PROJECT DESCRIPTION

1 Introduction

Flexibility of several protein receptors in the docking process is important for several biologically relevant systems, such as enzymes binding to substrates or inhibitors. The use of multiple protein structures (MPS) has been recognized as a useful approach to handle the flexibility of the receptor. Existing methods typically accommodate effectively side chain motions by creating an averaged representation of the protein. This proposal seeks to extend these methods by allowing more extensive motions, such as loop and domain motions. Instead of creating an averaged representation of the protein, our approach will select an ensemble of conformations using conformational sampling and analysis tools. These will be systematically searched against ligands. This approach promises to be more effective to describe docking when the motion upon binding is more extensive than simple side chain displacement.

Another question addressed by this proposal is the question of the selection of conformations for MPS docking protocols? How can one generate a set of conformations that sufficiently covers and represents the relevant conformational space of the molecule, and which is kept to a minimum size in order to keep the computational cost tractable? Some of these questions have been addressed by conformational analysis and sampling, but they have not been used extensively in the study of MPS docking protocols. We will use, extend, and develop techniques from conformational sampling and analysis to extend the coverage and study the optimal selection of a set of protein structures for docking.

Algorithmically, we will include data from NMR relaxation experiments (some of which will be performed in house) and simulations to determine the flexibility of different regions of the protein, and incorporate them into a conformational sampling protocol that we have developed, the Shadow Hybrid Monte Carlo, which combines efficiently molecular dynamics and Monte Carlo moves. These moves will include dihedral motions, as well as more sophisticated concerted motions, such as hinge motions.

The conformational sampling, selection of MPS structures, and docking protocols will be evaluated in three important model systems: HIV protease, HIV integrase, and DHFR. We will validate putative protein-ligand interactions experimentally using NMR techniques.

Because conformational sampling and docking generate massive amounts of information, and it is helpful to keep a lot of that information available for other researchers, statistical analysis, and reproducibility of results, we will develop a grid-based set of tools and databases for the simulation data generated.

The results of these efforts will help answer the questions of how to generate conformations that cover more conformational space using NMR and dynamical analysis; how to select optimally a subset of conformations for MPS flexible docking; and an evaluation of the relative merits of different approaches for performing the conformational analysis and flexible docking. Finally, these results will be disseminated through the use of the grid-enabled toolset, including the distributed database of molecular simulations. An important concern addressed explicitly in this proposal is the management of reliability and availability in the grid for simulation results: this facility is not available in current grid tools to the best of our knowledge, and may be a proof of concept for more ambitious biological databases in the future.

2 Background and Significance

2.1 Flexible protein docking

Incorporation of complete flexibility of protein structure is a major hurdle in protein-ligand docking. Modeling a fully flexible receptor involves huge computational complexity, which makes it impractical to implement. For this reason, many methods have been developed with simplified molecular description and limited motion of the receptor. Examples include, for instance, methods that incorporate side chain flexibility (Leach 1994, Jones et al. 1995), use of rotamer library (Schaffer and Verkhivker 1998) and harmonic modes (Zacharias and Sklenar 1999) or hinged domains (Sandak et al. 1998) to incorporate large motion of the receptor molecules in docking. An example of one of the first docking software is DOCK\(^1\), which considers receptors

\(^1\)http://dock.compbio.ucsf.edu
comprised of a single, rigid structure. DOCK identifies the pockets in the active site of the receptor and generates spheres. The ligand is then fitted into these spheres using least-square fitting. DOCK is suitable for small molecule docking, where prior knowledge of binding sites is available.

Recently, with the availability of fast and cheap hardware, different techniques have been developed to incorporate multiple protein structures (MPS) in protein-ligand docking as well as large motions in protein molecules. These techniques can be broadly divided into following categories: (i) Superimposed protein structure; (ii) average grid representation; (iii) pharmacophore modeling; (iv) systematic search; and (v) large domain motions.

(i) **Superimposed protein structure.** An ensemble of protein conformations are overlaid to generate a superimposed structure. Identical parts of backbone and side chains of the receptor are merged, whereas the dissimilar regions are retained to represent flexibility much like a rotamer library. A popular docking software, FlexE\(^2\), has been implemented with superimposed protein structure (Clauben et al. 2001).

(ii) **Average grid representation.** Another way of using MPS in docking is average grid representation. Multiple structures can be combined to generate a composite grid of interaction energies. The composite grid can be obtained from the grid representation of the overlaid protein structures by different techniques. A mean grid takes average at each point while a minimum grid takes the least value at each point. A better way to calculate the composite grid is to take a weighted average at each point across all the structures. The average grid contains the information about the shape and specificity of the receptor site and helps to simulate docking with feasible computational effort (Broughton 2000). Autodock\(^3\) is a popular docking software based on grid representation of receptors (cf. Osterberg et al. 2002). It uses MC simulated annealing technique for interaction configuration exploration and a rapid energy evaluation technique as a scoring tool based on grid based molecular affinity potential.

(iii) **Pharmacophore modeling.** Pharmacophore modeling is another method of averaging multiple target structures (Masukawa et al. 2000). This method is similar to the average grid representation. In this method, complementary regions of active site are identified using small probe molecules and then these structures are overlaid to develop a dynamic pharmacophore model.

(iv) **Systematic search** Averaging protein conformations makes docking computationally efficient with the price of decreased accuracy. Systematic search (docking each protein conformation with each ligand) can better incorporate flexibility of the receptor structure in protein ligand docking. Bouzida et al. (1999) first used this systematic search technique. They docked eleven different conformations of HIV-1 integrase with two different ligands. They found that one of the ligands bound preferentially to one of the conformations, whereas the other bound equally to four of the protein conformations. Their results clearly indicate the necessity of MPS in docking.

