

Chapter 5

In Silico Drug Design and Mechanistic Enzymology

The robust prediction of protein structure outlined in the previous two chapters ties directly into our ability to model to rationally develop highly specific therapeutic drugs and to understand enzymatic mechanisms. Calculations of affinities with which drug molecules bind to proteins important in metabolic reactions and molecular signaling processes, and the mechanisms of catalytic function of certain key enzymes, can be obtained with computational approaches outlined in this chapter.

An important future development is that much of enzymology can be simulated

using methods based on essential physical laws found in common with computational chemistry, materials science, and combustion. However, there are many challenges to the simulations of these biological processes that arise from the large molecular sizes and long timescales relevant for biological function, or the subtle energetics and complex milieu of biochemical reactions. These challenges are greatly amplified by the need for high-throughput efforts to annotate the large number of genome sequences to include an understanding of their biological mechanisms and functions.

Only a small percentage of a pool of viable drug candidates actually lead to the identification of a clinically useful compound, with typically over \$200 million spent in research costs to successfully bring it to market. On average, a period of 12 years elapses between the identification and FDA approval of a successful drug, with the major bottleneck being the generation of novel, high-quality drug candidates. While rational, computer-based methods represent a quantum leap forward for identifying drug candidates, substantial increases in compute power are needed to allow for both greater selection sensitivity and genome-scale modeling of future drugs.

Some of these same computational methods and bottlenecks arise in the evaluation of the binding affinity of drugs to specific targets, but the task of the drug designer is further complicated by the need to identify a small group of compounds out of a virtually limitless universe of combinatorial possibilities.

In rational ligand-docking approaches, the interaction between two molecules is evaluated by computer simulation to quickly identify compounds that bind to a target with high affinity and specificity. Enormous chemical databases must be rapidly sifted through in order to identify a

small number of candidate drugs which can be further evaluated by in vitro assays and animal studies. Using this technique, it is feasible to evaluate hundreds of thousands of potential compounds within a matter of months or weeks, and model customized features such as improved binding geometry and pharmacokinetic properties.

While rational, computer-based methods represent a quantum leap forward for identifying drug candidates, substantial increases in compute power are needed to allow for both greater selection sensitivity and genome-scale modeling of future drugs.

Simulating the Reaction Mechanism of Malate Dehydrogenase

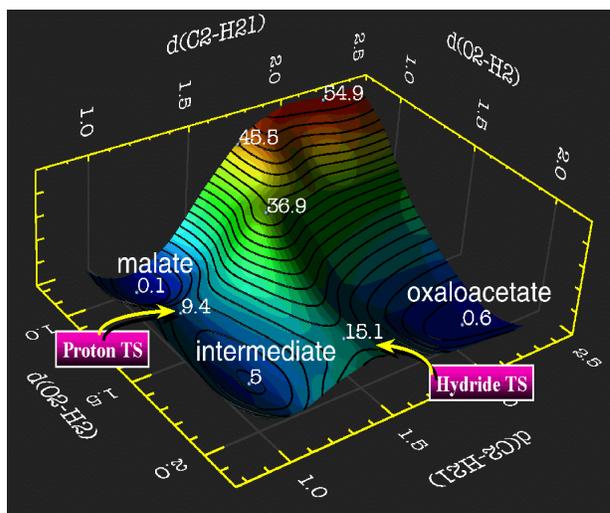
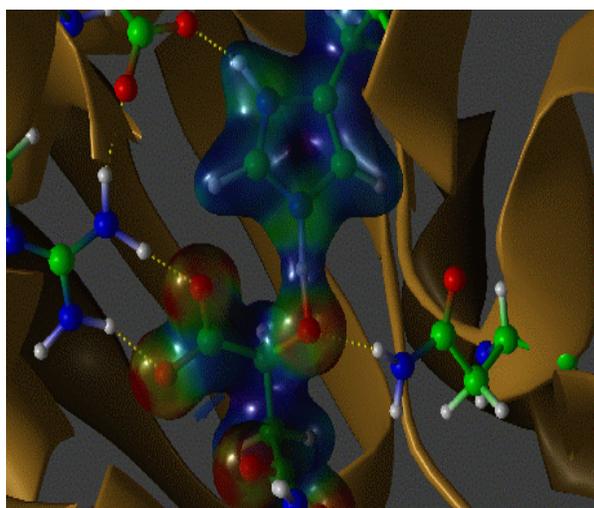
The minimum energy surface and reaction pathway for the interconversion of malate and oxaloacetate catalyzed by Malate Dehydrogenase (MDH) was simulated using semi-empirical QM/MM methods. Analysis of the energy profile shows that solvent effects due to the protein matrix dramatically alter the intrinsic reactivity of the functional groups involved in the MDH reactions. The enzyme effectively changes the reaction from an exothermic reaction in the gas phase to a nearly isoenergetic one in the protein-solvent environment of MDH.

