protein databank (RCSB PDB—with 106,293 structures to date), which collect structures from X-ray crystallography and NMR spectroscopy (13)—are the normal sources for such initial structures. It is also possible sometimes to model the initial atomic coordinates based on the structure of other proteins with similar sequences via homology modeling (14). The extent to which this is accurate of course depends on how close the unknown protein is to the known proteins, which is generally not known. As such, simulations almost always start from structures obtained from the RCSB protein databank. A promising development that could have impact in the near future is the possibility of building up structures from a type of quantum mechanical method known as Density Functional Theory (DFT). Variants of this method have proven useful in materials science and computational chemistry (15).

2.2 Force Fields

Force fields are the potential energy functions used to calculate the accelerations of the atoms and subsequently update the coordinates and the velocities at each step of the simulation. This parameterization of the energy surface of a protein or other biopolymers is conceptually straightforward, but complicated in practice. In principle, the energy surface of even a small protein has 100,000s of dimensions even without solvent. However, since the aim is to simulate the dynamics of connected and folded proteins, this surface can be simplified using conventional terms from chemistry, such as bonds, angles, dihedrals, and other terms related to chemical connectivity, and long-range interactions as modeled by van der Waals interactions and electrostatics. Among the many force fields that exist, the most popular families of force fields include CHARMM (16), AMBER (17), and GROMOS (18). The energy

equation from the CHARMM 27 force field is shown in Eq. (1), where V is the total potential energy.

$$\begin{aligned} V \quad &= \sum_{\text{bonds}} k_{\text{B}}(b - b_0)^2 + \sum_{\text{angles}} k_{\theta}(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_{\phi}[1 - \cos(n\phi - \delta)] \\ &+ \sum_{\text{impropers}} k_{\omega}(\omega - \omega_0)^2 + \sum_{\text{UB}} k_{\text{u}}(u - u_0)^2 \\ &+ \sum_{i>j} \varepsilon_{ij} \left[\left(\frac{R_{\min_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{\min_{ij}}}{r_{ij}} \right)^{6} \right] + \sum_{i>j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \end{aligned} \tag{1}$$

Many of the bonded interactions are effectively modeled as simple harmonic oscillator potentials, including bonds, angles, the Urey-Bradley term, and impropers, i.e., the first, second, fourth, and fifth terms in Eq. (1). In each of these terms there are force constants that control the stiffness of the bonds, angles, impropers, and Urey-Bradley terms. In principle, every single such interaction can have its own minimum and force constant, but in practice there is a great of similarity. Bonds, the first terms, are 1–2 interactions that occur between all atoms that are directly connected via chemical bonds. Angles, the second term, are 1–3 interactions that occur between all atoms that share a common bonded atom. The impropers, the fourth term, are 1–4 interactions that occur between atoms that share common angles. They occur between some atoms, those in which dihedrals are insufficient to constrain the torsional angle. The Urey-Bradley term, the fifth term, is a 1–3 interaction energy, i.e., an interaction between atom pairs that share a common bonded angle, that some atom pairs have and is designed to control angle-bending for particularly stiff angles. Dihedrals, the third term, are 1–4 interactions between all atoms that share common angles and are modeled with a cosine approximation. The last two terms are the non-bonded interactions, and are modeled via the Lennard–Jones potential and the Coulomb potential, where every atom pair that does not occur in a bond, angle, or dihedral, possesses these long-range interactions. The nature of the $1/r$ Coulomb potential is a long-range interaction, and is computationally limiting, since it does not go rapidly to zero as the Lennard–Jones potential does over longer ranges. However, methods have been developed to approximate the Coulomb potential accurately over longer ranges, such as the particle mesh Ewald method (19).

Although force fields are complicated approximations, these models are constantly being vetted and compared to experiment to improve the force field parameterization for proteins, nucleic acids and lipids. The force fields have been refined over the years to correct issues where, for example, AMBER over-stabilized alpha-helices (20, 21) or CHARMM tended toward pi-helices (22). There is little consensus to suggest that one force field is better

than the rest for protein simulations, and simulations performed on the same structure with different force fields generate consistent results, for example ref. (3) vs refs. (4) and (23). The success of these force fields has been recently highlighted when Martin Karplus, Michael Levitt, and Arieh Warshel won the 2013 Nobel Prize in Chemistry "*for the development of multiscale models for complex chemical systems*" (24).

2.3 Simulation Programs

Various simulation suites exist and the most popular include NAMD (25), CHARMM (26), AMBER (27), and GROMACS (28). These suites share common basic features but vary in their capacities and underlying philosophies.

The most user-friendly of these suites is NAMD, built upon C++ and TCL programming and scripting languages, but has the least functionality. However, it contains all the functionality needed for all-atom simulations. Conversely, the most versatile package is CHARMM, but it comes with a steep learning curve, and resembles a Fortran-based language. GROMACS is the only one of the four suites that is open source, and has been converted from its original FORTRAN implementation to C. Of these four packages, NAMD is the most capable of performing large, classical all-atom simulations on CPUs, and has been used to simulate particularly large proteins and protein complexes (for example (4, 29, 30)). GROMACS has the advantage of a large number of external tools for trajectory analysis; it is generally the second-fastest. CHARMM is the most flexible for analysis and for performing different simulations. For the simulations described in this chapter while running on CPUS, arguably the "best" combination would be to use NAMD to run simulations and CHARMM for analysis, while using GROMACS for both simulation and analysis would be a close second. However, over the last few years, GPU-enabled codes have emerged, especially ACEMD (31), which is similar to NAMD in its functionality. Until and unless other suites emerge that are as GPU-enabled, the ideal simulation technology at present is ACEMD on GPUs.

2.4 Running a Simulation

Given a particular biomolecular system of interest and a simulation package, the next step is to set up the simulation parameters. Many of these are default configuration parameters that should be understood.

First, a choice needs to be made as to which thermodynamic ensemble should be approximated. Since the isothermal–isobaric (NPT) ensemble has the Gibb's free energy as its thermodynamic potential, and usually corresponds to experimental conditions, this is currently the most common ensemble to simulate. Simulating the NPT ensemble requires using a thermostat and barostat to approximate constant temperature and constant pressure respectively.

Simulation packages typically offer the option to run other ensembles, minimally the canonical (NVT), and microcanonical (NVE) ensembles. Although it seems logical that one would simulate in the NPT ensemble for best agreement with experiments, it is not clear how different simulations are in these various ensembles.

To best represent physiological conditions, water molecules and ions that surround biomolecules in vivo are either explicitly or implicitly modeled in simulation; this is an important enough topic to warrant its only section below. In the most common case of explicit solvent and ions, periodic boundary conditions are implemented and then long-range electrostatic interactions are approximated using a particle mesh Ewald summation method with Fast Fourier transforms (32).

Embedded in each simulation code are numerical integration methods that are used to update the positions and velocities of each atom in the system from the accelerations determined by the force field for each simulated atom at each time-step. This time-step, or interval over which the forces are considered constant, and which determines how often configurations change, is an important consideration. If the time-steps are too small, computer time and disk space will be wasted. If the time-steps are too large, the simulation is no longer energy conserving and accuracy will suffer. However, simulation packages typically have good default choices of integrators, such as velocity Verlet (33) with time-steps of 1–2 fs.

After the simulation has been set up, usually a brief minimization is performed to remove any clashes between atoms. Post-minimization the system is simulated for a given number of time-steps—depending on the timescale necessary to address the biomedical problem as well as the computational power available. With current GPU-enabled codes, simulations on the 100s of nanosecond to microseconds are feasible depending on system size and patience of the user. Also, typically multiple simulations are performed where each atom starts with a different random velocity, taken from a Boltzmann distribution, to allow for better coverage of phase-space.

3 Solvation Techniques

In order to accurately simulate biomolecules, it is imperative to recreate the local environment as best as possible. As such, biomolecules are simulated in an aqueous environment to approximate physiological conditions. Modeling solvation is important, as it has been shown that solvent fluctuations can be directly related to protein motions (34, 35). Additionally, the layer of water surrounding the biopolymer, i.e., the water molecules closest to the sample, has properties different than that of the bulk solvent (36). It is clear that solvent interactions are critical to properly functioning biomolecules,

and when simulating such systems, the choice of solvent approximation is an important issue. There are two main approaches to simulating solvents, with explicit or implicit solvent, and many models within each implementation. While none of these is perfect, some are advantageous particularly dependent on the simulation in question.

An explicit solvent is exactly that, including a box composed of an oxygen bound to two hydrogens, each with updated coordinates and velocities calculated at each time-step. These models are often characterized by the number of site interaction points considered and go from 2-site models up to 6-site models. TIP3P and TIP4P are common 3 and 4-site models, respectively (37, 38), that have been studied extensively (39, 40).

The simplest explicit water models assume rigid bonds and only calculate non-bonded interactions including van der Waals and electrostatic interactions. In many explicit models, water bonds are maintained via the SHAKE algorithm, in order to speed up calculations as these bonds are typically not interesting, yet are of high frequency (41).

