High performance computing for the application of molecular theories to biological systems.

Laura J. Douglas Frink, Marcus G. Martin, Andrew G. Salinger, and Micheal A. Heroux
Sandia National Laboratories, Albuquerque, New Mexico 87185
E-mail: ljfrink@sandia.gov

Abstract. In this paper we summarize our work to bring modern numerical methods to bear in computations for density functional theories (DFTs) of inhomogeneous fluid systems. We present the general mathematical structure of the problem, and briefly discuss different strategies for solving the problems. Finally, we present a few recent results from calculations on complex peptide assemblies in lipid bilayers to demonstrate the application of the methods in one complex 3-dimensional system. We find that while solver strategies developed and optimized for partial differential equations (PDEs) can be applied to these systems of equations, they do not provide optimal solutions with respect to speed, memory use, or parallel partitioning.

1. Introduction
A significant amount of effort has been invested in high performance computing over the past 20 years. The result has been the development of new computing platforms (from fast vector machines to modern parallel computing platforms that come in many varieties), novel algorithms (from linear solvers to parallel partitioning algorithms to multiscale methods), and a general increase in both size and complexity in the systems that can be studied with numerical methods. Most of this effort has focused on optimization of methods in the context of continuum mechanics where partial differential equations (PDEs) are dominant. This emphasis is entirely appropriate when the engineering challenges at hand are macroscopic in scale (mm to meters), and when smaller length scales can be represented by simple equations of state. However, in the past 5 years, there has been growing interest in nanotechnology and the physics of materials at very small length scales. Concurrent with these developments, there has been an explosion in our ability to catalog the information content and complexity (e.g. DNA and proteomes) in biological systems. However, physics based computing at nanoscopic and biological length scales (Angstroms to microns) is still a major challenge.

Probably the most widely utilized method at the nanoscale has been molecular dynamics simulation. These methods are used regularly to simulate a variety of problems from materials to biology. Often times full atomistic detail is included[1], sometimes coarse-grained models are used in order to achieve simulations over longer length and time scales[2, 3]. While very rigorous MD simulations can be achieved for simple systems (e.g. atomistic fluids that are not too dense), there is always a question as the system complexity increases about whether the simulation has done an adequate job of sampling the ensemble, and whether the results (either trajectories or time averaged observed properties) are meaningful. Monte Carlo simulations may be used to
sample ensembles independent of dynamic trajectories; however, it is only recently that good sampling algorithms have been developed that allow decent statistical sampling in vacuum of the amino acid side chains found in proteins[4]. We note that a new geometrical method has been recently developed that can sample much larger changes in protein conformations; however, the connection of this sampling procedure to the rigorous underlying statistical mechanics is not clear[5].

Recently we demonstrated just how severe this sampling problem can be by performing Monte Carlo simulations on a 10-bead polymer chain immersed in fluids where the density varied from gas like ($\rho \sigma^3 = 0.1$) to liquid like ($\rho \sigma^3 = 0.7$) at high temperature $kT/\epsilon = 1.2$ [6]. Note that all of biology and all simulations of biological systems are performed at high fluid densities and low temperatures (relative to the critical point of water) making them even more difficult. We showed that reasonable conformational sampling of this very simple chain could only be achieved for low to moderate solvent densities $\rho \sigma^3 \leq 0.5$ if the solvent particles are included explicitly in the simulation. While our simulations were based on MC methods, we note that the same core sampling issues arise in MD simulations making it very unlikely that MD simulations can adequately sample conformational space of complex chain molecules. Given these challenges many groups use implicit solvent models rather than including solvent explicitly in their simulations[7]. We considered three implicit solvent models in our investigations. In all cases good statistical sampling of the 10-mer chain was achieved at all solvent densities and temperatures. The first implicit model treated the solvent as a continuum, and yielded very poor physical results for the chain structure (radius of gyration). The second was based on an accurate 3-dimensional molecular theory for the solvent to compute the solvation structure and energy of any given trial conformation of the chain. These calculations were performed at each step of the MC simulation, and were quite expensive; however, they produced very accurate results for the chain structure. A pair potential based approximation to this approach was also considered and yielded results of intermediate quality. We revisit these results to highlight one example where fast algorithms for these complex molecular theories (i.e. an accurate implicit solvent) could have a major impact in advancing the state of the art for simulating complex fluids generally, and particularly for the classes of problems found in biological systems.