The minimum energy profile was determined by 675 separate energy minimizations, consisting of 1000 steps of Adopted Basis Newton-Raphson each, using parallel computers at NERSC and ANL. Each processor was assigned an MDH model with a different and independent set of distance parameters that define the reaction mechanism in the enzyme. The

resultant energy profile showed that the MDH enzyme reaction is sequential, with the proton transfer preceding the hydride transfer.

In addition to giving a detailed mechanistic description of the MDH reaction, computer experiments with a QM/MM approach can also provide insights into the reasons that MDH is able to effectively catalyze the interconversion of malate and oxaloacetate. In particular, examination of the reaction profile shows that electrostatic effects in specific regions may enhance the transfer potential.

We anticipate that improved methods to solve Schrodinger's equation, coupled with advances in computer technology, will provide the means to simulate the electronic properties of complex systems at unprecedented levels of accuracy and reliability.



Proton transfer transition state (left) and minimum energy surface for the proton and hydride transfer reactions (right) in the enzyme Malate Dehydrogenase.

Predictions of Damage-Recognition and Mechanisms for DNA-Repair Enzymes

Human apurinic endonuclease (APE) is a DNA-repair enzyme that plays an essential role in maintaining the fidelity of the DNA sequence by recognizing and repairing sites where a single DNA base has been lost from the sequence. The structure and function of APE has been extensively studied, producing detailed kinetics data on a wide variety of substrates and an X-ray crystal structure for APE and related bacterial endonucleases.

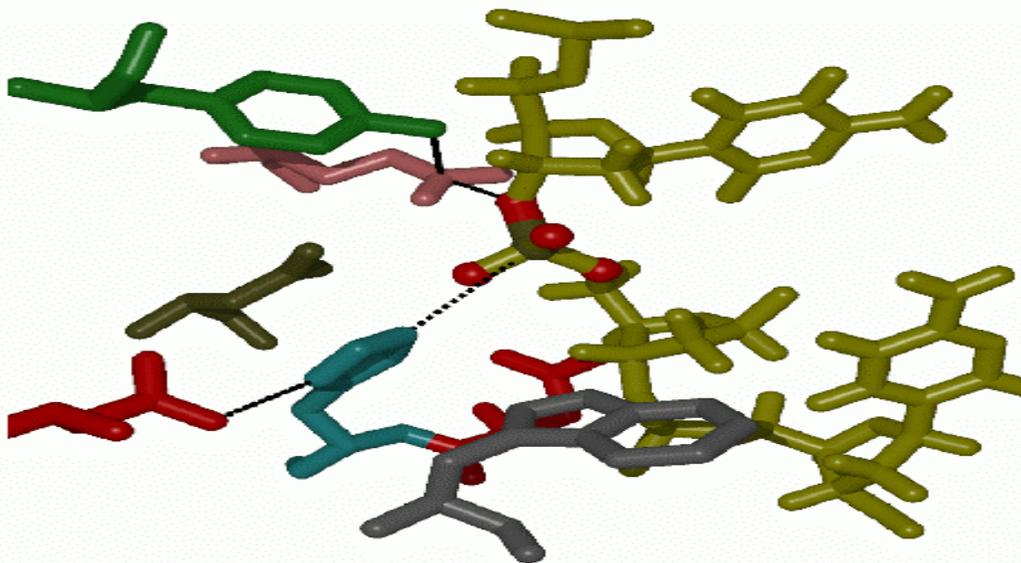
Nevertheless, the mechanism by which APE recognizes the DNA is unknown. Numerous mechanisms have been proposed, including one that APE acts by inserting an amino acid into the abasis site or by recognizing the bare deoxyribose ring. Recent site-directed mutation results and kinetic studies using synthesized abasic DNA substrates show that these mechanisms for APE damage recognition are incorrect.

Instead a mechanistically simpler hypothesis has been proposed that involves the conformational flexibility of the damaged DNA after it is bound to the repair enzyme. Since there is no experimental method for

directly monitoring such a mechanism, simulation is required to directly validate such an hypothesis.