(v) **Large domain motions** Another aspect of protein flexibility is attributed to large domain motions. Different techniques have been developed so far to predict large domain motion in protein molecules through hinge-bending (Maiorov and Abagyan 1997, Wriggers and Schulten 1997, Kellet et al. 2000). Using harmonic modes, mobility of large regions of the protein molecules can be modeled using vibrational frequencies. While hinge bending and harmonic modes has been incorporated separately into docking for modeling large domain motions, it is essential to incorporate MPS along with domain motions and loop fluctuations to catch full flexibility of protein molecules in protein-ligand docking (Carlson and Mccammon 1999).

Current paradigm describes proteins in a pre-existing ensemble of conformational states (Luque and Freire 2000, Ma et al. 2002). The protein conformations can be obtained either from simulation (conformational sampling) or from experiments (NMR). This suggests that a systematic search over MPS is more accurate than the averaging methods. But it is still computationally expensive to incorporate all these conformations into docking. For this reason, we are going to choose a “representative set”. The idea is to select a set

\(^2\)http://www.biosolveit.de/software
\(^3\)http://www.scripps.edu/pub/olson-web/doc/autodock/
of conformations which represents the complete space of conformations, sufficiently, where the set is small enough so that docking simulations produce results with feasible computational effort. The algorithms to find the “representative set” are described later, and follow in the category of conformational analysis and sampling.

2.2 Grid Enabled Computation and Databases

The emerging field of grid computing has been applied to a multitude of computationally complex scientific challenges. Mature grid projects related to molecular simulation have focused on computation. The Folding@home (Vijay Pande PI, Stanford) project and similar server-client endeavors have taken biomolecular grid computing into households around the world employing over 1,000,000 CPUs. Grid interfaced computations have permitted the modeling of protein folding over time lengths thousands of times larger than previously achieved. However, their computational focus has not addressed the dynamic utilization of molecular simulation and conformational data between geographically and semantically (algorithms, data representations, and architectures) separated research groups.

Our proposed research focuses on managing the integration of molecular simulation and experimental data as well as runtime configuration parameters; to enable conformational analysis and flexible docking on the grid. Figure 1 demonstrates the innovative workflow which will enable the rapid evolution of docking simulations. The Sampling and Conformational analysis tools will enable the intelligent selection of experimental and simulation data. The novel transient and meta-data utilization tools will transform the database from a simple data depository into a dynamic repository from which new simulations can be automatically generated and complementary simulations can be compared. A biochemist performing docking simulations will, for the first time, be able to parameters in one synergized abstraction that delivers computational and storage capabilities beyond that of their local cluster.

![Figure 1: Computational Tool Overview](image)

Complementary projects by researchers in the field are currently in their infancy. BioSimGrid (Mark Sansom PI, University of Oxford) seeks to provide the British research community with grid distributed access to biological simulation data. Their research team has appropriately identified and sought to address differential platform and data format challenges; however, we consider the development of standard formats, filters, and interfaces a secondary benefit to our primary goals: to use transient data to efficiently extend existing simulations, prove the capability to use meta-data to manipulate existing simulations for exploration via parameter variation, and invent grid distributed data primitives that ensure guaranteed levels of reliability. It is worth noting that there are also a few projects such as MolecularMechanicsOpenStandards (MMOS, 2004); which are solely focused on the standardization of molecular simulation configurations using XML to create a Molecular Mechanics Markup Language. The SimDB (Lennart Johnson PI, University of Houston), (Abdullah, 2002) project also identifies the challenges presented by the broad range of highly
differentiated data and computation configurations. In contrast, our research proposal is highly targeted to provide both a proof of concept for novel data sharing utilities, and an incremental release of open source implementations designed to work with NSF funded open source NMI middleware.

3 Research Plan

3.1 Flexible docking Subprojects

These subprojects attempt to characterize the effectiveness and efficiency of different docking protocols that use MPS for modeling flexible docking. Conformational sampling and analysis will be used to generate optimal sets of MPS for the flexible docking protocols. There has been extensive work in both conformational analysis and flexible docking. We expect the combination of these two areas of study will be synergistic: by better exploring conformational space and by selecting representative conformations for MPS docking, the efficiency and accuracy of flexible docking can be enhanced, irrespective of the protocol used. We will evaluate the most common flexible docking protocols. We think that a systematic search over an optimally chosen set of MPS will give an accuracy enhancement over other protocols. On the other hand, the introduction of docked structures may aid conformational sampling, by biasing the process towards biologically relevant conformations. It is our hope that these sampling, analysis, and docking protocols can also be applied in the future to evaluate putative protein-protein interactions where the structure is not known, which is beyond the capabilities of current docking protocols. Tools that perform these tasks will be made available to the research community through our GEMS toolset, as an added benefit of this research. Currently, our sampling methods described here are available in an open source framework called PROTOmol, (Matthey et. al. 2004).

3.1.1 Conformational Sampling

Conformational sampling will be performed from a known X-ray or NMR structure. Two types of simulations will be done: unbound molecules and docked molecules. The latter will be obtained from either an experimental complex (preferably) or from a simulated docked structure. Sampling will be performed using our Shadow Hybrid Monte Carlo protocol, with suitable modifications to include non-continuous motions of essential torsion angles determined by simulation or through NMR experiments.