Because explicit solvents can handle representative motions at a global and local scale, they are often preferred. Alternatives include implicit solvent approximations in which the electrostatic properties of the water are calculated approximately without including the explicit presence and motion of water molecules. This reduces the computational expense by removing explicit water atoms. Various implicit solvent models have been compared (42–44), and Generalized Born (GB) models (23) show the most promise of the implicit models (42, 43). The struggle is to reproduce the solvent behavior consistent with experiment, and while there has been some success (23, 45), comparison of the TIP3P water model to GB implicit models show an over-stabilization of secondary structure in implicit solvent models over explicit solvent models. Namely, it has been shown that alpha-helices are over stabilized in GB models over TIP3P (20), and ion pair interactions are sometimes over stabilized leading to the trapping of molecules in non-native states (39). Overall, implicit solvents reduce computational time, yet they pay a penalty in accuracy. However, explicit solvents, while generally more accurate, require additional computer resources. In the era of GPUs and parallelization of calculations, explicit models are preferable when possible, due to their accuracy.

Regardless of the solvation technique, in order to model in vivo or in vitro conditions, ions need to be added to the simulation. If the ions exist in the X-ray structure, then they can be added in explicit in the positions in the structure, as such ions are likely to be structurally important. Otherwise, there are automated processes for doing so in software packages such as VMD (46), which place sufficient ions randomly in the water box to match conditions desired; such as 0.15 M ionic strength with NaCl as is common to match experiment conditions, or just sufficient Na^+ or Cl^- to neutralize the protein system.

4 Analysis Methods

Once simulations have been performed, they must be analyzed to check their validity and also to extract useful information. Since the results of a simulation are the coordinates of all the atoms in the system simulated over the timescale of the simulation, a wide variety of analysis methods can be applied to extract virtual any type of structural and provide the most dynamical information. Below, the most common–and typically the most useful—of these analysis methods are discussed.

Simulation Check: *Structural relaxation*
Given that simulations can run from hours to months depending on the system size and timescale desired, performing energy and structural checks shortly after starting a simulation minimizes time wasted on unstable or improperly setup simulations. If a simulation's log file has been configured to report energies, the user can read out the energies after a relatively small number of time-steps to see if the energies reported are reasonable. Similar checks can be performed on the pressure and temperature, which should be relatively constant for biological systems and fluctuate around the values set for the thermostat and barostat (usually 300 K and 1 atm, although 310 K can be used to better match physiological conditions). As an additional validity check, a user can calculate the Root Mean Square Deviation (RMSD) of a subset of the simulation's atoms. This measure quantifies how much the polymer of interest has changed from a reference structure over time. Such checks allow a user to judge the physical reasonableness—relative to the system and thermodynamic ensemble chosen—of a simulation (12). The reference structure is usually the initial structure that has been obtained from experimental work, in order to gauge the stability of the simulation and structure itself. Beyond understanding how realistic a simulation is, measures of energy, pressure, temperature and RMSD versus time are indicators of a successfully equilibrated system. An initial period of rapid change followed by relative stability with small fluctuations around a mean value indicates a successfully equilibrated system. For example, in Fig. 1, there is a rapid growth of RMSD in the first 100 ns. Afterward, the system reaches an equilibrated state with small fluctuations around a mean of about 5 Å.

This RMSD measures the average difference of all selected atoms from one frame to the next via

$$R_{\mathrm{RMSD}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\vec{r}_i - \vec{r}_i'\right)^2}$$

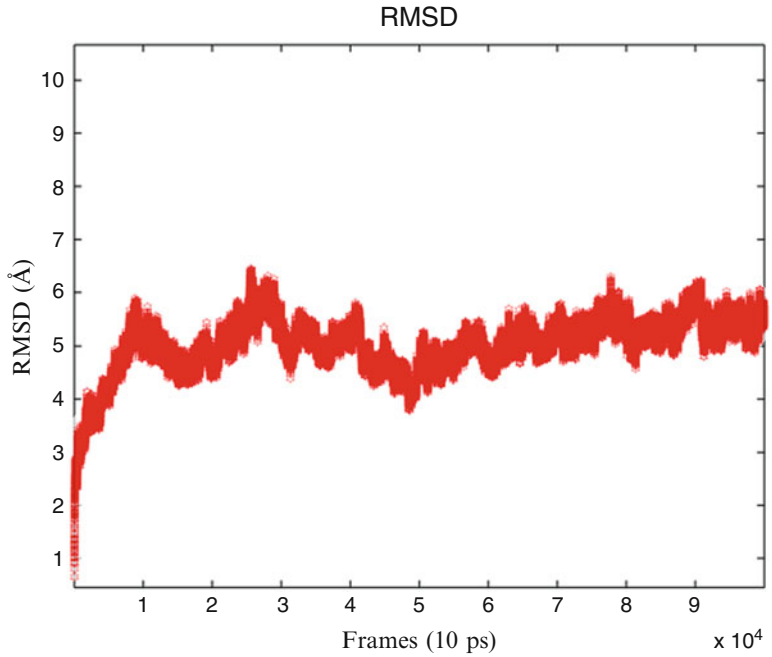

Fig. 1 All-atom MD simulation of a Zinc-Finger structure has a 100 ns equilibration phase

where N is the number of atoms in the selection, r is the position vector (x,y,z) of the atom at time t, and r' is the position of the atom in the reference frame. Before a meaningful measure of RMSD (or any other quantity that depends on translational or rotational differences in position) can be made, the atoms of interest in the trajectory must first be aligned to some reference structure so as to remove the overall motion of the protein. Without this alignment, conformational changes will be conflated with rigid body motions of the protein that is the diffusion and overall rotation of the entire protein that occurs during the simulation. To align the atoms, analysis software, such as VMD or scripts in CHARMM or GROMACS, minimizes the RMSD of selected atoms between a reference structure and every frame in the trajectory using only rigid-body rotations and translations. This alignment focuses the analysis of protein conformations and dynamics.

4.1 Clustering: Searching Conformation Space

Given that simulations can run from hours to months depending on the system size and timescale desired, clustering analysis simplifies the comparison of structures output from an MD simulation by classifying thousands to ten thousands of frames into a smaller taxonomy with representative conformations. Figure 2 shows how finely these clusters can distinguish structures from simulations, while still reducing the complexity from, in the case, the structural information contained in a microsecond scale simulation to just 50 representative structures along with how often these structures are sampled. Two of which are depicted in Fig. 2 for illustrative purposes.

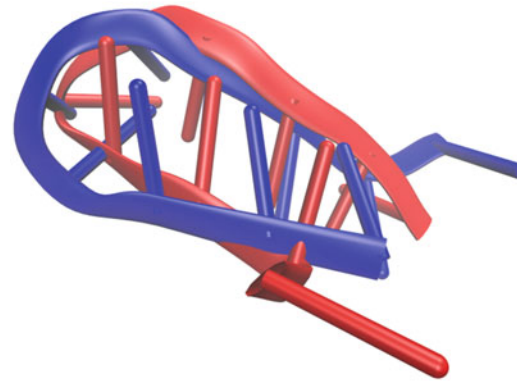

Fig. 2 Two representative structures of a 10-residue FdUMP chain show small conformational differences between some clusters. From the *red* to the *blue* representative, F10's termini have spread apart

The clusters are defined by their size in a parameter space—typically pair-wise RMSDs between structures—so that clustering effectively partitions configuration space. Algorithms for deciding what conformations fall in a given cluster come in two categories, hierarchical clustering and nonhierarchical clustering. The method for determining the distance between clusters distinguishes algorithms within each category. Hierarchical clustering methods partition the conformation space into a tree by iteratively connecting neighboring elements in a dataset. Selecting a level at which to divide the tree forms clusters. Hierarchical clustering, Fig. 3, is simple and fast since once an element of the dataset has been placed in a cluster it is ignored for the remainder of the clustering process. Slower, nonhierarchical methods optimize each cluster based on some desired parameters set by the user. Nonhierarchical clustering, Fig. 4, allows for moving data among clusters as part of the optimization process, making them slower than their hierarchical counterparts (47, 48). Nonhierarchical, iterative methods are implemented in two popular analysis software packages, VMD (46) and CHARMM (49).

VMD uses a Quality Threshold (QT) clustering algorithm (47). The method begins by assembling a cluster based on every element in the dataset. For example, if an MD trajectory contains 10,000 frames, the first iteration of the QT algorithm will have 10,000 clusters. In this iteration the *N*th cluster begins with the *N*th frame and is compared with all other frames, regardless of whether those frames have already been placed in another cluster. The frame that causes the smallest increase in cluster diameter is accepted into the group. This process is repeated for the *N*th cluster until no frame can be added without taking the cluster diameter past the threshold specified by the user. At the end of this iteration,

