Abstract

The objective of this proposal is to design mathematical methods for computational drug design and protein folding that directly address the multiscale nature of these problems. The approach combines ideas from conformational dynamics (which reduces the dimensionality of the system), multilevel methods (to create a multiscale description of the dynamical system), and computational geometry (to assist in the derivation of high level molecular descriptions).

These ideas may allow a more efficient dynamics simulation and sampling of configuration space in which the systems move. The dynamics are important because the biochemical functionality of some complex systems such as enzymes depends on dynamical processes like gating. Also, molecular dynamics simulations are a natural way to permit ligands (and receptors) to explore their configurations in space, which determines the efficacy of binding or docking (and is related to the promise of a compound to be used as a drug). Also, dynamics simulations can sample phase space to determine the folded states of proteins and pathways among them.

A limitation of conformational dynamics is the approximate nature of its solution, since the inert part of molecules is ignored, which limits its predictive power. Extensions to this method will be explored at several levels: at the finest level, symplectic integrators that allow long time steps will be developed. This is a continuation of my work on developing more stable symplectic integrators for Hamiltonian systems. Next, the development of linear time methods to compute the electrostatics force in the presence of periodic boundary conditions will be attempted using multigrid summation and wavelet transforms. At a higher level, conformational dynamics will be coupled with multiscale molecular dynamics, which will provide a more accurate description of the dynamical system, while taking advantage of the high-level picture of the conformational or essential dynamics. This will provide a way to handle the inert part of a molecule. The proposed solution will find an average contribution from the dynamics of all degrees of freedom, and will incorporate this contribution into the essential dynamics as a Langevin-type term. The main difference with Langevin dynamics is that this method will provide an automatic and optimal—in a well-defined sense—way of computing the Langevin coefficients.

The use of computational geometry to assist in the detection of essential variables, and in the production of other high-level descriptions of drug compound and proteins will be explored.

The support of this grant will allow an initial exploration of these ideas through literature surveys and the analysis and prototyping of simple models and solutions. It may also lead to the establishment of formal collaborations with several researchers in academia and industry.

Given the multidisciplinary nature of this research, the proposed research will support graduate students in science, engineering, and computer science. This Fall I will direct Mr. Marc Ma (M.S., Mech. Eng.), Mr. Hong Hu (M.S., Chemistry), Mrs. Alice Ko (B.S., Biology), and Mr. Virgil Andronache (B.S., Math and Computer Science).

Also, this research may be helpful in establishing an embedded center for High Performance Computing for Bio/molecular modeling at Notre Dame, which could also involve some of the local pharmaceutical and chemical industries.

Finally, this grant may enhance the possibilities of obtaining an NIH Bioengineering grant, or Whitaker Foundation Biomedical Engineering Research Grant, and an NSF Career Award, which I will submit this summer.
1 Description

The objective of this proposal is to design and implement improved mathematical methods for protein folding and computational drug design. In protein folding, one wants to predict the tertiary or globally folded structure of globular proteins from the primary structure (a sequence of aminoacids) or from the secondary structure (information about local folding). A fourth level of structure is that of aggregation of proteins or quaternary structure. We will not consider the latter in this proposal. This hierarchy is illustrated in Fig. 1. Folded states of proteins, along with the dynamics that determine access to their binding sites, and the chemically active groups on their surfaces, determine their biochemically interesting function. Since sequencing is occurring at a much faster rate than the experimental resolution of the structure of proteins, partly because it is difficult to crystallize proteins to determine crystal structure, there is interest in predicting folding by computer. This would be helpful in interpreting the data from the human genome project, understanding the mechanism of some diseases, designing drugs with specific therapeutical properties, and growing polymers with specific material properties (cf. survey paper by [Neu97]). A related application is the rational design of drugs. Here, a potential drug is intended to interfere with normal or aberrant biochemical processes. There are usually two steps: find a lead compound, which then may be modified to improve its properties. In both steps there is need from help of computer simulations as will be explained below [CP00, BCM93].