Graph theoretic algorithms and analysis of dynamics and thermodynamics will be used to characterize flexible and rigid residues. NMR relaxation data, when available, will be used to identify flexible residues in a binding site. Fast (picosecond to nanosecond timescale motions) will enable the selection of essential or most flexible dihedrals in the backbone and side chains. In particular, this data might enable the recognition of dynamics that are affected upon ligand binding (cf. Ishima and Torchia, 2000). Those rotatable bonds whose dynamics are flexible according to the experimental data, will become possible moves in the MC step of our sampling protocol.

The rate of conformational changes on the microsecond to the millisecond scale can give clues into the function of different enzymes. This information may come from NMR experiments – which could be corroborated by looking at the structure. Unlike the fast timescale data from NMR experiments, this information is less specific, and points more to neighborhoods or the milieu in which slow motion is occurring. Thus, more simulations and experiments will be needed to determine the cause of such motion. When two or more structures are available, hinge motion detection algorithms or more general transition path sampling will be used to generate further conformations.

3.1.2 Conformational Analysis

Conformational analysis will be performed on sets of conformations to select subsets that are most representative according to different criteria, which are explained below. These algorithms will be used to select MPS that are input to the flexible docking protocols, and also to select which conformations to store with higher reliability in the grid. Different techniques will be evaluated in this proposal.
3.1.3 Evaluation of Flexible Docking Protocols

Some of the most important flexible docking protocols will be evaluated in representative data sets. The accuracy of predictions, measured using different metrics, and the efficiency of the protocol will be determined through simulations. For an important subset of these results, experimental validation using NMR techniques will be performed.

3.1.4 Grid Portal for flexible protein docking

In concert with our algorithmic developments, we will develop a grid portal to offer an intuitive interface to for the GEMS toolset. The grid portal will leverage the rich base of existing work by the NSF NMI (National Middleware Initiative) to create a portlet inside of the Open Grid Collaborative Environment (OGCE) 4. The grid portal will be constructed to provide individual portlets encompassing the aspects of conformational sampling, conformational analysis, docking, and data storage.

3.1.5 Distributed Database on the GRID

In order to provide an intuitive interface to the end user, an implicit requirement for the portal will be to manage the massive amounts of data produced by the grid. While the grid offers the opportunity to access resources on an unheralded scale, the sheer magnitude of data produced can be overwhelming without proper organization. To that end, we propose to develop novel techniques capturing the reliability and availability of data as well as the relational nature of transient data produced by the simulations. By taking into account the special relations of the data and incorporating additional meta-data, we will be able to offer a large scale, reliable database of molecular simulation conformations and trajectories via the portal.

3.2 Preliminary Results

3.2.1 Conformational Sampling Methods

Shadow Hybrid Monte Carlo (SHMC) is a new method developed by the PI's group for sampling the phase space of large molecules, particularly biological molecules. It improves sampling of Hybrid Monte Carlo (HMC) by allowing larger time steps and system sizes in the molecular dynamics (MD) step. The acceptance rate of HMC decreases exponentially with increasing system size $N$ or time step $\delta t$. This is due to discretization errors introduced by the numerical integrator. SHMC achieves an asymptotic $O(N^{1/4})$ speedup over HMC by sampling from all of phase space using high order approximations to a shadow or modified Hamiltonian exactly integrated by a symplectic MD integrator. SHMC satisfies microscopic reversibility and is a rigorous sampling method. SHMC requires extra storage, modest computational overhead, and a reweighting step to obtain averages from the canonical ensemble. This was validated by numerical experiments that compute observables for different molecules, ranging from a small $n$-alkane butane with 4 united atoms to a larger solvated protein with 14,281 atoms. In these experiments, SHMC achieves an order magnitude speedup over HMC in sampling efficiency for medium sized proteins.

Let $\tilde{\rho}(\mathbf{x}, \mathbf{p})$ be the target density of SHMC, where

$$\tilde{\rho}(\mathbf{x}, \mathbf{p}) \propto \exp \left( -\beta \tilde{H}(\mathbf{x}, \mathbf{p}) \right),$$

$$\tilde{H}(\mathbf{x}, \mathbf{p}) = \max \left\{ \mathcal{H}(\mathbf{x}, \mathbf{p}), \mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p}) - c \right\}. \quad (2)$$

Here, $\mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p})$ is the much smoother shadow Hamiltonian, defined in [Skeel and Hardy 2001], and $c$ is an arbitrary constant that limits the amount by which $\mathcal{H}_{[2k]}$ is allowed to depart from $\mathcal{H}(\mathbf{x}, \mathbf{p})$.

Algorithm 1 lists the steps for calculating SHMC. The first step is to generate a set of momenta, $\mathbf{p'}$, often chosen via a Gaussian distribution. $\mathbf{p'}$ is accepted based on a Metropolis criterion step proportional to the difference of the total and shadow energies. This step is repeated until a set of momenta are accepted. Next, the system is integrated using MD and accepted with probability proportional to Eq. (1). Finally, in order to calculate unbiased values, the observables are reweighted.

---

4 http://www.ogce.org
Algorithm 1 Shadow Hybrid Monte Carlo (SHMC)

1. **MC Step**: Given $X = x$, generate $P$ with p.d.f. $\tilde{\rho}(X, p)$, using the acceptance-rejection method:
   
   (a) Generate $P$ having p.d.f. $\rho_P(p)$
   
   (b) Accept with probability $\min\left\{1, \frac{\exp(-\beta(\mathcal{H}(X, P) - c))}{\exp(-\beta\mathcal{H}(X, P))}\right\}$
   
   (c) Repeat (1a) - (1b) until $P$ is accepted.