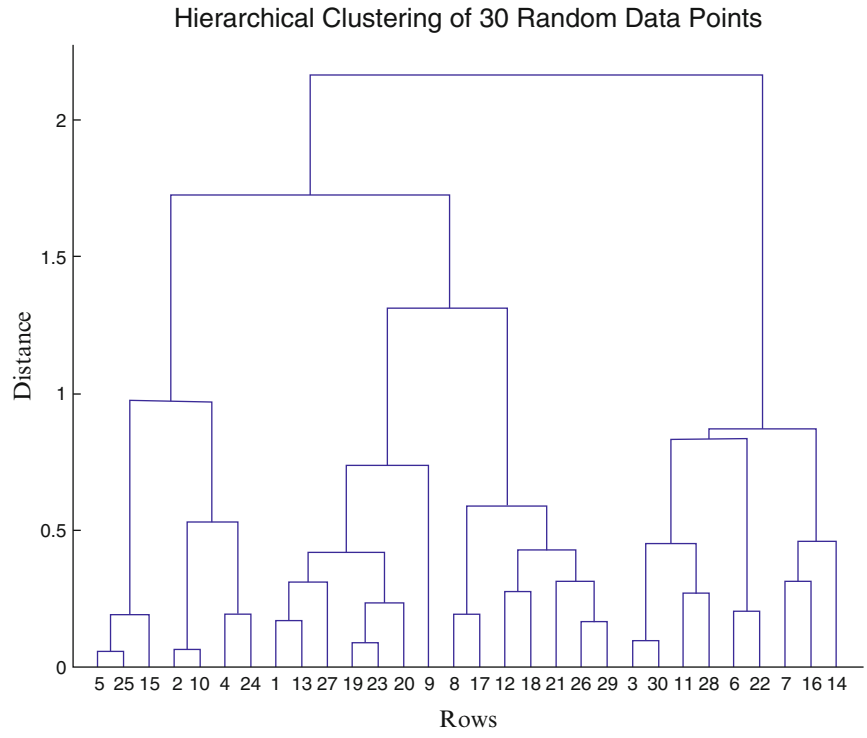

Fig. 3 Once all data points are grouped in the clustering tree, a cutoff distance is chosen. Any lines that join at a distance equal to or greater than the cutoff distance are clustered; all other lines are unclustered. Row values are the representative cluster numbers

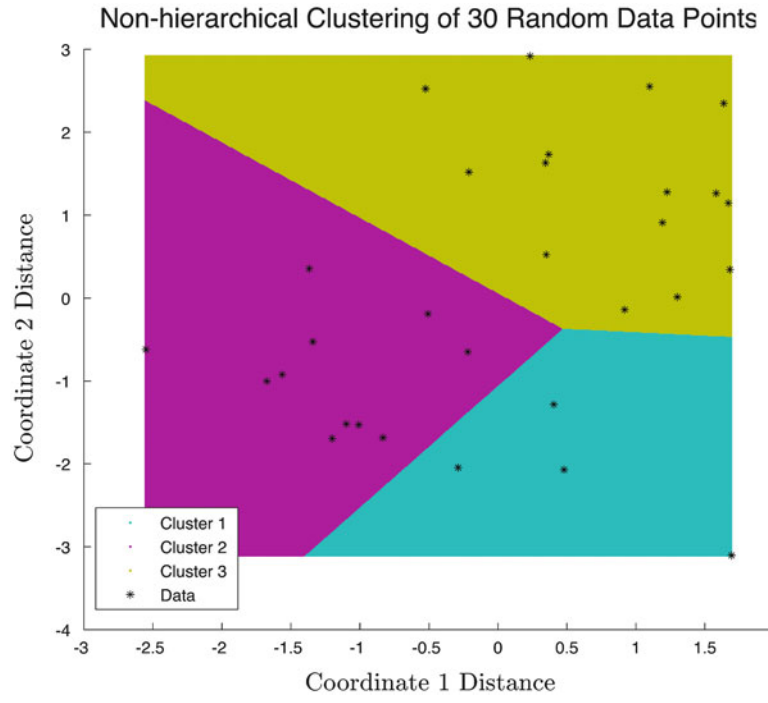

Fig. 4 Non-hierarchical clustering interactively searches for partitions within conformation space that minimize the distance between clusters in each partition

all frames in the largest cluster are removed from consideration. This largest cluster is now fixed, and the process iterates until either no frames remain or the number of fixed clusters equals a target number of clusters set by the user. In the latter case the remaining frames are simply left unclustered. In VMD specifically, if M is the target number of clusters, all unclustered frames are labeled as cluster $M + 1$. CHARMM uses the ART-2′ clustering Algorithm, which is based on a self-organizing neural network (50, 51). Similar to QT, this algorithm optimizes each cluster based on a constrained cluster radius. However, ART-2′ starts with one cluster rather than the largest possible number of clusters. The first of two phases in the clustering determines the number of clusters and their respective centers. To begin, ART-2′ selects the first frame of the trajectory as the center of a single cluster. Then, the Euclidean distance to each (in the space of the user-selected cluster parameter) frame is calculated. If the distance from the cluster center to a conformation is within the cutoff radius, that frame is added to the cluster and the cluster center is recalculated before comparison with the next frame. If the distance is outside the cutoff radius, the rejected frame is assigned as the center of a new cluster. The second phase of clustering is still done with Euclidean distance but is performed in multiple iterations; furthermore, at each iteration, the number of clusters and cluster centers are fixed. Once all frames are assigned to a cluster in a given iteration, the cluster centers are recalculated. Finally, the clustering assignment process is repeated. This cycle of the second phase continues until no changes occur between iterations. An obvious pitfall of this method is that the order of the frames influences the cluster assignment. Therefore, the user may wish to check the stability of the clusters by doing a second round of clustering with a randomized frame order (48).

Regardless of the clustering method used, the user must set input parameters based on some analysis criteria based on user preference. For example, when using VMD's RMSD clustering method, a cutoff distance and number of clusters must be set. These parameters should be chosen in such a way that balances the number of frames placed in clusters and the number of clusters themselves. Obviously, if the number of clusters is set to the number of frames, the user is guaranteed that all frames will be clustered. However, no information is gained in this example as the point of clustering is simplifying the analysis of the trajectory. To this end, the user may first decide a reasonable number of clusters, e.g., 50, to analyze and then adjust the cutoff parameter to minimize the number of unclustered frames. The strategy in that case is to begin with a low cutoff, e.g., 2 Å, and gradually increase it until the point when either VMD reports fewer clusters than the number desired or an increase results in no or a very small change in the number of unclustered frames. Such a procedure balances

approximations, but not requiring a large number of clusters to analyze while including as much of the simulation data available for in clusters for further analysis.

4.2 Markov Analysis

In addition to identifying representative structures of clusters, such as those in Fig. 2, plots of cluster vs frame, Fig. 5, can show the transitions among states. In this representation it is easy to identify long-lived states and to find the populations of different conformational states. However, it is also easy to miss short-lived stable states, and it is difficult to accurately see the transition states by eye. Accessing these transitions requires reconstructing the kinetics of the system, a task for which Markov State Models (MSMs) are well suited.

MSMs are network models that convey the rates of transition among states. These models typically assume the system is memoryless, though they can be generalized to systems with memory. For example, a memoryless process is a Markovian process of order 1. In this case, the state of a system at time-step N depends only on the system's state at the previous time-step $N-1$. To generalize to include memory, one uses a Markovian process of order M. In this case, the state of the system at time-step N depends on the system's state at time-steps $N-M$, $N-(M-1)$, ... and $N-1$. Using

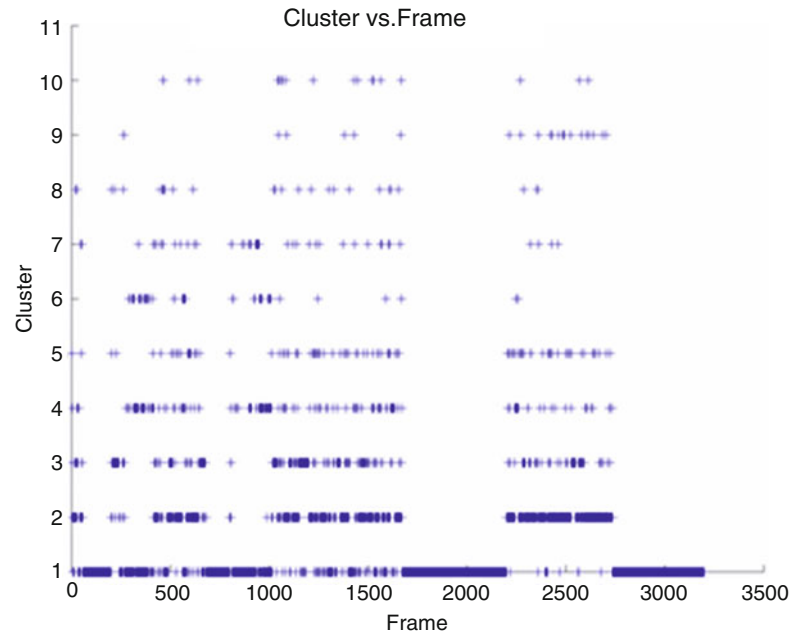

Fig. 5 During a 16-μs all-atom MD trajectory, a 10-mer of FdUMP cluster analysis shows there is a preferred, low energy state with frequent transitions to higher energy states

these models requires defining states based on physical parameters, typically based on RMSD. For example, a state might consist of all structures within 2 Å of each other. In which case, these states are taken from clustering analysis. These Markov models also allow for further simplifications based on kinetic definitions of macrostates, which are combinations of microstates. A macrostate might be all microstates that transition among each other in less than 20 ps. Once these states are, it is possible to apply statistical mechanics to estimate various thermodynamics quantities, such as the free energy of a macrostate using kTlog(P) where P is the population of the states contained in a macrostate.

Recently, software packages for assisting in dynamics-based clustering and construction of MSM have been developed. Two such packages are EMMA (52) and MSMBuilder2 (53). The latter software package has a companion application, MSMExplorer, for visualizing MSMs (54).