In the remainder of this paper we summarize our efforts to leverage where possible and develop where necessary modern computing approaches for molecular theory based studies of complex fluid systems. Our work has focused on a class of theories known as density functional theories (DFTs); however, there are other integral equation theories (IETs) of importance as well [8]. In fact these classes of theories are intimately related. DFTs are formulated at the level of a free energy functional while IETs are formulated at the level of correlation functions which are the second functional derivatives of the DFT free energies. We note that the DFTs we discuss here address inhomogeneous classical fluid systems where an exact Hamiltonian is not available. The more common quantum mechanical DFTs, where one computes electron densities, have the advantage of an exact Hamiltonian, and a different mathematical structure [9]. In the next section, we first give a brief overview of one class of mathematics found in this field, we then discuss our work on numerical methods, we present one recent set of calculations that have been facilitated by the algorithms we have developed, and then summarize and briefly discuss future work needed in this area.

2. Mathematics and Molecular Theories

In Fluid-DFTs a free energy functional, $\Omega$, depends on a set of critical fields, $\psi$, in the problem of interest, $\Omega[\{\psi(r)\}]$. The minimization of this free energy with respect to all fields results in a system of residual equations to be solved. Generally any of these residual equations can be
\[
\frac{\delta \Omega}{\delta \psi_i} = 0 = I(\psi(r)) + D(\psi(r)) + F(\psi(r))
\]  

(1)

where \( I \) is a general integral operator, \( D \) is a general differential operator, and \( F \) is some function of the fields involving neither integral or differential operators. For example a problem with a charged atomistic fluid model will have two critical fields, the fluid density and the electrostatic potential. In this case the numerical problem consists of an integral equation coupled to a familiar PDE, Poisson’s equation.

Often, some contributions to the free energy functionals are not local. Rather, the free energies are written as functionals of both the critical fields and some nonlocal variables, \( n \) with \( \Omega[\psi, \{ n_\gamma \psi \}] \). The nonlocal variables are themselves functionals of the critical fields of interest. If the free energy functional contains an integral operator term,

\[
\frac{\delta \Omega}{\delta \psi_i} \delta \psi_i (r) = \int \sum_\gamma \partial G_\gamma (\delta n_\gamma (r^\prime)) \delta n_\gamma (r^\prime) \delta \psi_i (r) \, dr'.
\]  

(3)

While the free energy minimization is always defined by the general expression in Eq.1, the formulation of the matrix problem may proceed in two ways [10].

The first approach is to consider a solution vector that contains only the critical variables. The Jacobian (or really Hessian) used in the matrix problem is then

\[
J_{ij}(r, r') = \frac{\delta^2 \Omega}{\delta \psi_i (r) \delta \psi_j (r')} = \int \sum_\epsilon \sum_\gamma \frac{\partial^2 G_\gamma (\delta n_\epsilon (r^\prime)) \delta n_\epsilon (r^\prime) \delta \psi_i (r) \delta \psi_j (r')} {\delta \psi_i (r) \delta \psi_j (r')} \, dr''.
\]  

(4)

Often these nonlocal variables are simple linear functionals of the critical fields defined as

\[
n(r) = \int w(r, r') \psi (r') \, dr'
\]  

(5)

with \( w(r, r') = \delta(r - r') - R \) or \( w(r, r') = \theta(|r - r'| - R) \). where \( R \) is some characteristic dimension (a particle size or a bond length). The Jacobian can then be written

\[
J_{ij}(r, r') = \frac{\delta^2 \Omega}{\delta \psi_i (r) \delta \psi_j (r')} = \int \sum_\epsilon \sum_\gamma \frac{\partial^2 G_\gamma (\delta n_\epsilon (r^\prime)) \delta n_\epsilon (r^\prime) \psi (r') w_\epsilon (r', r')} {\delta \psi_i (r) \delta \psi_j (r')} \, dr''.
\]  

(6)

Clearly in order to compute a Jacobian entry with this structure will require a second order \((N^2)\) operation in order to locate the intersection of the weight functions \( w_\gamma (r^\prime r, r) \) and \( w_\epsilon (r^\prime r, r') \). In some Fluids-DFTs the definition of nonlocal variables can be even more complex resulting in multiple integrals for each Jacobian entry. In real space, this approach leads to matrix coefficient calculations that range from time consuming to completely impractical.