2. **MD Step**: Given $\Gamma = \begin{pmatrix} X \\ P \end{pmatrix}$,
   
   (a) $\Gamma' = R\Psi(\Gamma)$ (where $\Psi$ nearly conserves $\mathcal{H}_{[2k]}$)
   
   (b) Accept $\Gamma'$ with probability $\min\left\{1, \frac{\tilde{\rho}(\Gamma')}{\tilde{\rho}(\Gamma)}\right\}$
   
   (c) If rejected, choose $\Gamma$

**Reweighting Post-processing**: Given sequence of $\{A, \Gamma\}$, reweight observable $A$ using $\rho(\Gamma)/\tilde{\rho}(\Gamma)$ before computing averages.

The purpose of the constant $c$ is to minimize the difference in the energies so that the reweighted observables of $\mathcal{H}_{[2k]}$ are unbiased. Currently, $c$ is chosen proportional to the expected value of the discretization error, $\langle \Delta\mathcal{H} \rangle$. This value is obtained after running a sufficient number of steps and monitoring $\Delta\mathcal{H}$ at each step.

**Testing**  SHMC was tested on molecular systems ranging in size from the 4 united atom alkane $n$-butane to the more complex solvated protein BPTI, with 14,281 atoms. Many different tests were performed, see Izaguirre and Hampton, (2004), with sampling efficiency being the most pertinent. Let the sampling efficiency of SHMC be defined as the computational cost per newly discovered conformation. The number of molecular conformations is determined using a method suggested by (Kirchoff et. al. 1996). This method works by classifying key dihedral angles of the molecule and watching how they evolve over the simulation. In a preprocessing step, the maxima of a dihedral angle are determined and used to create wells which can be identified by a simple character. At each simulation step, the current angle of a dihedral is used to classify which well it occupies and the corresponding character is recorded. This generates a string of characters which can then be compared.

**Results**  Results of sampling efficiency of SHMC versus HMC were recorded in Izaguirre and Hampton (2004). Figure 2 shows the values for a solvated protein, BPTI, with 14,281 atoms. The total time per new discovered conformation is plotted against several different timesteps. From the figure, we can see that SHMC is generating new conformations a full order of magnitude quicker than HMC. Since the speedup of SHMC to HMC is proportional to the size of the system, we would expect even better results for larger systems.

A preliminary set of tests comparing SHMC and MD (using Leapfrog) was conducted to further gauge the sampling ability of SHMC. The tests were run on a solvated protein, BPTI, with 1,101 atoms. Simulations were run with full electrostatics using particle mesh Ewald (PME) and a timestep of 1 fs. Samples were taken every 100 fs. A method to calculate the distance between conformations from Leach (p. 492, 2002) is the metric used for comparison. The mean distance and standard deviation are recorded in Table 1. The values are statistically similar, but SHMC shows more variation. We hypothesize that this is due to SHMC sampling from a wider array of conformations. It is well known that MD methods tend to sample in local basins, whereas stochastic methods are more likely to jump from one basin to another. The final column of Table 1 has the calculated efficiency for each method. The Leapfrog algorithm produced conformations slightly faster than SHMC. However, SHMC has the advantage that it is an exact method and does not suffer from the bias of MD. In addition, other MC type moves could be used with SHMC such as directly...
perturbing dihedral angles. There is enough evidence from these tests for us to consider comparing these two methods in greater detail.

<table>
<thead>
<tr>
<th></th>
<th>Mean (Rad)</th>
<th>Standard Deviation</th>
<th>Efficiency (sec/conformation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leapfrog</td>
<td>374</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>SHMC</td>
<td>340</td>
<td>38</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 1: Comparison of SHMC and Leapfrog for a solvated BPTI protein with 1101 atoms.

3.2.2 Grid Portal - GIPSE Toolset

In (Striegel et al. 2004), we proposed the concept of a generic framework for parameter-driven simulations that links the entire data production process (task creation, task submission, data parsing, and data storage) in a result-centric fashion. The Grid Interface for Parameter-driven Simulation Environments (GIPSE) toolset offers an alternative interface to the grid that is specifically tailored to simulation-based research. The motivation for GIPSE arose from the problem that although the grid provides the tools necessary for locating and utilizing resources, the grid is targeted towards managing tasks rather than the overarching result. While recent work tools (G von Laszewski et al. and D. Gannon et al. 2003) have reduced the complexity associated with interfacing with the grid, the interface presented to the user is still typically task-oriented. For parameter-driven simulation environments with large parameter spaces (such as with flexible protein docking), the development gap between the existing tools and the actual results can be quite significant. In short, GIPSE sits on top of existing grid tools (NSF NMI toolset) and employs XML metadata to link the service-centric nature of the grid with the data-oriented nature of the researcher.

The experience of the PI (Izaguirre) and co-PI (Striegel) on the GIPSE project will offer valuable experience for building GEMS on the NSF NMI framework. Furthermore, GEMS moves significantly beyond GIPSE in terms of storage by virtue of its fault tolerant virtualized storage database for the Molecular Simulation Database.

3.3 Proposed Work and Timeline

3.3.1 Sampling Methods

The use of MPS is a way of including both flexibility (and experimental uncertainty) into the modeling of the receptor in protein-ligand docking. We propose to improve in a systematic way the coverage and selection of
the conformational space spanned by the MPS ensemble for docking. This is an iterative process involving two stages. The first stage attempts to explore conformational space for docking, and consists of three steps: (1) the increase of coverage of conformational space, by generation of new conformations from an initial set of bound and unbound conformations, which in many cases may be just one experimental structure; (2) the selection of a set of representative conformations; (3) the docking of this initial set of MPS to ligands of interest. The second stage refines the docked structures from the previous stage by repeating the same steps, but in a more local way. Thus, different methods are used in each stage for the same conceptual steps. New algorithms are proposed for several of these steps, and all of them will be evaluated and their results made available to the research community using the grid GEMS infrastructure proposed here.