A convenient way to convey the information in a MSM is with a rate matrix and heat map thereof, Fig. 6. There are three primary steps in the construction of such a matrix for an MD trajectory. First, order clusters by their corresponding frame number. This

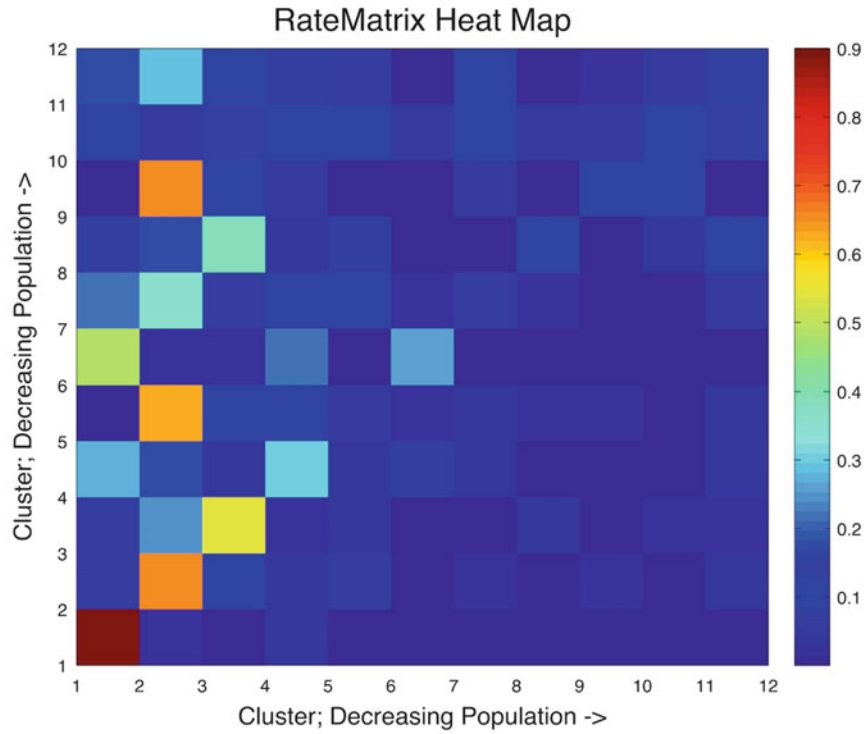

Fig. 6 Whenever this FdUMP polymer enters cluster 1 (*blue* in Fig. 2), it is 90 % likely to remain there during the next time step (in this case 5 ns). This molecule frequently transitions from cluster 5 to cluster 2 and cluster 9 to cluster 2

sorted list is called a Markov chain. Next, count how often state i is followed by state j where both i and j run from 1 to the number of macrostates. Finally, place those counts in A, which is an $M \times M$ matrix, where M is the number of states, and normalize each row so that it sums to unity. Now, the matrix is read "when the system is in state i, the probability it will transition to state j in the next time-step is Aij." Using Markov State Models in this fashion quickly reduces the task of analyzing the kinetics of thousands to billions of conformations to reading an $M \times M$ matrix, where M is much smaller than the number of frames in the trajectory. More details on such modeling are beyond the scope of this review, but a simple yet comprehensive review of them has been published (55).

4.3 Analysis of Protein Motions and Dynamics

Clustering combined with Markov analysis provide information about the configuration space accessed by a biopolymer during the timescale, typically nanosecond to microsecond, of the MD simulation. Markov analysis also provides information about kinetics via quantifying transitions between configurations. However, there is additional information contained in a simulation, including information about protein motions through analysis of dynamics via examination of fluctuations.

Dynamical studies of protein motions are important in understanding regulation and function of a cell. Naturally, cellular function is an extremely active process requiring a myriad of properly functional components. To use this information for predictive purposes, it behooves us to have some knowledge of the motions that dictate proper function and understand how physical laws drive motions. Simulations provide enough insight to hypothesize the important functional processes, as conformational changes have helped to identify drug targets, for example (...). Mechanisms such as protein–protein and protein–surface interactions have recently gained more traction in the drug targeting process (56).

Understanding how protein motions are affected by ligand binding and the impact that may have on proper function of a biomolecule suggest the importance of dynamics in the drug discovery process (57). However, there is a great deal of remaining investigation of the dynamics of protein, DNA, and RNA interactions, and dissecting these dynamics may yet inform the development of therapeutics.

Studies suggest that conformational changes at one site of a protein, for example, affect distant regions of a protein and its ability to bind properly, despite no noticeable changes in the binding region (58). This form of general or hidden allostery is a dynamical component often overlooked in the drug discovery process (9) and is proving useful in predicting functional molecular mechanisms (59).

4.4 Root Mean Square Fluctuations

Root mean square fluctuation (RMSF) is a useful analysis tool to examine the behavior of a protein with atomic precision across the whole trajectory by measuring the average mobility of each atom in the simulation. RMSFs are a measurement of time-averaged fluctuations from a reference frame, typically the first frame. The RMSF is a useful estimation of the rigidity of various parts of the biomolecule, with higher RMSF indicating a more flexible region. Values are typically on the order of a few to tens of Angstroms and are calculated using the following equation,

$$R_{\mathrm{RMSF}} = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T}\left(\vec{r}_i(t_j) - \vec{r}_i'\right)^2}$$

Where r is the x, y, z coordinates of the atom, for example, and r' is that of the reference structure. T is the total number of frames, and i represents the index over atoms and j represents the index over time. Figure 7 provides an example of RMSFs for a small 28-residue zinc finger, NEMO (60). In this example, the flexibility of this protein is studied over different timescales, and surprisingly for such a small protein, the flexibility was radically increased on longer-time scales. The RMSFs plots also show changes in flexibility along the protein backbone, indicating regions of increased flexibility and regions of increased rigidity.

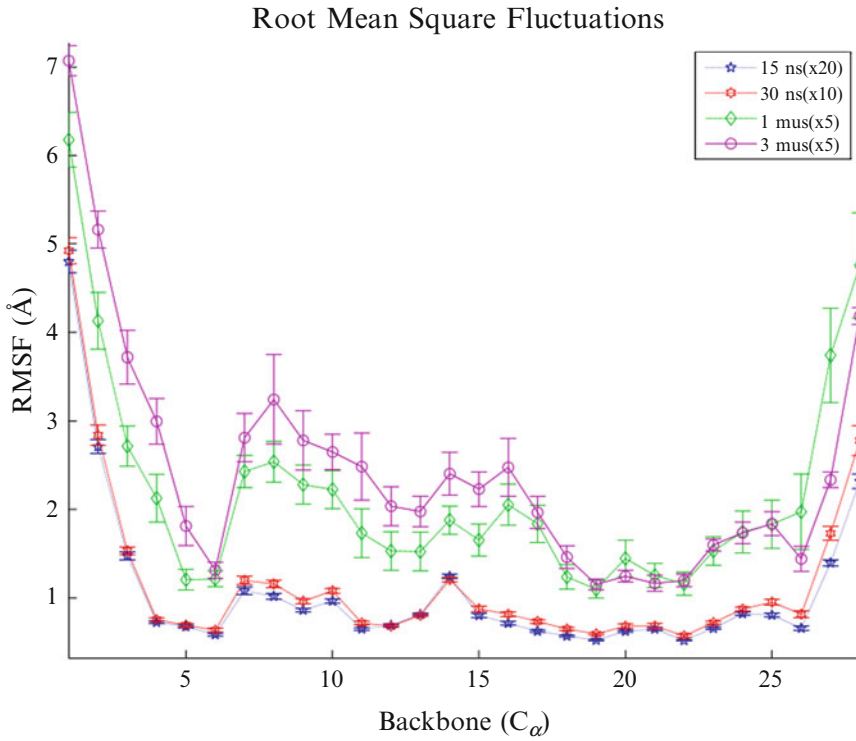

Fig. 7 A plot of RMSF for each alpha carbon for four different timescales of simulations showing variation in flexibility on a residue basis

4.5 Covariance and Correlation Analysis

Whereas RMSFs provide information about flexibility at the atomic level, they provide no information about coupled motions. Covariances, or their normalized counterparts, correlations, however, provide an indication of coupled dynamics by indicating what parts of the system show correlated motions; these could be motions in the same direction, in the opposite directions, or most often, uncorrelated motions (61). Covariance analysis is a useful technique for detecting the motions of a protein that might, for example, be responsible for a particular interaction, or to elucidate long-range interactions within a sample that may be responsible for allosteric regulation or other functional behavior. If r_i and r_j are position vectors of two atoms in the sample, then the covariance is calculated using

$$\widetilde{C}_{ij} = \sum_{\alpha=1}^{N} \frac{\left(\vec{r}_i^{\,\alpha} - \left\langle \vec{r}_i^{\,\alpha} \right\rangle\right) \cdot \left(\vec{r}_j^{\,\alpha} - \left\langle \vec{r}_j^{\,\alpha} \right\rangle\right)}{N}$$

Here N indicates the total number of frames and alpha is the index over each frame of the trajectory. These covariances are then typically normalized into correlations by dividing by the square root of the product of C_{ii} and C_{jj}, so that the diagonals are one; an atom always fluctuations with itself. In the molecular dynamics literature, correlation matrices, e.g., Fig. 8, are often referred to as covariance