One way to overcome this problem is to to formulate the real space matrix problem in terms of not only the critical fields, \( \psi \) but also all of the nonlocal density variable, \( n \) [10]. This approach is akin to the transformation of a higher order PDE to a system of first order PDEs by introduction of additional variables in the problem. Taking a case where the only critical field is the density \( \rho \) and where there are linear nonlocal density variables as defined above, the
contributions of the nonlocal free energy term to the extended matrix problem written in block form is

\[
\begin{bmatrix}
-\frac{1}{\partial n_i n_i} (\mathbf{r}) & w_\gamma (\mathbf{r}, \mathbf{r}') \\
\sum_\gamma \frac{\partial G}{\partial n_i \partial n_j} (\mathbf{r}) & 0
\end{bmatrix}
\begin{bmatrix}
\Delta n_i \\
\Delta \bar{\rho}
\end{bmatrix}
= - \begin{bmatrix}
R_{NL}(\mathbf{r}) \\
R_{EL}(\mathbf{r})
\end{bmatrix},
\]

where \( \mathbf{r} \) is a row of the matrix, \( \mathbf{r}' \) is a column of the matrix, \( R_{NL} \) is the residual for the equation that defines \( n \) (Eq.5), and \( R_{EL} \) is the residual of the Euler-Lagrange equation (Eq.3). Note that there are no integrals in the Jacobian, and so the complexity of filling the matrix has been considerably reduced. We will refer to Eq. 7 as the first order formulation of Fluid-DFTs. In contrast Eq. 6 is an example of a higher-order formulation.

3. Computing for Molecular Theories

Our efforts to develop numerical methods for DFTs for inhomogeneous fluids began with the development of a Newton’s method real space approach. We utilized algorithms (partitioning and parallel iterative solvers) that had been optimized for the solution of PDEs (particularly reacting flow codes) on large distributed memory parallel computers[10, 11]. This code was robust and in fact applied to a variety of complex systems including wetting of chemically heterogeneous surfaces[12], capillary condensation in disordered porous media[13], the distribution of ions in gramicidin A (a simple ion channel protein)[11], and the structure of lipid bilayers based on coarse-grained models[14]. One piece of critical understanding that came out of those investigations is that the state complexity of these DFTs can be quite large. In several different kinds of systems we found many metastable (and unstable) solutions by application of arc-length continuation (ALC) algorithms that had originally been developed for PDEs. Our initial matrix based implementations of DFTs included both first order and higher order formulations as discussed above[10]. However, the extended size of the matrices in the first order formulation resulted in poor numerical behavior so it was not possible to realize any advantage at that time.

To summarize, our initial implementation provided complete access to 2-dimensional systems for the first time, but it was not fast enough for routine investigation on 3-dimensional systems particularly given that the most powerful applications of the methods (i.e. either as implicit solvents in coupled simulation or in full investigations of problems with complex state behavior) require generation of O(100-10000) solutions to the DFT equations. Given this limitation, we then developed a matrix free method with fast Fourier transforms (FFTs) to compute certain convolutions in the theory[15]. This approach led to a code that could be applied to 3-dimensional problems using very modest computer resources (single processor workstations). However the FFTs limit the application space of the DFTs to cases with periodic boundary conditions, and since no matrix is stored, matrix based preconditioning methods cannot be applied to converge difficult nonlinear problems.