Computational methods in these areas suffer two fundamental shortcomings: on the one hand, the predictive value of the simulations is not sufficiently high for computational drug design or determination of protein function or structure, and on the other hand, the applicability of the methods, particularly for protein folding, is limited due to the computational expense. The presence of multiple spatial and time scales in the systems simulated is responsible to a great extent for both problems: it is necessary to resolve the models at very fine scales in space and time, and this limits the size of systems that can be simulated, and more importantly the length of simulations to observe processes of interest. For example, folding of proteins occurs in the order of microseconds whereas fast vibrations occur in the order of femtoseconds. Also, the accuracy of the models is lowered to make computational solutions feasible. We will try to address both problems in what follows.

Let us state some theoretical assumptions about protein folding. Folded structures correspond to an energy minimum under physiological conditions. If nature followed a brute search we would confront the Levinthal paradox (1968): a chain of about 100 amino acids such as ribonuclease with 10^{100} possible conformations for each one, would need to explore 10^{50} configurations! Even if the molecule could explore one conformation every 10^{-13} second, it would take 10^{30} years to test a significant number of them. Instead, it is believed that proteins fold along a funnel, as illustrated in Fig. 2. Although there is disagreement in the details of the theory, folding seems to involve well-defined intermediate states, which correspond to local minima. However, the folding process can get delayed by trapping of molecules in “off-path” states. This is illustrated in Fig. 3. Indeed, testing different theories of how folding pro-
Figure 2: Energy surface to visualize protein conformations. The energy landscape tends to funnel the protein toward the native state (a minimum). Some paths lead "downhill" to the minimum (like A) whereas others (like B) may lead to secondary energy minima. From K. Dill and H. S. Chan, Nature Struct. Biol. (1997) 4:10-19.

ceeds is in itself one of the attractions of modeling protein folding.

2 Methods

The approach advocated in this proposal combines ideas from conformational dynamics (where the objective is to reduce the dimensionality of the problem), fast computation of electrostatic forces, multiscale molecular dynamics (which attempts the efficient evaluation of all scales, by providing well-defined ways of interacting among multiple scales), and computational geometry (to discover invariants in the conformational dynamics and derive 3D high-level descriptions of compounds and proteins).

2.1 Conformational dynamics

Conformational dynamics has arisen as an answer to the dimensionality problem in molecular dynamics. Its key insight is that the dynamical behavior of mechanical systems moving on multi-minima energy surfaces occurs along so-called essential dynamic variables [Ber98, HM98]. Thus, if one is able to identify these essential variables, it is possible to directly compute metastable conformations in the reduced (or essential) space, and the probability of transitions among them. A conformation is a set of configurations in space that are close to an equilibrium point. A metastable conformation corresponds to a local minimum that is close to the global minimum, and in which a protein may remain for a long time.

A Hybrid Monte Carlo (HMC) algorithm has been used to compute this conformational information [Sch99]. HMC does a better sampling of the conformational space than either pure Monte Carlo or molecular dynamics. It uses a molecular dynamics simulation to propose the next Monte Carlo move. The new configuration is then evaluated according to the Metropolis algorithm. This is better than traditional Monte Carlo, which typically allows only random changes at one particular location, and better than pure molecular dynamics since it allows longer time steps and does not require conservation of energy. A sufficient condition to satisfy the property of detailed balance in HMC is that the integrator used be symplectic, which implies that it preserves volume in phase space. Research into improving the HMC approach of Drs. Schütte and Deuflhard at the Konrad-Zuse-Zentrum für Informationstechnik in Berlin [DDJS98] will extend the results of my previous work in symplectic multiple time stepping integrators. I have increased the time steps possible using several mollified impulse multiple time stepping methods [IRS99, Iza99, ICWS00, AI00, Iza00]. Further development of these integrators will allow for even longer time steps to be taken for the simulations, which will be particularly important for biological systems. Symplectic integrators are an excellent match for Hamiltonian systems for several reasons: (a) symplectic integrators produce the exact solution to a slightly perturbed Hamiltonian; (b)
symplectic methods are generally more stable than their non-symplectic counterparts; and c) they allow integration for longer periods of time than non-symplectic methods. Good introductions to symplectic integration are the book by [SSC94], and more specifically to symplectic integration in MD, the paper by [Ske99].