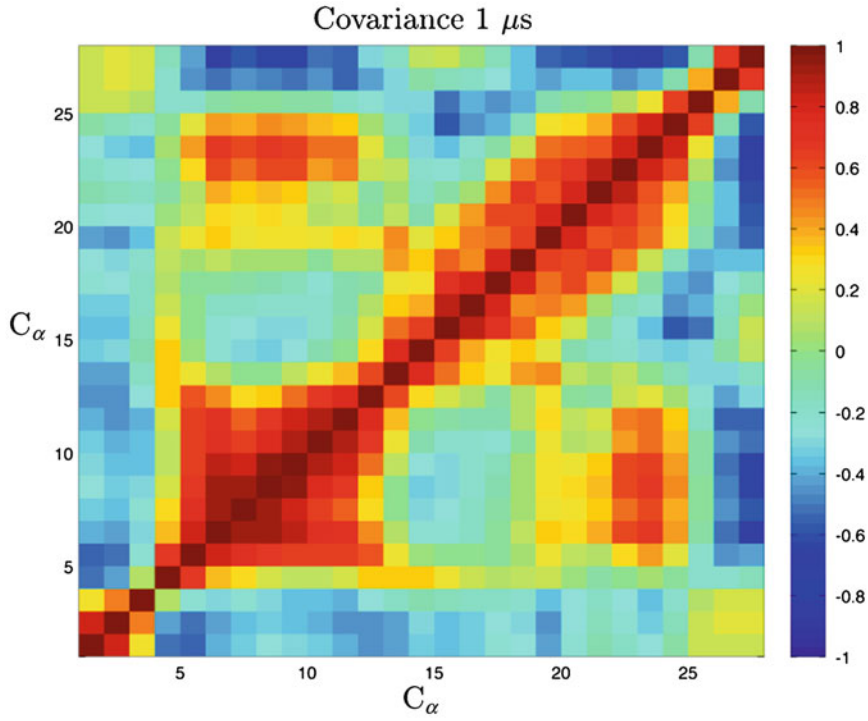

Fig. 8 Alpha carbon covariances for the same 28-residue zinc finger protein, NEMO, as in Fig. 7, as simulated on the microsecond timescale

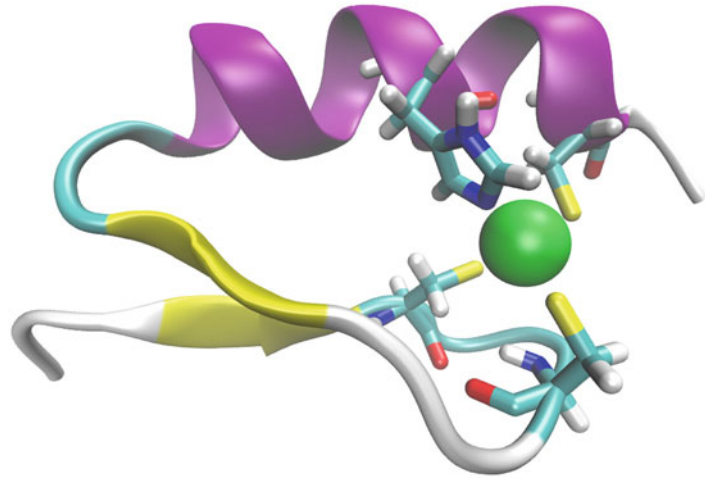

Fig. 9 Cartoon drawing of NEMO. Based on the striker 2JVX from the RCSB. The alpha helix in *magenta*, the beta sheet in *yellow* with the zinc in vdW representation in *green*. The binding residues of the zinc are in a bonded representation

matrices, where they should properly be referred to as correlation matrices. The difference between the two is that a covariance matrix has not been normalized so that the diagonal elements are one, whereas a correlation matrix is a normalized version of the covariance matrix. Such correlation matrices aka normalized covariance matrices are useful in extracting the essential degrees of freedom often hypothesized responsible for the primary function of a system (62). Additionally, entropy estimates, and subsequently heat capacity, can be derived from the covariance matrix using harmonic approximations (63, 64). Figure 8 provides an example of correlations for the alpha carbons a small 28-residue zinc finger, NEMO (60). In this example, there are some unsurprising covariances, those within the secondary structural elements, the beta structure (residues 5–12), and the alpha helix (residues 16–24), but also shows correlated motions between the beta sheet and portions of the alpha helix, and anti-correlated motions between the beta-sheet and the loop connecting the helix and the sheet., c.f. Fig. 9 for the structure of NEMO. Whether these are due to zinc binding is a subject of further study, but this illustrates that non-trivial covariances can exist in even small proteins, and provides an example of the sort of information available from correlations plots, even by just visual inspection.

5 Small Molecule Docking

The analysis tools discussed so far provide information about the conformations, fluctuations and dynamics of a protein or other macromolecular system. Clustering and Markov analysis have

particular uses also in syncing with small molecule docking for lead generations. Protein flexibility is a challenge for molecular docking. Rather than allow docking software to simulate small conformational changes, it is often more efficient to use MD software to assemble an ensemble of structures as initial structures for docking studies, since simulations of polymer flexing are squarely in the realm of MD. Once these conformations are selected from an MD trajectory, each one can be used as a starting structure for small molecule docking, for which there are multiple efficient software packages (65–72).

One such docking software, arguably the most efficient, that is popular in a wide variety of uses is AutoDock (73, 74). In 2004, an open source generation of the software was released—AutoDock Vina (75). Previous generation software focused on analytic approaches to docking. For example, AutoDock 4.2 calculated free energies of association for bound conformations using an empirical force field, Lamarckian Genetic Algorithm (76), and explicit modeling of side chains in receptors (77). While AutoDock Vina maintains some of these strategies, the energy function has been refined using machine learning and the PDBbind database (78, 79) of known binding affinities. The key advantage of AutoDock Vina is the calculation of both a scoring function and the gradient thereof. By calculating the gradient, the software knows which direction the next iteration of the search should move a given local set of atoms (75, 80–82). As a result, AutoDock Vina is far faster than previous software packages, while still retaining considerable precision.

In order to use docking software for identify potential lead candidates for drug discovery, ligand libraries must be used. While a small ligand library could be hand constructed using common molecular drawing and export software, a strategy to take advantage of the strength of molecular dynamics and of virtual screening would be to use a large general screening library of ligands. These can be obtained from various online chemical databases such as BindingDB (83), ChEMBL (84), DrugBank (85), PubChem (86), TCM Database@Taiwan (87), and ZINC (88, 89). This last database, ZINC, is arguably the best for general purpose docking and was originally developed with drug discovery in mind and has kept that focus while growing to sample 34,000,000 unique molecules from 134 commercial and 36 annotated catalogs. The attempt is to have a library of all commercially available compounds. This focus on drug discovery manifests itself in two ways. First, ZINC's structure files are generally selected for their biological relevance. Second, the subsets into which these structures are grouped have been curated with screening and discovery in mind so that datasets using standard definitions of "drug-like," "lead-like," and "fragment-like" are readily available for download. They also maintain subsets of these datasets containing "currently" available compounds;

usually updated every 6 months or so. Currently available means a delivery window of 0–10 weeks, with a target price of $100 or less per sample. Within each of the latter lead-like datasets, there are also different versions that have had different levels of similarity analysis performed on them. For example, one could download currently available lead-like molecules, from a set of 3,687,621 molecules at the time this was written. Or one could download a subset filtered at the 90 % similarity level—so that any two compounds that are more than 90 % similar are filtered out and only one selected—which is less than a tenth the size of the whole library; 322,638 molecules.

The ZINC database has other features that make it particularly useful for drug discovery using virtual screening. For example, to ensure the quality of structures and groupings, the creators of ZINC have what they term a "hit picking party" (89) from time to time, during which they run docking trials on structures in the ZINC database and compare the output structures to experimental data. Beyond providing structures for virtual screening, the physical compounds can be purchased through ZINC. For time-sensitive studies ZINC is able to sort purchasable compounds by estimated delivery time. While the physical compounds are sold commercially, searching for and downloading structure files are free services (89).

If it is not desirable to use a general library for virtual screening, for example, if there is a lead available and the goal is to refine or expand outward in the space of molecules, or if particular chemistries are desired, then libraries can be constructed combinatorially. That is particular core structures can be defined along with functional groups to modify those groups, and they can be combined computationally to generate a personalized library. There are commercial libraries that can be used for this purpose, but Simlib is a freely available code that can be used to easily generate libraries (90).

6 Timescale Considerations

It should not be surprising that longer timescale simulations require more computational resources, and choosing the appropriate timescale to simulate is an important consideration. The decision is motivated by the resources available and what is of biophysical interest. Nanosecond timescale simulations are valuable to elucidate low energy conformations and nearby fluctuations from that minimum. These simulations can be especially useful for identifying dynamical motion sufficient to hypothesize corresponding biological function (59, 91), and afford sufficient time to observe conformational changes such as motions of a lever arm, for example. These types of simulations may indeed be sufficient for lead identification (11, 92).

And yet, many important molecular processes occur at longer timescales, so if these are of primary interest, either for biological interest or to obtain rarer conformations for docking, then longer simulations are required. Larger conformational arrangements may require longer simulations as slower processes are responsible for some allosteric regulation or other conformational selection events (93). Even for small systems, such as a 28 residue zinc-finger motif, rare events, that are not available in shorter simulations, occur in the microsecond regime (60). These less-common events may well have significance for drug discovery, and the fact that they happen on the order of microseconds means they still happen 100s–1000s times per millisecond, and they may provide pockets for drug docking, as suggested by those who argue that allostery is an intrinsic property of all proteins (9).