More recently we have returned to the first order mathematical approach in another matrix based real space implementation of the problem[16]. We have developed and implemented a new method based on segregated Schur complement techniques. This approach exploits the fact that some variable couplings are simple algebraic expressions and act as constraints on the remaining equations. Proper reordering and specialized preconditioning can reduce solver memory costs and implicit problem dimensions by an order of magnitude or more, while at the same time provide a truly scalable solver. This approach is a hybrid in that it leverages many lessons from the linear algebra of PDEs and multiscale simulations, but it specializes the formulation for each new DFT that is implemented. These algorithms are fast enough and robust enough to be applied to 3 dimensional systems of significant complexity and modest size. They can be interfaced cleanly with other matrix based analysis software (ALC and optimization software for example). Finally there are clear opportunities for both improvement of the methods, and extension of the general framework to a broad suite of DFTs of varying complexity.
4. One example problem: peptide assemblies in lipid bilayers

To demonstrate the utility of these methods for studying complex problems we present some results from recent studies on the structure of peptide assemblies in lipid bilayers. Specifically we consider the assembly of alpha-helix peptide fragments. Such fragments are often the anchors that bind proteins to membranes. However they can also assemble into bundles that promote pore formation in the membrane. Previous computational investigations of this problem have applied a synthesis of molecular theory and phenomenological theory[17] or have neglected the bilayer solvent interface altogether [18]. Our investigations are the first rigorous 3-dimensional molecular theory based studies of this problem.

Our calculations are based on coarse-grained (CG) models of the lipid and water molecules as well as the peptides. The CG lipid molecule is a linear chain with 18 total CG beads per molecule with two central head-group beads. CG solvent model we use is a single bead. In order to ensure that this lipid-solvent mixture will form bilayer assemblies, we set both the like interactions (solvent-solvent, tail-tail, head-head) as well as the solvent-head group interactions to be favorable. In contrast the tail-head and tail-solvent interactions are purely repulsive. The details of the model interaction potential can be found elsewhere [14]. The CG peptides are cylinders that can have amphipathic chemistry where one side of the cylinder is polar (modeled as attractive to head groups and solvent).

When the numerical problem is formulated using the first order method described above, there are 44 unknowns per node in the problem at hand. Computations were done for a variety of system sizes ranging from $1.5 \times 10^6$ to $4 \times 10^6$ total unknowns. Calculations were performed on 60-160 processors of the Sandia system red squall, a 258-node, dual processor Opteron-based system ( 2.2 GHz processors with 4 GB memory per node) using a Quadrics Elan4 high-speed interconnect. Solutions were generated using arc-length continuation algorithms, and convergence was obtained in 3-10 nonlinear iterations for most cases. More than 1000 solutions were generated in a study to fully investigate the nature of these peptide-lipid assemblies, and the effect of the amphipathic chemistry of the peptides on the structure of the assemblies. The details of the physical results will be presented elsewhere. Here we demonstrate that complex assemblies such as the toroidal membrane structure were found in our investigations. In addition, we found that like other problems in inhomogeneous fluids, this system has a complex state space involving multiple solutions, metastable, and unstable states. Figure 1 shows one example of a toroidal pore found in these investigations. Figure 2 shows a plot of the total number of solvent particles per unit area of the uniform membrane in the system as a function of the amphipathic chemistry on the cylinders.

![Figure 1. Density profiles for tail groups (left), head groups (middle), and solvent (right) in two slices of a 3-dimensional calculation on CG peptide assemblies in lipid bilayers. This case shows a membrane spanning pore. The top row is a slice through the center of the bilayer perpendicular to the peptides. The bottom row is a slice parallel to the long axis of the peptides showing the membrane-solvent interface.](image-url)
Figure 2. The number of solvent beads per unit area of uniform membrane as a function of the interaction potential parameter, $\epsilon/kT$ between the polar region on the CG peptides and the solvent and lipid head groups of the fluid. The curve was generated with an arc-length continuation algorithm. As the polar interactions become more attractive (increasing $\epsilon$) a first order phase transition is found (at $\epsilon/kT = -0.096$) between a state with no pore and a state with a membrane spanning pore.