3.3.2 Coverage of conformational space

The objective of this subproject is to generate new conformations to increase coverage of conformational space, from a given set of conformations. We describe first the important case in which only one conformation is available initially, and then extensions when more than one conformation is available.

The proposed work will enhance the efficiency of conformational sampling through: (i) inclusion of information from conformational analysis obtained from either simulation or NMR relaxation data regarding the essential dynamics, as explained below; (ii) multiscale extension of the sampling protocol; and (iii) use of techniques to enhance sampling of MC procedures, such as the use of control variates.

**Essential dynamics into SHMC.** Several techniques will be evaluated to determine the degree of flexibility of torsion angles. This classification will be used to build a list of possible dihedral moves for the MC step of SHMC, considering their frequency of occurrence whenever this information is available. The objective is to enhance the occurrence of rare dihedral motions that may be inferred from NMR relaxation data or low mode analysis of slow motions.

The degree of flexibility in the protein will be evaluated by the following methods: (i) the method of Luque and Freire (2000), that produces a network of flexible and rigid residues; (ii) the graph theoretic method of Jacobs et al. (2001), which uses the connectivity in the system to identify structurally underdetermined regions; and (iii) principal component analysis.

If NMR relaxation data is available, the degree of flexibility of dihedral angles for the fast motions (ps to ns) can be determined by measuring the dihedral angle order parameter $S_{\text{NMR}}$, (cf Philippopoulos and Lim 1999). In the case of slow motions, the data points to a neighborhood or milieu of movements, that may involve one or more residues. Thus, more than one dihedral motion will have to be generated, and even so there is no guarantee that one will be able to account for slow motion, since some of these have concerted motions of more than one side chain.

When more than one conformation is available, we will explore the use of Chandler’s transition path sampling (TSP). Since we are interested in coverage of conformational space rather than in the correct dynamics, we can use a simpler and cheaper version of this method. In particular, TSP requires the generation of initial trajectories between two conformations of interest, the identification of regions of importance in the transition path sampling, and the computation of free energies for conformations. We will perform the first two steps of the protocol to identify a set of intermediate conformations. However, we will not need to perform the free energy estimation, which takes more than half of the computing time spent in TSP in biomolecules (cf. Radhakrishnan and Schlick 2004, and references therein). We will also use targeted SHMC instead of targeted MD in order to generate the initial trajectories. Targeted SHMC will use a targeted MD step instead of pure MD, with restraints given by NMR, the X-ray structure, or more generally the target conformation.

In the case of multiple initial conformations, we will apply the hinge method of Sandak referenced above to generate other conformations.

Finally, when docked structures are available, either during exploration or refinement of conformational space, the simulations will be performed with the docked structures, in order to sample the local basins and the effect of ligand binding upon the conformations of the protein. The MD portion of the simulations will include explicit solvation using fast multigrid electrostatic methods, which scale linearly and have been shown to be faster than PME for relatively small system sizes of 6000 or more atoms (Ko 2002, Izaguirre and Matthey 2004).
It is important to highlight that the exploration of conformational space will be automated as much as possible, to relieve the user from method parameter exploration and simulation setup, and allow him or her to concentrate on modeling problems. For this purpose, we will extend a recommender methodology that we have developed for selecting parameters for fast electrostatic methods (Crocker et al., 2004).

**Multiscale extension of the sampling protocol.** Multiple time stepping (MTS) algorithms work by evaluating the fastest varying forces more frequently than the slowly varying forces, resulting in a longer MD time step. MTS methods such as r-RESPA (Tuckerman and Berne 1992) have been used in conjunction with HMC. We are interested in combining r-RESPA with SHMC. In addition, we are considering a more stable, symplectic MTS algorithm known as MOLLY (García-Archilla et al. 1998, Ma and Izaguirre 2003). Our goal is to extend these methods to work with SHMC. This will require modification of shadow Hamiltonians for use with MTS integrators.

**SHMC sampling efficiency enhancement.** SHMC has shown great promise in finding new conformations, but it is currently limited by the variance of reweighted observables. In order to maintain statistical consistency, it is currently not possible to explore the full capabilities of SHMC. To overcome this, we will pursue algorithms that decrease the variation of the reweighted values. One method that we are examining is control variates (CV) (Fisher 2000). CV uses information about the expected value of an auxiliary variable and its correlation with the value of interest to improve the variance.

3.3.3 Selection of conformations and evaluation of docking protocols

This subproject is interested in choosing an optimal set of conformations from the conformational space for flexible docking. Several techniques will be used and evaluated to select conformations for docking, including random selection, uniform in space or time, most distinct, most representative by clustering or conformational space spanning metrics, and uniform within one cluster.

**Filtering scheme** Let us assume that we want to select $k$ conformations as a “representative set” for $n$ different conformations available. Random selection is one of the easiest procedure to select any $k$ out of $n$ different choices. But this does not guarantee that we will get a good representative set.

Conformations can be selected from sampling after certain number of timesteps. This approach requires $O(n)$ but still, it may not give us a good representative set that spans optimally over the entire space.