Beyond the elucidation of the damage-recognition mechanism of APE, a detailed simulation will insight into the general mechanisms of the DNA repair enzymes. For example, a particularly interesting property of APE is that for certain damaged DNA substrates (those lacking the ribose ring), Mg^{++} ion is required for the catalytic activity, while for others it is not. MD simulations involving first principles force fields can help determine the location and role these ions have in the APE active site.

These quantitative models of the active site of APE will provide predictions for experimental validation, and eventually design, including suggestions for single amino acid mutations that would increase or decrease the reaction rate, and new damaged DNA substrates with novel synthetic backbones based on their backbone flexibility and activity in the simulated active site.



Models and Algorithms for Computational Enzymology

Modeling enzyme reactions at present involves two simulation approaches. One is semi-empirical or *ab initio* quantum mechanical (QM) methods that possess sufficient or even high accuracy for various biochemical properties of interest, but are currently limited to relatively small chemical systems and non-dynamic simulations. Another is classical molecular dynamics (MD), which simulates the motions of atoms in their chemical context for relatively large systems and long timescales, but with empirical force fields that often have insufficient accuracy, and altogether fail to treat the breaking and forming of bonds, that is especially important for enzymatic reactions. We have outlined much of the methodological and computing kernels of classical MD algorithms in the previous chapter, and focus this section on QM methods.

The simplest level of *ab initio* QM simulation is the Hartree-Fock (HF) method. This method produces very accurate bond lengths and angles and reasonable reaction energies. Potentially more accurate structures and reaction energies can be determined with Density Functional Theory (DFT) that shares HF's favorable scaling properties. Promising new algorithms, such as the MP2 method, should allow for very accurate energetic calculations on the chemically significant segments of many biochemical reactions. However, for certain properties, such as reaction barriers that are particularly important in non-equilibrium biochemical processes, more sophisticated QM methods such as CCSD and CCSD(T) may be required.

If we consider the series of theoretical models, HF or DFT methods, MP2, CCSD, CCSD(T), for a given size basis set, and varying molecular size, M , then in the simplest analysis their computational requirements scale as M^4 , M^5 , M^6 , and M^7 respectively. However, recent research has contributed to a

rapid breaking down of the computational bottlenecks in HF, DFT and MP2 calculations.

Two steps are involved in one HF/DFT energy and derivative calculation. The first step is the construction of the effective one-electron Hamiltonian matrix, usually termed the Fock matrix, given a density matrix. The second is the evaluation of a new density matrix, usually via the generation of new molecular orbitals or Kohn-Sham orbitals.

That HF and DFT methods naively scale as the fourth power of molecular size arises because of the evaluation of electron-electron interactions via *four* center two electron integrals. However, the number of non-negligible two electron integrals does not grow quartically with the size of the molecule, but grows as M^2 when the molecular size is large enough (i.e. the two atomic orbitals (AO's) comprising each pair must overlap in order to make a distribution containing non-negligible charge). This realization, together with advances in the speed of two-electron integral evaluation (integrals are generated as they are needed rather than stored), combine to permit routine calculations on systems approaching the 100 heavy (i.e. non-hydrogen) atom range.

The next generation of quantum chemistry algorithms will exploit new theories and technology that will reduce the scaling requirements of the HF, DFT, and MP2 methods. For example, linear scaling in the assembly of the Fock matrix follows directly from the collectivization of distant electron-electron interactions via multipole expansions with controlled error bars known as Fast Multipole Methods. In the face of linear scaling methods for electron integral evaluation, the generation of a new density matrix via diagonalization that scales as M^3 will eventually become dominant for large molecular sizes. Current effort has been directed toward methods for updating the density and/or orbitals without explicit

diagonalization, taking advantage of the fact that most molecules, such as proteins, the density matrix is spatially localized.

It is important to emphasize that DFT and existing functionals capture only certain types of electron correlation, and therefore the quality of DFT calculations are still under debate. We note that the development of new DFT functionals is an active area of research. A potentially feasible alternative is the MP2 method that is the simplest wavefunction-based theory of electron correlation. In most current quantum chemistry program packages MP2 scales as M^5 ; the M^5 scaling is a consequence of the formulation of MP2 using delocalized MO's which arise from standard HF calculations.