7 Recent Applications of Molecular Dynamics and Docking

7.1 Molecular Dynamics and Docking: MSH2/6 and Rescinnamine

An example of pharmacological use of virtual screening via docking to clusters from MD simulations comes from a search for small molecules that would selectively bind to an apoptosis-inducing conformation of the MSH2/MSH6 proteins (92). These two proteins form a complex and recognize DNA defects resulting from improper replication. The proteins then enter either a repair conformation when the lesions are reparable or a death-signaling conformation when the lesions are irreparable (94–99). In both cases, this protein complex senses the defect or damage and recruits other proteins to either repair the defect or initiate cell death. It is also likely that there are multiple cell death pathways. The goal of this example study was to find cytotoxic agents that would bind to MSH2/MSH6 while the protein is in the death-signaling conformation, thereby triggering apoptosis (91, 99).

The researchers started with an X-ray structure of a complex of DNA and *Escherichia coli* MutS—a homolog of MSH—modified to include the cisplatin adduct cross linking DNA with hydrogens added via the CHARMM software package and solvated with TIP3P water using VMD's "Add Solvation Box" extension. They performed a 250 ps equilibration and subsequent 10 ns production run in NAMD with the CHARMM27 force field, having pressure set to 1.01325 bar and temperature to 300 K. Frames in the final trajectory were clustered using a 1 Å cutoff radius K-means clustering. They input the resulting ensemble of conformations into AutoDock 3 (see (76) for details of this version) for docking trials with a small library of commercially available compounds. They then tested the compounds with highest binding affinities for the E. coli MutS-DNA complex in vitro on MSH2/MSH6 and found that they could indeed use a selectively binding ligand to select out the death-signaling conformation of the proteins (92, 100).

Beyond their specific goal, this work demonstrated the predictive power of in silico molecular dynamics and virtual screening to select compounds for in vitro trials. Subsequent work building on this has focused on using nanoscale simulations to study further aspect of binding of damaged DNA to the human MSH2/6 (59, 101, 102).

7.2 Docking Without Molecular Dynamics

Docking trials on their own are powerful predictive tools for drug discovery. For example, they were used in the development of a novel phosphatidylinositol-3-kinase (PI3K) inhibitor that specifically targets prostate cancer (103). PI3K activity is currently understood to inhibit apoptosis of prostate cancer cells and allow them to continue multiplying even in local environments that would be unfavorable to healthy cells, specifically areas of low androgen concentration (104). In order to inhibit PI3K activity and allow induced apoptotic signaling, the researchers were searching for the most energetically favorable binding site on PI3K for a quercetin analog (LY294002). Using AutoDock 3 (76) with a pool of 30,000 initial dockings, the researchers found that the activated prodrug (L-O-CH2-LY294002) had a high affinity for PI3K's ATP binding site, leading them to conclude that the activated prodrug would indeed inhibit PI3K activity. They then confirmed experimentally that this compound was successful in inhibition and induced apoptosis in prostate cancer cells. Discovering this drug's affinity for PI3K's ATP binding site through 30,000 in vitro experiments would have been prohibitively expensive, in terms of both time and compound synthesis required, but was made tenable with the aid of in silico trials.

7.3 Sequence Similarity Motivates Drug Discovery with MD: Tamiflu and Relenza

A study of multiple drugs and their related proteins via MD simulation have proven useful in understanding how drug binding mechanisms work (29), and has implications for new drug discovery, as well as understanding mechanisms of drug resistance (105). In the Le study, the pathogenic avian H5N1 type-I neuraminidase, which is the target for drugs such as Tamiflu and Relenza, is compared to other sequence-similar proteins. These all-atom simulations investigated drug–protein interactions, including both conserved and unique interactions, with particular emphasis on hydrogen bonding and electrostatics. Their findings suggest how conserved networks of hydrogen bonds across the three structural variations elucidate a possible mechanism for how certain mutations might lead to drug resistance.

Such investigations on how mutations affect drug resistances or create genetic diseases are increasingly common, as conformational and dynamical changes comparing similar structures or systems highlight the specific mechanisms likely responsible for proper, or improper, function. This form of mechanism hypothesis is just one of the many ways MD simulations help to push the ability to generate effective drugs, namely, ones that are less susceptible to mutation based resistance.

8 Areas in Which Molecular Dynamics Shows Promise to Impact Drug Discovery

In addition to explicitly providing conformations, especially rarer conformations for ligand screening, there are multiple areas in which molecular dynamics appears poised to make an impact; these areas include protein–protein interfaces, an elusive, yet potentially profitable arenas for drug discovery, and in understanding more complex, longer-range allosteric interactions.

8.1 Protein–Protein Interfaces and Interactions via Molecular Dynamics: Peroxiredoxins

Another study involving molecular dynamics, enhanced by additional calculations, in this case electrostatic-based p*K*a calculations, have been ones on the peroxiredoxins (Prxs) shows how detailed p*K*a values, combined with MD simulation, can be combined to gain additional insight into the modeling chemical contributions to the oligomerization (106, 107). This family of proteins is responsible for catalyzing the reduction of hydrogen peroxide, alkyl hydroperoxides and peroxynitrite. They control levels of cytokine-induced peroxide and act as a regulator of signal transduction in mammals (108–110).

These interactions have proven elusive in the drug discovery process because of the complexity in finding information about specific sites for ligand binding (111), although progress is being made (56, 112). The transient nature of the PPI makes identification of regions for small molecule binding difficult to find, and understanding dynamics will be critical for effective predictions. Recent efforts combining NMR with MD have proven useful in this regard (112).

8.2 Molecular Dynamics and Long-Range Allostery: MetRS/tRNA Complexes

In a different research study, the authors used MD to probe action at a distance allostery involved in enzymatic reactions (113). They simulated E. coli methionyl-tRNA synthase (MetRS) with 9 ns long trajectories, using the CHARMM27 force field and a TIP3P explicit water model. Using RMSFs, they compared simulation results directly against experimental values from X-ray crystallography and compared global mobility for the different mutations. Additionally, covariance analysis helped reveal how certain amino acid substitutions alter the conformations and dynamics. Through this careful analysis of the results, the authors showed that substitution of a certain tryptophan residue (trp-461) results in specific changes in protein correlation and dynamics. Effects of this substitution to the local region are not surprising, but there is clear evidence that the mobility-correlated motions of a region 40–50 Å away are reduced in the absence of the conserving tryptophan. It appears that the conserved residue has function in addition to codon recognition that is mediating the conformational structures available to the protein. These simulations along with

previous ones discussed (59, 91, 101, 102) show the utility of nanosecond scale simulations to study protein dynamics around the ground state of the folded structure.

Here we see another example of how MD is used to inform the drug discovery process. While this particular system might not have direct application for drug discovery, it is useful in showing how MD can help us realize the mechanism of action in these extremely complex, highly dimensional systems. It is difficult to conceive a better method of understanding elusive biomolecular processes, such as this type of hidden allostery, than MD simulations.

9 Further Reading

Although we have surveyed the basics of molecular dynamics as it pertains to drug discovery and discussed several illustrative studies, there are other articles which review different aspects of molecular dynamics simulations and their use in understanding drug–protein interactions, binding mechanisms, and protein–protein interactions. Additional case studies can be found in (12), and a discussion of homology modeling and a possible method for accelerating molecular dynamics can be found in (114).

The use of homology modeling in molecular simulation and drug discovery is somewhat controversial. However, homology has gained traction in recent years, as a means to alleviate the time consuming and expensive tasks of X-ray crystallography and NMR, especially in protein families for which there are many structures available (114). In addition becoming the basis for iterative processes for model refinement as more information becomes available, homology is commonly used for comparison of families of proteins for bioinformatics. As discussed in (114) homology has been used to help fill the gap between sequence and structure, for G-protein coupled receptors (GCPRs). They are among the most prominent targets for small molecule drugs, and, due to this abundance, they are a prime target for homology modeling as a means to structure based drug design. While there are roughly 100 GPCR structures in the PDB there are still many more in the family, and due to their popularity as a drug target for a variety of disease, this problem is particularly well suited to homology modeling, despite the difficulties (115, 116). There is even some evidence that homology modeling is more successful with GPCRs than de novo techniques (117). The success of homology modeling provides hope that computationally techniques will be able to draw upon the increasing number of experimental structures available to apply molecular dynamics and virtual screening techniques to the even larger protein universe, which is increasing even faster.

Additionally, Kalyaanamoorthy discusses implementations of enhanced sampling techniques to access longer timescales.

One such technique is promising, namely, that of random acceleration molecular dynamics (RAMD). Used for investigating ligand dissociation, RAMD applies a small, additional random force to the center of mass of the ligand allowing it to search the protein for potential egresses. This technique is appealing because it can provide ligand dissociation information on nanosecond timescales, it unbiasedly searches possible molecular channels, due to the stochastic nature of the extra force, and may be able to help identify key amino-acid residues in the ligand (un)binding process. Although a disadvantage of course is that this requires a structure with a ligand, but could be used, for example, to study conformational changes that occur due to ligand escape which could then be used to inform, or even to provide structures, for virtual screening. A complementary technique is steered MD (SMD) and it is analogous to an atomic force microscopy or optical tweezers, in silico. Much like RAMD, it could be used to look at small molecule dissociation. Both of these techniques show promise in further enabling drug discovery by using experimental structures with ligands, and providing information about the binding process.