5. Summary and Future Work

This paper presents our efforts to bring modern numerical methods approaches to bear on nanoscale systems that can be studied with modern molecular theories. One of the key difficulties is that many different kinds of integral equation based molecular theories exist, and many of them have unique mathematical components. As a result each of them may present new challenges from a computing perspective. It is clear that a generic approach developed in the context of PDEs is not optimal for these problems. However, some algorithms that have been developed primarily for engineering analysis can be applied to great advantage in this field. Given the importance of interfacial physics for biology and nanoscale systems, we believe that significant advances will be made as a result of the application of molecular theories in general and density functional theories in particular in these fields.

References

[1] P.S. Crozier, M.J. Stevens, L.R. Forrest, and T.B. Woolf. Molecular dynamics simulation of dark-adapted rhodopsin in an explicit membrane bilayer: Coupling between local retinal and larger scale conformational change. *J. Mol. Biol.*, 333:493–514, 2003.
[2] S.O. Nielsen, C.F. Lopez, G. Srinivas, and M.L. Klein. A coarse grain models and the computer simulation of soft materials. *J. Phys. Cond. Matter*, 16:R481–R512, 2004.
[3] M. J. Stevens. Coarse-grained simulations of lipid bilayers. *J. Chem. Phys.*, 121:11942–11948, 2004.
[4] M.G. Martin and A.L. Frischknecht. Using arbitrary trial distributions to improve intramolecular sampling in configurational bias monte carlo. *Mol. Phys.*, in press, 2006.
[5] M. Lei, L.A. Kuhn, M.I. Zavodsky, and M.F. Thorpe. Sampling protein conformations and pathways. *J. Comp. Chem.*, 25:1133–1148, 2004.
[6] L.J.D. Frink and M.G. Martin. A combined molecular simulation-molecular theory method applied to a polyatomic molecule in a dense solvent. *Cond. Matter Phys.*, 8:271–280, 2005.
[7] B. Roux and T. Simonson. Implicit solvent models. *Biophysical Chemistry*, 78:1–20, 1999.
[8] D. Henderson, editor. *Fundamentals of Inhomogeneous Fluids*. Marcel Dekker Inc.s, New York, 1992.
[9] R.G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford Science Publications, New York, 1989.
[10] L.J.D. Frink and A.G. Salinger. Two and three dimensional nonlocal density functional theory for inhomogeneous fluids i. algorithms and parallelization. *J. Comp. Phys.*, 159(2):407–424, April 2000.
[11] L.J.D. Frink, A.G. Salinger, M.P. Sears, J.D. Weinhold, and A.L. Frischknecht. Numerical challenges in the application of density functional theory to biology and nanotechnology. *J. Phys.-Cond. Matter*, 14(46):12167–12187, November 2002.
[12] L.J.D. Frink and Salinger A.G. Wetting of a chemically heterogeneous surface. *J. Chem. Phys.*, 110:5969–5977, 1999.
[13] L.J.D. Frink and Salinger A.G. Rapid analysis of phase behavior with density functional theory ii: Capillary condensation in disordered porous media. *J. Chem. Phys.*, 118:7466–7476, 2003.
[14] L.J.D. Frink and A.L. Frischknecht. Density functional thoery approach for coarse-grained lipid bilayers. *Phys Rev E*, 72:Art. No. 041924, 2005.
[15] M.P. Sears and L.J.D. Frink. A new efficient method for density functional theory calculations of inhomogeneous fluids. *J. Comp. Phys.*, 190:184–200, September 2003.

[16] M.A. Heroux, Frink L.J.D., and Salinger A.G. Segregated Schur complement approaches to solving density functional theories for inhomogeneous fluids on parallel computers. *SIAM J. Sci. Comput.*, 2006. Submitted.

[17] Assaf Zemel, Avinoam Ben-Shaul, and Sylvio May. Perturbation of a lipid membrane by amphipathic peptides and its role in pore formation. *Eur Biophys J.*, 34:230–242, 2005.

[18] P. Lagie, M. J. Zuckermann, and B. Roux. Lipid-mediated interactions between intrinsic membrane proteins: A theoretical study based on integral equations. *Biophysical Journal*, 79(6):2867–2879, 2000.