### 2.2 Fast electrostatic methods for periodic boundary conditions

A related area of research will be the study of fast methods for the computation of electrostatic forces with periodic boundary conditions. This is very important in biological systems, since long range electrostatics play a role in the interactions of proteins and DNA, and in highly charged systems [KBS97, HD78]. It has been shown that the use of cutoffs in these computations is a source of severe inaccuracies [GHWS91, Sai94, YYL+95]. For a discussion on periodic boundary conditions the reader is referred to the book [AT87, pp. 24–32, 156ff.]. The direct computation of electrostatic forces scales as $O(N^2)$. Fast methods such as the fast multipole method, which is $O(N)$, and the Particle Mesh Ewald (PME), which is $O(N \log N)$, are difficult to implement or to parallelize. I intend to generalize a multilevel fast summation of Coulomb electrostatic forces to handle periodic boundary conditions. This method was originally proposed by A. Brandt in [Bra95], and has been implemented by R. D. Skeel at UIUC [Ske00]. We will collaborate with the latter in extending the method. Another possibility that will be explored is to use a wavelet expansion, instead of the traditional Fourier expansion to compute an Ewald sum on a grid. Both methods could in theory be made $O(N)$. The multilevel fast summation is easier to implement than the fast multipole method, whereas a fast method based on a wavelet expansion with compact support should be easier to parallelize than PME. This work may involve the collaboration of Dr. Godehard Sutmann, who will be visiting me at Notre Dame in August 2000 (wavelet-based electrostatics, Research Centre Juelich, Germany).

### 2.3 Multiscale dynamics

Another improvement to conformational dynamics will be to model the interaction among the essential dynamics and the rest of the system (the inert part of a biomolecule), in order to obtain a more accurate dynamical description of the system (thus improving the predictive power of computer simulations). This will be done by constructing a multilevel description of the dynamics. Multilevel methods (such as multigrid PDE solvers) have allowed the efficient solution of equations in a great number of applications (cf. [Bra95]). Thus, for example, in the finest level of description, the dynamics of the whole system will be considered. From these dynamics, a Langevin type coefficient will be incorporated in the next coarse level, and so on. The coarsest level is that of the essential dynamics that may be resolved by HMC. The correction from the finer levels will be incorporated in the MD component of the essential dynamics. The whole approach can be seen as a generalized form of Langevin dynamics. However, the amount of Langevin damping will not be decided arbitrarily; it will rather be computed from simulations at finer levels. The general ideas will be similar to those of Dov Bai at the Mathematical Sciences Institute at Cornell and Achi Brandt at the Weizmann Institute [BB00], but the approach will be different. Whereas they use Monte Carlo in the finest levels, I will use dynamics at most levels. We will collaborate on comparing the results of our differing approaches.

### 3 Applications

Multiscale conformational dynamics will be applied in several problems:

1. **Protein folding and dynamics**, so as to find metastable conformations (using multiscale HMC) or to study the actual dynamics (using multiscale MD only) [ZWM98]. For example, Duan and Kollman at the University of California in San Francisco simulated a Villin protein for a microsecond. They observed helix formation and compactness followed by conformational adjustments, and were able to recognize metastable states and pathways. This simulation is the longest of a protein to this day [DK98]. A particularly interesting system is the **Shaker** potassium ion channel, an important membrane channel whose ion conductivity through the channel is turned on or off by changes in membrane potential (voltage). Such voltage-controlled ion channels are found in neurons, muscles, and other cells, and are involved in nerve impulses. The structure
and conformational changes of these channels in multi-cellular organisms (like humans) are not well understood. I would like to explore the use of multiscale conformational dynamics to model results from experiments done using fluorescence resonance energy transfer by Prof. Paul Selvin, department of Physics, University of Illinois at Urbana-Champaign [CSSB99]. A starting point to model such systems may be Dr. Daniel Gezelter’s attempts to model diffusion through membranes using a simplified potential that incorporates dipole moments. Efficient numerical evaluation of this potential would make the model even more attractive and useful in this application (cf. [GRB99]).