However, the MO's can be localized, and there has been some preliminary progress towards developing versions of MP2 theory based on localized orbitals. The "local-MP2" method scales only quadratically with molecular size, and comes to within a few percent of reproducing the exact MP2 energy with a given basis.

The simulation of certain enzyme catalyzed reaction mechanisms involving homolytic bond breaking (i.e. the breaking of electron pairs) or if transition metal atoms are involved in the active site, will require more accurate electron-correlated quantum chemical methods, e.g. coupled cluster (CCSD) that presently scale as M^6 - M^7 . The development of parallelized electron-correlated QM methods under the Strategic Simulation plan will allow the application of computational chemistry to these more complex enzyme mechanisms. For example a CCSD energy calculation should be feasible for a 40 atom system on a teraFLOP computer, which is sufficiently large to include a typical enzyme substrate and several catalytic amino acid residues.

Despite the great value of the static properties that can be calculated using QM

methods, many biological processes are inherently dynamical. Such problems include processive reactions (DNA or protein synthesis), and processes such as macromolecular conformational changes (DNA unwinding and allosteric enzyme regulation). Empirical force fields, without the inclusion of electrostatic polarization, cannot accurately describe the solvation of highly charged biomolecules and such force fields are inherently unable to treat bond making/breaking reactions. Improvements will be made to these classical force fields, but a shift to quantum mechanical force fields (*vide infra*) will be required to achieve quantitatively accurate enzymatic simulations.

The primary advancement will be the merging of the QM and molecular dynamics methods to allow so-called First Principles MD, where quantum mechanical forces will be used to drive the classical motions of the atoms. Extension to dynamics simulations requires considerable methods development; the DFT force calculation must be converted to a linear-scaling method, and the entire molecular dynamics simulation must be implemented on a massively parallel computer.

Even with these improvements, first principles MD will not yet be feasible for long timescales and large molecular sizes such as that outlined in the previous chapter. However, this capability will allow the solving of a large number of fundamental biophysical problems that have been inconclusively addressed by existing classical MD methods. These problems include the determination of the hydration structure of the DNA nucleoside bases; the energetic factors leading to DNA base pairing; the hydration of the DNA backbone and basic sites; and the role of polarization in the stability of protein helices.

Models and Algorithms for In Silico Drug Design

Docking methods are computational algorithms developed to both predict the three-dimensional structures of ligand-receptor complexes, and to evaluate the relative affinity or free energy of binding for these bound ligands or drugs. The need for improved docking and scoring methods is now especially acute given the future direction toward high-throughput annotation of genomes to generate new protein structural targets, combined with revolutionary advances in combinatorial synthesis of small molecule docking candidates.

The current combinatorial library paradigm is to design diverse drug libraries aimed at multiple but unknown targets or directed ligand libraries aimed at optimizing hits against individual targets. An inverted procedure is possible in which one or many libraries are screened on the computer against many targets to determine which libraries have the most desirable characteristics for which targets. This general approach can also include an "optimization" cycle where augmented libraries are scored against the best targets.

The fundamental attractiveness of this approach is that potential targets for all compounds can be addressed at a much earlier time and at much lower cost per compound, and is consistent with genome-scale drug design efforts. The basic challenge is how to improve the accuracy of the fundamental docking algorithms themselves while rapidly screening increasingly growing drug databases both in house and in the public domain.

Docking methods for geometric optimization of a candidate drug into a target active site is a solved problem when both the ligand and the target are treated as rigid objects. In some cases, limited flexibility is introduced by dividing the ligand into several rigid fragments that are docked separately. In either case, these binding complexes are then evaluated with an empirical scoring function, which we discuss further below. This

represents the level of sophistication that is currently available from commercially available software packages. This approach is likely to identify, at most, one weakly binding compound per database of 100,000 chemicals, i.e. there are too many false negatives generated.

Introduction of full flexibility of at least the ligand for docking into a rigid target to refine the binding geometry has been shown to lead to better binding energetics, and therefore finding better drug leads in general. Flexible ligand and rigid target represents the upper limits of what can be attempted with current computational resources. When the peptide backbone and side chains of the target molecule are also treated as flexible, allowing the molecule to undergo locally induced conformational changes upon ligand binding, the resulting induced fit seems to be essential to understand ligand specificity. Large-scale screening with full flexibility of both ligand and localized areas of the target is well-beyond reach with current computational resources, since only a few compounds can be screened in a realistic time. Essentially increased use of geometric refinement with at least full ligand flexibility, and ideally at least localized target flexibility, via standard optimization techniques for large libraries of drug compounds is an accessible and desirable goal in using future 100 teraflop computing.