One of the remaining challenges for the field is the perception of the scientific community. It seems that confidence in MD simulations has not gained the traction that it has in other scientific disciplines such as meteorology, fluid dynamics and astrophysics (118). Despite the success in other disciplines, MD simulations are often disregarded as insufficient, even though there is a wealth of data showing consistent results between simulation and experiment including many of the examples reviewed herein.

References

1. Mccammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590
2. Radkiewicz JL, Brooks CLI (2000) Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase. J Am Chem Soc 122:225–231. doi:10.1021/ja9913838
3. Salsbury F (2001) Modeling of the metallo-β-lactamase from B. fragilis: structural and dynamic effects of inhibitor binding. Proteins 44:448–459
4. Salsbury FR, Crowder MW, Kingsmore SF, Huntley JJ (2009) Molecular dynamic simulations of the metallo-beta-lactamase from Bacteroides fragilis in the presence and absence of a tight-binding inhibitor. J Mol Model 15:133–145. doi:10.1007/s00894-008-0410-0
5. Kumar S, Ma B, Tsai C-J et al (2000) Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci 9:10–19
6. Freire E (1999) The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody. Proc Natl Acad Sci 96:10118–10122
7. Kern D, Zuiderweg ER (2003) The role of dynamics in allosteric regulation. Curr Opin Struct Biol 13:748–757. doi:10.1016/j.sbi.2003.10.008
8. Pan H, Lee JC, Hilser VJ (2000) Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. Proc Natl Acad Sci U S A 97:12020–12025. doi:10.1073/pnas.220240297
9. Gunasekaran K, Ma B, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57:433–443. doi:10.1002/prot.20232
10. Tsai C, Ma B, Nussinov R (1999) Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci U S A 96:9970–9972

11. Vasilyeva A, Clodfelter JE, Rector B et al (2009) Small molecule induction of MSH2-dependent cell death suggests a vital role of mismatch repair proteins in cell death. DNA Repair (Amst) 8:103–113. doi:10.1016/j.dnarep.2008.09.008
12. Salsbury FR (2010) Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. Curr Opin Pharmacol 10:738–744. doi:10.1016/j.coph.2010.09.016
13. Berman HM (2000) The protein data bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235
14. Saxena A, Sangwan RS, Mishra S (2013) Fundamentals of homology modeling steps and comparison among important bioinformatics tools: an overview. Sci Int 1:237–252. doi:10.5567/sciintl.2013.237.252
15. Szalewicz K (2014) Determination of structure and properties of molecular crystals from first principles. Acc Chem Res 47:3266–3274. doi:10.1021/ar500275m
16. MacKerell A, Bashford D (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem 5647:3586–3616. doi:10.1021/jp973084f
17. Ponder JW, Case DA (2003) Force fields for protein simulations. Adv Protein Chem 66:27–85. doi:10.1016/S0065-3233(03)66002-X
18. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656–1676. doi:10.1002/jcc.20090
19. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N log(N) method for Ewald sums in large systems. J Chem Phys 12:10089–10092
20. Roe DR, Okur A, Wickstrom L et al (2007) Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. J Phys Chem B 111:1846–1857. doi:10.1021/jp066831u
21. García AE, Sanbonmatsu KY (2002) Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc Natl Acad Sci U S A 99:2782–2787. doi:10.1073/pnas.042496899
22. Feig M, MacKerell AD, Brooks CL (2003) Force field influence on the observation of π-helical protein structures in molecular dynamics simulations. J Phys Chem B 107:2831–2836. doi:10.1021/jp027293y
23. Lee MS, Salsbury FR, Brooks CL (2002) Novel generalized born methods. J Chem Phys 116:10606. doi:10.1063/1.1480013
24. The 2013 Nobel Prize in Chemistry
25. Phillips JC, Braun R, Wang W et al (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802. doi:10.1002/jcc.20289
26. Brooks B, Brooks C (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614. doi:10.1002/jcc.21287.CHARMM
27. Case DA, Cheatham TE, Darden T et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688. doi:10.1002/jcc.20290
28. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447. doi:10.1021/ct700301q
29. Le L, Lee E, Schulten K, Truong TN (2009) Molecular modeling of swine influenza A/H1N1, Spanish H1N1, and avian H5N1 flu N1 neuraminidases bound to Tamiflu and Relenza. PLoS Curr 1:RRN1015. doi:10.1371/currents.RRN1015
30. De Meyer FJ-M, Venturoli M, Smit B (2008) Molecular simulations of lipid-mediated protein-protein interactions. Biophys J 95:1851–1865. doi:10.1529/biophysj.107.124164
31. Harvey M, Giupponi G, Fabritiis G (2009) ACEMD: accelerating biomolecular dynamics in the microsecond time scale. J Chem Theory Comput 5:1632
32. Schlick T (2010) Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide. Springer Science & Business Media, New York, NY
33. Frenkel D, Smit B (2001) Understanding molecular simulation: from algorithms to applications. Academic, San Diego, CA
34. Fenimore PW, Frauenfelder H, Mcmahon BH, Parak FG (2002) Slaving: solvent fluctuations dominate protein dynamics and functions. Proc Natl Acad Sci U S A 99:16047–16051
35. Frauenfelder H, Fenimore PW, Young RD (2007) Protein dynamics and function: insights from the energy landscape and solvent slaving. IUBMB Life 59:506–512. doi:10.1080/15216540701194113

36. Tarek M, Tobias DJ (2000) The dynamics of protein hydration water: a quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. Biophys J 79:3244–3257. doi:10.1016/S0006-3495(00)76557-X
37. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Springer, Berlin, pp 331–342
38. Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926. doi:10.1063/1.445869
39. Zhou R (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins 161:148–161
40. Mark P, Nilsson L (2001) Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J Phys Chem A 105:9954–9960. doi:10.1021/jp003020w
41. Ryckaert J, Ciccotti G, Berendsen H (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 341:327–341
42. Knight JL, Brooks CL (2011) Surveying implicit solvent models for estimating small molecule absolute hydration free energies. J Comput Chem 32:2909–2923. doi:10.1002/jcc.21876
43. Pu M, Garrahan JP, Hirst JD (2011) Comparison of implicit solvent models and force fields in molecular dynamics simulations of the PB1 domain. Chem Phys Lett 515:283–289. doi:10.1016/j.cplett.2011.09.026
44. Zhou R, Berne B (2002) Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? Proc Natl Acad Sci U S A 99:12777
45. Lee MS, Feig M, Salsbury FR, Brooks CL III (2003) New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. J Comput Chem 24:1348
46. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33
47. Heyer L, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9:1106–1115
48. Karpen M, Tobias D, Brooks C III (1993) Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. Biochemistry 32:412–420
49. Brooks BR, Bruccoleri RE, Olafson BD et al (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217. doi:10.1002/jcc.540040211
50. Carpenter GA, Grossberg S (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. Comput Vis Graph Image Process 37:54–115. doi:10.1016/S0734-189X(87)80014-2
51. Pao Y-H (1989) Adaptive pattern recognition and neural networks. Addison-Wesley Longman Publishing Co., Inc., Boston, MA
52. Senne M, Schütte C, Noé F (2012) EMMA: a software package for Markov model building and analysis. J Chem Theory Comput 8:2223–2228
53. Beauchamp K, Bowman G (2011) MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. J Chem Theory Comput 7:3412–3419. doi:10.1021/ct200463m.MSM Builder2
54. Cronkite-Ratcliff B, Pande V (2013) MSMExplorer: visualizing Markov state models for biomolecule folding simulations. Bioinformatics 29:950–952. doi:10.1093/bioinformatics/btt051
55. Pande V, Beauchamp K, Bowman G (2010) Everything you wanted to know about Markov State Models but were afraid to ask. Methods 52:99–105. doi:10.1016/j.ymeth.2010.06.002.Everything
56. Kozakov D, Hall DR, Chuang G-Y et al (2011) Structural conservation of druggable hot spots in protein-protein interfaces. Proc Natl Acad Sci U S A 108:13528–13533. doi:10.1073/pnas.1101835108
57. Peng JW (2009) Communication breakdown: protein dynamics and drug design. Structure 17:319–320. doi:10.1016/j.str.2009.02.004
58. Mauldin RV, Carroll MJ, Lee AL (2009) Dynamic dysfunction in dihydrofolate reductase results from antifolate drug binding: modulation of dynamics within a structural state. Structure 17:386–394. doi:10.1016/j.str.2009.01.005
59. Negureanu L, Salsbury FR (2012) Insights into protein - DNA interactions, stability and allosteric communications: a computational study of mutSα-DNA recognition complexes. J Biomol Struct Dyn 29:757–776. doi:10.1080/07391102.2012.10507412
60. Godwin RC, Gmeiner WH, Salsbury FR (2015) Importance of long-time simulations

for rare event sampling in zinc finger proteins. J Biomol Struct Dyn (In press). doi:10.1080/07391102.2015.1015168