Ideally, we need an algorithm which will select conformations uniformly over space. This can be obtained by modeling the set of conformations as points in higher dimensions (given by the number of dihedrals) and computing the convex hull of these set of $n$ points. An advantage is that the volume of the convex hull quantifies the span of the conformational space. But computing the convex hull of $n$ points in $\mathbb{R}^d$ requires $O(n^{\frac{d}{2}})$ time, in the worst case (de Berg et al. 2000), which makes this approach impractical to implement. Using an output sensitive algorithm, we can achieve some improvements. The method by Chan (1996) can compute the convex hull of $n$ points in $\mathbb{R}^d$ in running time $O(n \log k + (nk)^{1\over d(2^{d-1})^{\log O(1)}})$, where $k$ is the complexity of generating the convex hull. This has also exponential complexity.

The idea is to approximate the convex hull using a bounding box. A bounding box of $n$ points in $\mathbb{R}^d$ can be computed in $O(nd)$ running time. We can select the points or conformations by slicing the bounding box. Although the bounding box is not very accurate, it is a good approximation to the convex hull.

Cluster analysis is another approach to selecting a representative set (Leach 2001). The points in $\mathbb{R}^d$ are usually scattered in groups. By cluster analysis, these groups can be identified and then the representatives can be extracted for docking from each of these groups.

Our plan is not to just use one of these techniques. Rather we will study the representative set generated by the different techniques described above. One useful metric for this study can be to check the difference in dihedrals in each of the conformations selected and finding the mean. This will quantify the span of the representative sets for each filtering technique. We will also study the effect of span of the set for docking.
Evaluation of Docking Protocols  Optimality will be evaluated by the ability of a docking algorithm to dock known inhibitors of test systems (sensitivity) and the ability to discriminate these from inactive compounds (specificity). The efficiency of docking will also be determined by the computational cost to generate a docked structure.

The docking protocols that will be evaluated will be the following:

1. FlexE, which produces an average structure for parts that are similar across different structures, and fragment libraries for parts that change.

2. Autodock and DOCK, which use multiple grids.

3. Restricted systematic search: flexible ligand, rigid receptor docking of each MPS vs. each ligand of interest. This technique should be more accurate than (2) above, since it keeps the ensemble of conformations available. If the number of MPS is kept relatively low, with the help of the selection procedure, this method might also be computationally competitive. This approach is also optimal for the grid with our GEMS toolset, since it involves multiple independent computations. The results will give probabilities for the conformations based on their Boltzmann weight.

4. A voting scheme will also be tested, which combines the results from the approaches above in order to determine whether or not a compound binds to the receptor.

3.4 Validation experiments

The data sources used will be HIV protease and DHFR, which exhibit motions in several timescales and for which abundant data exists (cf. Bouzida et al. 1999; Broughton 2000; Jacobs et al. 2001; Verkhivker et al. 2002; Osterberg et al. 2002; Ishima and Torchia 1999); and HIV integrase, for which a flexible docking algorithm has been validated experimentally (Carlson et al. 2000). A subset of the Active Compounds Database (ACD) will also be screened to seek potential new compounds.

For the coverage, selection, and docking protocols described above, and for a set of potent and moderate inhibitors, and inactive compounds taken from the literature, docking experiments will be performed to answer the following questions:

1. What is the sensitivity, specificity, and computational efficiency of different combinations of sampling, selection, and docking protocols for the exploratory and refinement phases? Our hypothesis is that where large loop or domain motions are involved, the systematic search approach will be more accurate. However, for the refinement, it is likely that one of the averaged methods will be most efficient.

2. What is the optimal size of the MPS set? Simulations will be performed varying the size of the MPS according to the selection criteria described above, to determine some practical values and analytical bounds if possible (for example, as a fraction of coverage of conformational space).

3. What is the effect in accuracy and efficiency of simulating the following: (i) both docked and unbound structures in the generation of MPS; (ii) explicit solvent; (iii) large motions from NMR or dynamics simulations analysis.

To measure predictive capabilities of the optimal combinations of selection, coverage, and docking above, the ACD will be screened. For the best putative interactions, NMR experiments of the combined system will be performed to determine the activity of the compounds (cf. Peng et al. 2004).

3.4.1 Grid Portal

The framework for the grid portal will be based on the NSF NMI suite of software tools. Figure 3 shows a high level view of the software packages employed for our grid portal. Conceptually, the user will log on to the portal through the OGCE (Open Grid Collaboratory Environment). The user will be presented a view of the grid resources (via OGCE) as well as a portlet representing the available GEMS tools.

From the portlet, the user will be able to manage, view, and create new results using tools properly developed to work with GEMS or via well-published interfaces to the Molecular Simulation Database (MSDB).
Inside of the MSDB, the various data results will be kept along with meta-data representing how the results are generated. By virtue of this meta-data, the portlet will offer several novel approaches to working with data that can offer significant improvements to the research process (see Figure 4):

- **Model exploration:** While a data run may produce the exact results desired, oftentimes it may be desirable to examine interim steps in the simulation and make alternative assumptions. To that end, the portlet will provide the capacity to fully explore the model by modifying inputs to drive the transient data in an alternative direction. Since GEMS fully understands the required original inputs to the data, it is quite trivial to allow the user to modify inputs to the tools in transient steps and create a new branch of data with alternative results. Figure 4 shows a conceptual screenshot of the user modifying the simulation input to explore an alternate branch midway through the transient results.