2. Molecular docking for drug design. Molecular docking is a tool to identify binding configurations of ligands (e.g., a small drug) and a receptor (e.g., a binding site in a protein), test modifications of lead compounds for drug design, and to explore systems that are difficult to model experimentally (such as enzymes). Most docking algorithms model the receptor as a rigid body. However, an ideal docking algorithm should include flexible ligand and receptor, allowing both to explore their conformational spaces [Lea96]. Multiscale conformational dynamics may make this possible.

3. Virtual drug screening. I plan to collaborate with Prof. Danny Chen in exploring uses of his work in computational geometry in two applications: First, in the derivation of 3D pharmacophores, that is, sets of features that are common to a series of active molecules. Databases of 3D pharmacophores are typically used to search for possible lead compounds for drugs (a process called virtual screening) [BM96, SG90]. Second, it will be used to detect geometric similarity among different configurations of biological systems so as to determine essential variables for essential dynamics and coarsening in multiscale MD. These applications may benefit from collaboration with Dr. Peter Deuffhard, president of the Konrad-Zuse-Zentrum in Berlin, and Dr. Michael Chaney, from TransTech Pharma in North Carolina.

4. Methodology

- Model refinement: The model for multiscale conformational dynamics will be refined in a manner similar to [BB00]. The iterative procedure involves the following:
  1. Build coarser models of the protein.
  2. Test that the finer model probability distributions converge using only the coarser variables.
  3. Iteratively correct the coarse model using the above measure.
  4. Add correlation terms. Repeat procedure.

- Development: the methods mentioned above will be prototyped using the molecular dynamics program SAMD 2, an object oriented program developed in our group for easy prototyping. It has scripting capabilities and will be parallelized soon [IWSV00, WI].

- Analysis
  1. Stability analysis of the symplectic integrators developed. Given the nonlinear nature of molecular dynamics, this has to be done partly with empirical evaluation and linear analysis extrapolations.
  2. Statistical testing of the mildly stochastic methods like HMC.

References


[BB00] D. Bai and A. Brandt. Multiscale algorithms in molecular dynamics. In A. Brandt, K. Binder, and J. Bernholc, editors, Multiscale Computational


5 Discussion of seed grant support

As mentioned above, the seed grant support would enhance the prospects of the research effort in several levels: materially, because it would allow for partial support of a graduate student to explore these ideas during the Fall semester of 2000, and it would support me during the summer to explore a few of the above ideas. Formally, it would also enhance the possibility of funding, particularly because this will be my first proposal where I am the principal investigator. Finally, the intellectual collaboration with other members of the Center for Applied Math, which would thus be strengthened and formalized, would most definitely be helpful to me in carrying out this research effort.

This may also lead to formalized collaborations with some of the following researchers: Dr. Atul Bahel (drug design, postdoctoral research associate in my group, CSE, Notre Dame); Dr. Danny Chen (computational geometry, CSE, Notre Dame); Dr. Dan Gezelter (membrane models, Chemistry, Notre Dame); Dr. Peter Deuflhard (conformational dynamics, President, Konrad-Zuse-Zentrum, Berlin, and Free University of Berlin); Dr. Achi Brandt (multiscale molecular dynamics, Weizmann Institute of Science, Israel); Dr. Paul Selvin (ion-channel experimental structure and dynamics determination, Physics, Dr. Robert D. Skeel (fast electrostatics, Beckman Institute, University of Illinois at Urbana-Champaign); and Dr. Godehard Sutmann (wavelet-based electrostatics, Research Centre Juelich, Germany). I have also a potential collaborator in industry (drug design, Dr. Michael Chaney, TransTech Pharma, North Carolina).