61. Ichiye T, Karplus M (1991) Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins 11:205–217
62. Amadei A, Linssen A, Berendsen HJC (1993) Essential dynamics of proteins. Proteins 17:412–425
63. Schäfer H, Mark AE, Van Gunsteren WF (2000) Absolute entropies from molecular dynamics simulation trajectories. J Chem Phys 113:7809–7817. doi:10.1063/1.1309534
64. Andricioaei I, Karplus M (2001) On the calculation of entropy from covariance matrices of the atomic fluctuations. J Chem Phys 115:6289–6292. doi:10.1063/1.1401821
65. Huang Z, Wong C (2009) Docking flexible peptide to flexible protein by molecular dynamics using two implicit-solvent models: an evaluation in protein kinase and phosphatase systems. J Phys Chem B 113:14343–14354. doi:10.1021/jp907375b.Docking
66. Knegtel RM, Kuntz ID, Oshiro CM (1997) Molecular docking to ensembles of protein structures. J Mol Biol 266:424–440. doi:10.1006/jmbi.1996.0776
67. Claussen H, Buning C, Rarey M, Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 308:377–395. doi:10.1006/jmbi.2001.4551
68. Lin J-H, Perryman AL, Schames JR, McCammon JA (2002) Computational drug design accommodating receptor flexibility: the relaxed complex scheme. J Am Chem Soc 124:5632–5633. doi:10.1021/ja0260162
69. Frembgen-Kesner T, Elcock AH (2006) Computational sampling of a cryptic drug binding site in a protein receptor: explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase. J Mol Biol 359:202–214. doi:10.1016/j.jmb.2006.03.021
70. Tatsumi R, Fukunishi Y, Nakamura H (2004) A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. J Comput Chem 25:1995–2005. doi:10.1002/jcc.20133
71. Lin J, Perryman A (2003) The relaxed complex method: accommodating receptor flexibility for drug design with an improved scoring scheme. Biopolymers 68:47–62
72. Totrov M, Abagyan R (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins 220:215–220
73. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202. doi:10.1002/prot.340080302
74. Goodsell D (1996) Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 9:1–5
75. Trott O, Olson AJ (2010) Software news and update AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461. doi:10.1002/jcc
76. Morris G, Goodsell D (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
77. Morris G, Huey R (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791. doi:10.1002/jcc
78. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 47:2977–2980. doi:10.1021/jm030580l
79. Wang R, Fang X, Lu Y et al (2005) The PDBbind database: methodologies and updates. J Med Chem 48:4111–4119. doi:10.1021/jm048957q
80. Abagyan R, Totrov M, Kuznetsov D (1994) ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15:488–506
81. Blum C, Roli A, Sampels M (2008) Hybrid metaheuristics: an emerging approach to optimization. Springer, New York, NY
82. Nocedal J, Wright SJ (1999) Numerical optimization. Springer, New York, NY
83. Liu T, Lin Y, Wen X et al (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35:D198–D201. doi:10.1093/nar/gkl999
84. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. doi:10.1093/nar/gkr777
85. Knox C, Law V, Jewison T et al (2011) DrugBank 3.0: a comprehensive resource for "omics" research on drugs. Nucleic Acids Res 39:D1035–D1041. doi:10.1093/nar/gkq1126

86. Li Q, Cheng T, Wang Y, Bryant S (2010) PubChem as a public resource for drug discovery. Drug Discov Today 15:1052–1057. doi:10.1016/j.drudis.2010.10.003. PubChem
87. Chen C (2011) TCM Database@ Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS One 6:e15939. doi:10.1371/journal.pone.0015939
88. Irwin JJ, Shoichet BK (2005) ZINC - a free database of commercially available compounds for virtual screening. J Chem Inf Model 36:177–182. doi:10.1002/chin.200516215
89. Irwin JJ, Sterling T, Mysinger MM et al (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52:1757–1768. doi:10.1021/ci3001277
90. Schüller A, Hähnke V, Schneider G (2007) SmiLib v2.0: a Java-based tool for rapid combinatorial library enumeration. QSAR Comb Sci 26:407–410. doi:10.1002/qsar.200630101
91. Salsbury FR, Clodfelter JE, Gentry MB et al (2006) The molecular mechanism of DNA damage recognition by MutS homologs and its consequences for cell death response. Nucleic Acids Res 34:2173–2185. doi:10.1093/nar/gkl238
92. Vasilyeva A, Clodfelter JE, Gorczynski MJ, et al. (2010) Parameters of reserpine analogs that induce MSH2/MSH6-dependent cytotoxic response. J Nucleic Acids, Article ID 162018, doi: 10.4061/2010/162018
93. Lange O, Lakomek N, Farès C (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320:1471–1475
94. Stojic L, Brun R, Jiricny J (2004) Mismatch repair and DNA damage signalling. DNA Repair (Amst) 3:1091–1101. doi:10.1016/j.dnarep.2004.06.006
95. Fishel R, Wilson T (1997) MutS homologs in mammalian cells. Curr Opin Genet Dev 7:105–113
96. Kolodner RD, Marsischky GT (1999) Eukaryotic DNA mismatch repair. Curr Opin Genet Dev 9:89–96. doi:10.1016/S0959-437X(99)80013-6
97. Bellacosa A (2001) Functional interactions and signaling properties of mammalian DNA mismatch repair proteins. Cell Death Differ 8:1076–1092. doi:10.1038/sj.cdd.4400948
98. Kunkel TA, Erie DA (2005) DNA mismatch repair. Annu Rev Biochem 74:681–710. doi:10.1146/annurev.biochem.74.082803.133243
99. Drotschmann K, Topping RP, Clodfelter JE, Salsbury FR (2004) Mutations in the nucleotide-binding domain of MutS homologs uncouple cell death from cell survival. DNA Repair (Amst) 3:729–742. doi:10.1016/j.dnarep.2004.02.011
100. Abdelhafez OM, Amin KM, Ali HI, Abdalla M, Ahmed EY (2014) RSC Adv 4:11569–11579. doi:10.1039/c4ra00943f
101. Negureanu L, Salsbury FR (2012) The molecular origin of the MMR-dependent apoptosis pathway from dynamics analysis of MutSα-DNA complexes. J Biomol Struct Dyn 30:1–15. doi:10.1080/07391102.2012.680034
102. Negureanu L, Salsbury FR (2014) Non-specificity and synergy at the binding site of the carboplatin-induced DNA adduct via molecular dynamics simulations of the MutSα-DNA recognition complex. J Biomol Struct Dyn 32:969–992. doi:10.1080/07391102.2013.799437
103. Baiz D, Pinder T, Hassan S (2012) Synthesis and characterization of a novel prostate cancer-targeted phosphatidylinositol-3-kinase inhibitor prodrug. J Med Chem 55:8038–8046. doi:10.1021/jm300881a
104. Cohen MB, Rokhlin OW (2009) Mechanisms of prostate cancer cell survival after inhibition of AR expression. J Cell Biochem 106:363–371. doi:10.1002/jcb.22022
105. Woods CJ, Malaisree M, Pattarapongdilok N et al (2012) Long time scale GPU dynamics reveal the mechanism of drug resistance of the dual mutant I223R/H275Y neuraminidase from H1N1-2009 influenza virus. Biochemistry 51:4364
106. Yuan Y, Knaggs MH, Poole LB et al (2010) Conformational and oligomeric effects on the cysteine pK(a) of tryparedoxin peroxidase. J Biomol Struct Dyn 28:51–70. doi:10.1080/07391102.2010.10507343
107. Salsbury FR, Yuan Y, Knaggs MH et al (2012) Structural and electrostatic asymmetry at the active site in typical and atypical peroxiredoxin dimers. J Phys Chem B 116:6832–6843. doi:10.1021/jp212606k
108. Rhee SG, Kang SW, Jeong W et al (2005) Intracellular messenger function of hydrogen peroxide and its regulation by peroxiredoxins. Curr Opin Cell Biol 17:183–189. doi:10.1016/j.ceb.2005.02.004
109. Sue GR, Ho ZC, Kim K (2005) Peroxiredoxins: a historical overview and speculative preview of novel mechanisms and emerging

concepts in cell signaling. Free Radic Biol Med 38:1543–1552. doi:10.1016/j.freeradbiomed.2005.02.026

110. Wood ZA, Schroder E, Robin Harris J, Poole LB (2003) Structure, mechanism and regulation of peroxiredoxins. Trends Biochem Sci 28:32–40. doi:10.1016/S0968-0004(02)00003-8
111. Kube S, Weber M (2007) A coarse graining method for the identification of transition rates between molecular conformations. J Chem Phys 126:024103. doi:10.1063/1.2404953
112. Bernini A, Henrici De Angelis L, Morandi E et al (2014) Searching for protein binding sites from Molecular Dynamics simulations and paramagnetic fragment-based NMR studies. Biochim Biophys Acta 1844:561–566. doi:10.1016/j.bbapap.2013.12.012
113. Budiman ME, Knaggs MH, Fetrow JS, Alexander RW (2007) Using molecular dynamics to map interaction networks in an aminoacyl-tRNA synthetase. Proteins 68:670–689. doi:10.1002/prot.21426
114. Kalyaanamoorthy S, Chen Y-PP (2014) Modelling and enhanced molecular dynamics to steer structure-based drug discovery. Prog Biophys Mol Biol 114:123–136. doi:10.1016/j.pbiomolbio.2013.06.004
115. Kobilka BK (2007) G protein coupled receptor structure and activation. Biochim Biophys Acta 1768:794–807. doi:10.1016/j.bbamem.2006.10.021
116. Patny A, Desai PV, Avery MA (2006) Homology modeling of G-protein-coupled receptors and implications in drug design. Curr Med Chem 13:1667–1691. doi:10.2174/092986706777442002
117. Bhattacharya S, Lam AR, Li H et al (2013) Critical analysis of the successes and failures of homology models of G protein-coupled receptors. Proteins 81:729–739. doi:10.1002/prot.24195
118. Borhani DW, Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 26:15–26. doi:10.1007/s10822-011-9517-y