- **Exploring the work of others:** A frequent issue associated with simulations is the difficulty in reproducing the data produced by other researchers. Typically one is often faced with significant amounts of time simply getting the original simulation source to work or to adapt the inputs for comparison. To that end, the MSDB will provide a common platform for evaluating results. For previous simulations that have been executed, images of the simulations themselves (the executable), the environment in which the simulation was run (i.e. RedHat Enterprise 3.0, x86-32), and the input arguments will all be preserved in the database. Thus, with appropriate modules to transform the output data for storage in the MSDB, one can easily explore and compare the work of others without significant effort.

### 3.4.2 Distributed Databases

A major sub-problem associated with introducing the concept of the Molecular Simulation Database (MSDB) is the ability to manage the scale of the data in an efficient and reliable manner. Consider the following scenario for a researcher:

* A researcher wishes to run a set of GEMS simulations that require 2 weeks to complete. The simulation generates 10 GB of output data and 1 TB of transient data from a 100 MB set of inputs. For the next two months, the researcher will need immediate access to the data and cannot afford to lose the data. Afterwards, the researcher wishes to make the data public but has no specific time constraints for access.

In essence, the researcher would be interested in three key qualities, namely reliability (do not completely lose the data), availability (how soon the data needs to be accessed), and efficiency (minimize the total storage consumed). On one extreme, one keeps multiple copies of the original data to help with both reliability and availability. However, this approach suffers from scalability issues due to the sheer scale of the data itself. In contrast, one could store only a single copy of the data itself to maximize efficiency. The problem with this...
approach is that a failure in the host holding information will result in a penalty of redoing the 2 weeks of simulation time. Furthermore, the information may only be regenerated if the original configurations have been preserved and were not lost with the data itself.

While there have been a variety of tools proposed to help virtualize the storage resources on the grid, the ability to consider the fault tolerance implications of the data (availability, reliability) have not been developed. Ideally, a researcher would simply like to request the portlet to automatically manage storage and balance the above metrics to keep within the time requirements of the researcher. Most notably, we observe three key observations about the simulation process itself. First, the data itself has different requirements for fault tolerance. The requirements for fault tolerance of the output data (high reliability/high availability) are significantly different than the transient data (reduced reliability/reduced availability). Since data can be re-created provided the inputs are kept intact, the system should ensure that the inputs are preserved and balance the redundancy in both the transient data and final output data versus the availability requirements for the data (See Figure 5).

Second, the data itself will undergo two key stages. In the first stage, identified as the workspace, the researcher needs timely and reliable access to the data since the researcher is actively using the data. Once the work is complete, the timeliness of the data becomes less critical and the data moves into the repository stage. In this stage, the data is made available for other researchers with the expectation that data may take some time to be fully regenerated.

Finally, the data itself is only valuable if it can be identified within the overall data repository. Since the grid introduces the potential to create tremendous amounts of data, a data storage mechanism must possess the ability to identify key properties about the data itself. Otherwise, the data is useless as it could take nearly as much time to either reproduce or identify useful as it did to simply re-run the entire simulation.
process itself. To that end, data should be identified with a meta-data to annotate critical aspects such as data significance and how the data was produced, which requires the standardization of descriptors for molecular simulation.

Over the course of developing the Molecular Simulation Database, we will address the following research problems:

- **Formalization of F-T aspects**: In order to balance the metrics of reliability, availability, and efficiency, the first step is formalize the metrics to govern allocation on the grid storage mechanisms. This step will also include methods for assessment of fault tolerant capabilities for grid resources.
- **Grid portlet**: The grid portlet will automate the process of providing fault tolerance to the user simulations. The portlet will offer both a simple F-T wizard for beginning users that automates the entire procedure and advanced configuration tools for more advanced users. The portlet will abstract all interactions with the MSDB to the user.
- **XML Meta-data**: In order to correctly retrieve data from the MSDB, the actual data will include meta-data to include the significance and mechanisms for creation. Examples for creation mechanisms would include input parameters, OS/architecture employed, and pointers to the executable images. This work will include the development of XML schemas and Java/C++ modules to govern storage to/from the virtual MSDB.

### 3.5 Project Timeline Benchmarks

**Year 1**
- Algorithms: SHMC with dihedral MC moves; Graph theoretics, conformation analysis, and selection metrics; systematic search with flexible docking
- Experiments: Selection and categorization of NMR data testbed; Configuration of NMR facility; PostDoc and student training
- GEMS Database: Fault Tolerance Formalization; Prototype F-T storage; Stand alone interface to GEMS

**Year 2**
- Algorithms: Multiscale SHMC using NMR relaxation data; Evaluation of Flex E docking protocol
- Experiments: NMR relaxation data for HIV protease (fast and slow) to establish a basis for subsequent validation
- GEMS Database: F-T with Globus/OGSA-DAI linkage; Grid Portal to GEMS

**Year 3**
- Algorithms: Transition path sampling to generate MPS hinge motion SHMC
- Experiments: Validation of putative protein-ligan interactions; existence of interaction and change in mobility
- GEMS Database: Automated F-T management via grid portal; Model exploration portlet

### 4 Broader Impacts

#### 4.1 Dissemination of Results and Computational Tools

**4.1.1 Open Source Computational Tools**

The PI has led open source initiatives targeted to accelerate the discovery of new algorithms for molecular simulation. The core of this endeavor is ProtoMol, an object oriented open source molecular simulation
tool for rapid algorithm testing and maturation. PROTOmOl is available to the research community and developers through the Source Forge repository. All of the novel computational tools developed as part of this research work will be open source and freely available to academic and government institutions. Open source distributions truly share the fundamental scientific knowledge revealed by the research team and provide the crucial details necessary for full understanding within the scientific community. The novel capabilities of our molecular simulation framework are then fully reproducible as the full implementation of algorithms are available to the research community. Open source grid management and database primitives will be available for peer review and refinement to immediately foster the next generation of innovation.