Once a drug is geometrically docked, scoring of the binding affinity of a drug-receptor complex ranges from statistical multivariate equations that correlate X-ray crystal structural data of ligand-receptor complexes with experimental free energies of binding, to physically-based molecular mechanics approaches, to computationally intensive free energy perturbation methods.

Overall the rapidly calculated multivariate functions perform as well as the computationally intensive free energy perturbation calculations, with estimated

relative binding free energies of about 2 kcal/mole, which corresponds to a binding affinity error ($\sim 10^{\text{Gerror}/1.4}$) of about 30-fold. The quality of these scoring functions provides a qualitative filter for ordering the binding affinity of drugs in large databases, but are not a reliable predictor of the most active drug molecules. Ranking drug affinity among many ligands for a given target is where multivariate functions work well, but it is unlikely that these existing scoring functions could determine specificity of a single drug against different receptors. It is even arguable that there is insufficient structural data for this case, so that multivariate statistical approaches are ultimately a dead-end.

Both molecular mechanics and free energy perturbation methods have the advantage of being based on physical interactions, so that alternative problems can be treated by the same approach. Molecular mechanics functions with solvent-accessible surface area descriptions for solvation have on average performed significantly worse than multivariate functions in the past, with binding

free energy errors typically being 3 kcal/mol/. However, quite good correlation of the enzyme inhibitor activity of 33 inhibitors of HIV-1 proteases were determined by a purely molecular mechanics scoring function, and recent reported results for some new empirical force fields perform as well as correlated MP2 QM methods.

The cost of evaluating a 100,000 compound library against 100,000 gene products would take on the order of a full year on a dedicated teraflop computer. Certain large pharmaceutical companies have in-house databases approaching 500,000 drugs, and revolutionary combinatorial synthesis approaches are going to expand these databases even further. Add on top of that additional modeling accuracy requirements to better screen drugs, i.e. better empirical force fields, longer refinement stages in the calculation, and greater target flexibility, the search for better drug candidates will utilize well 100 teraflop capabilities on a sustained basis, and are inherently scalable beyond this projected computing goal for 2003.

High End Computing Needs for In Silico Drug Design and Enzymology

The table below gives an estimate of the method and *current* computational requirements to complete a binding affinity calculation for a given drug library size. Going down the table for a given model complexity is a level of computational accuracy. The benefits of improved model accuracy must be

offset against the cost of evaluating a model over the ever increasing size of the drug compound library brought about by combinatorial synthesis, and further exacerbated by the high throughput efforts of the genome project and structural annotation of new protein targets.

Modeling complexity	Method	Size of library	Required computing time
Molecular Mechanics Rigid ligand/target	SPECITOPE	140,000	~1 hour
	LUDI	30,000	1-4 hours
	CLIX	30,000	33 hours
Molecular Mechanics Partially flexible ligand Rigid target	Hammerhead	80,000	3-4 days
	DOCK	17,000	3-4 days
	DOCK	53,000	14 days
Molecular Mechanics Fully flexible ligand, Rigid target	ICM	100,000	~1 year (extrapolated)
Molecular Mechanics Free energy perturbation	AMBER CHARMM	1	~several days
QM Active site and MM protein	Gaussian, Q-Chem	1	>several weeks

State-of-the-art quantum chemistry algorithms can expand the applicability of QM/MM methods to simulate a greatly expanded QM subsystem for enzymatic studies, or even estimation of drug binding affinities. An estimate of the cost of using QM methods for evaluating a single energy and force evaluation system of 10^4 heavy atoms, would require resources that can handle $\sim 10^{16}$ FLOPS; on a 100 teraflop machine this would require five minutes.

Parallel versions of these methods have been implemented and are available at a number of universities and DOE laboratories. On current generation teraflop platforms, QM

calculations of unprecedented size are now possible, allowing HF optimizations on systems with over 1000 atoms and MP2 energies on hundreds of atoms. The presently available traditional-scaling ($\sim N^3$) first principles molecular dynamics code is running efficiently on serial platforms, including high-end workstations and vector supercomputers. The ultimate goal is to develop linear scaling quantum molecular dynamics code. It will be important to adapt these codes to the parallel architectures, requiring rewriting parts of existing programs, and developing linear scaling algorithms.