4.1.2 Virtual Portal Access to Grid Computation

This project enables the distribution of results via the inherent nature of the research proposed. To deliver the novel grid enabled database functionality, we will provide a virtual portal to the scientific community through which they can access the computational power and novel distributed stored data as a single abstract resource. To ensure a higher level of QoS, the researcher will be able to unite their resources with the grid while simultaneously increasing the utility of the grid to the research community as a whole. Enabling a greater number of researchers the abstraction of grid magnitude computational power and data access coupled will accelerate the introduction of novel techniques and algorithms.

4.1.3 Grid Distributed Database (Simulation Images and Complex Conformations)

Efficient, selective, and reliable access to massive amounts of information will turn overwhelming data into fundamental knowledge. The proposed work takes massive amounts of molecular conformational data from the research community and implements novel utilization and computational regeneration techniques to provide ensembles of conformations that enable efficient docking simulations. In addition to the availability of conformational data we propose to collect, generate, store, and provide molecular simulation data. The implementation of standard molecular simulation configuration XML descriptors will both serve to further the community’s fundamental understanding of molecular simulation and provide an essential repository for simulation data enabling efficient reproducibility for scientific verification, statistical analysis, and adaptation.

4.2 Educational Contributions

4.2.1 Courses

The PI has developed a course in computational biology (CSE598). The course covers mathematical, computational, and scientific applications of biomolecular modeling. The most recent enrollment was about half computer science and engineering students and half students of physical sciences, including chemistry, physics, and biology, at both the graduate and senior undergraduate level. The PI has also incorporated computational science into the data structures course for juniors that he has taught the past four fall semesters. Some of the student projects included significant MD simulations, utilization of haptic interfaces, and testing of new algorithms for kinetic data structures. Dr. Striegel (co-PI) has included the use of grid computing into a course on Computer Security. The course utilizes Condor and a grid of Itanium2 workstations to analyze the effectiveness of passwords and encryption algorithms. The research described in this proposal will also be integrated into existing biochemistry courses and serve as a key cornerstone in a new interdisciplinary course focused on computational grid implementation and management. The collaborations described in this proposal will provide a wealth of possibilities for examples, course projects, and lab exercises.

4.2.2 Educational outreach

A significant effort has been undertaken to establish Research Experiences for Undergraduates (REUs). Over the last four years, the PI has directed over twenty undergraduate research projects. Supplementary funds for REU will be requested if this proposal is funded. Dr. Striegel has directed over seven undergraduate research projects. The projects have focused on components of GIPSE ranging from Java GUI development to XML database interaction to Globus GridFTP interactions. Undergraduates from Duke, Caltech, Goshen...
College, Tri-State, and Notre Dame have participated in the summer REUs sponsored by the PI and Co-PIs. The PI appreciates the need to bring the challenges and rewards of science and engineering to underrepresented minorities. Dr. Izaguirre has recently involved female and Hispanic students in his group’s research projects. He serves as the faculty advisor for the Notre Dame chapter of the Mexican American Engineering Society (MAES) and the Society of Hispanic Professionals in Engineering (SHPE). Directed efforts will be made to recruit and involve minority students in the challenging work presented in this proposal. In order to proactively engage minorities early in their academic development, we propose to provide educational presentations to middle school students in two local schools with disproportionately large minority populations. These biannual presentations will include multimedia simulation progressions and simple chemical reaction demonstrations. Follow-up dialog between the research associates and the middle-school students could be made available at their teacher’s request.

5 NSF prior support

J. A. Izaguirre, “CAREER: Scalable Mathematical and Computational Models for Biomolecular Modeling,” 2/1/02-1/31/07, $384,918

1. Theoretical understanding of limitations of multiscale approaches, underscoring the need to introduce stochasticity for long simulations.

2. Contributed new algorithms for molecular dynamic, that are 2 to 4 times faster than the state of the art r-RESPA

3. Contributed an O(N) algorithm for sampling of conformations of molecules, Shadow Hybrid Monte Carlo, that gives an order of magnitude or more speedup

4. Contributed a parallel O(N) algorithm for the N-body problem with periodic boundary conditions, which is 4 or 5 times faster or more than the fastest available algorithms

5. All these algorithms have been made available in a cross-platform, open source software, ProtoMol (http://protomol.sourceforge.net)

6. We have provided a web service, MDSimAid (http://mdsimaid.cse.nd.edu), that chooses optimal algorithm and parameters for a molecular simulation. We have shown that a hybrid optimization approach (rules give a guess, run-time optimization refines the choice)


While a wide variety of techniques have emerged to increase the efficiency of the network, the techniques have met with varying degrees of success. This career work aims to develop novel techniques for network efficiency through transparent bandwidth conservation while avoiding many of the pitfalls associated with previous approaches. Individual facets of the work include a paradigm-shifting approach to multicast (stealth multicast) whereby redundant packets are dynamically converted to/from multicast, inter-domain peering protocols for expanding the benefits of bandwidth conservation, and extensive analysis/exploitation of existing applications to increase redundancy but yet improve overall system efficiency.

In addition, this work will serve as a catalyst for the development of large-scale group oriented applications. Furthermore, the work will have a direct impact on the efficiency of the Internet (quality of service), pricing and resource management, and how applications are designed to cooperate with the underlying network. Preliminary publications as a result of this work have already appeared in (A. Striegel 2004 and A. Striegel 2004a).
References


