Exact calculation of peptide-protein binding energies by steered thermodynamic integration using high performance computing grids

P.W. Fowler,[1,*] P.V. Coveney,[1,†] S. Jha,[1] and S. Wan[1,2]

[1] *Centre for Computational Science, Department of Chemistry, University College London, Christopher Ingold Laboratories, 20 Gordon Street, London WC1H 0AJ, United Kingdom*

[2] *Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, United Kingdom*

We describe a grid-based method to dramatically accelerate from weeks to under 48 hours the calculation of differences in binding free energy and its application to Src homology 2 (SH2) protein cell signalling domains. The method of calculation, thermodynamic integration, is briefly outlined and we indicate how the calculation process works from the perspective of an application scientist using either the UK National Grid Service or the US Teragrid. The development of a PDA-based steering client is especially useful as it gives the application scientist more freedom. Finally, we discuss our experience in developing and deploying the application on a grid.

I. INTRODUCTION

The ability to rapidly and accurately calculate a difference in binding free energies between two drug candidates and their target (usually a protein) is of interest in structural biology and of importance to the pharmaceutical industry. Such calculations are difficult to perform, require a substantial amount of computational resource and consequently take weeks to months to complete. It is our aim to dramatically speed-up an existing technique to allow a user to easily complete a binding affinity calculation within 48 hours.

The RealityGrid [1] steering library provides a generic computational steering interface [2]. Since April 2004, it has been available to download under a liberal open source license [3]. We have integrated this steering library with NAMD2 and VMD, extending their existing steering capabilities [4]. NAMD2 is a highly-scalable classical molecular dynamics application used primarily for biomolecular simulations [5], and VMD is its sister visualisation package [6].

II. IN-SILICO DRUG DESIGN?

There has been a significant research effort, both commercial and academic, expended in the last ten years to understand the precise nature and mechanism by which small peptides bind to SH2 protein signalling domains (see Figure 1). The ultimate aim of this effort is to develop drug leads that inhibit specific SH2 protein domains. These systems are well-studied but remain poorly understood.

Different SH2 protein signalling domains are found within many cell signalling pathways. Inhibition of specific SH2 domains is expected to lead to control of specific pathways through the activation or deactivation of the genes that they regulate. Precisely because these pathways are so generic, the possibilites to influence a wide range of ailments, from osteoporosis to immunological disorders, are large [7].

When developing a drug it is vital to know the Gibbs free energy of binding (ΔG) [8]. Relating the strength of binding with structural information, such as how the candidate drug interacts with the SH2 domain, is an established and important method for gaining insight and thereby developing good drug leads. It is possible to measure experimentally both ΔG and its components, the enthalpy ΔH and the entropy ΔS, using isothermal calorimetry [8]. There also exist computational methods for computing the difference in binding free energy, $\Delta\Delta G$, between two peptides. We will discuss in this paper one such method, that of thermodynamic integration.

Thermodynamic integration requires the use of a thermodynamic cycle as shown in Figure 2. The difference in the free energy of binding between drug A and drug B is given by:

$$\Delta\Delta G_{AB} = \Delta G_A - \Delta G_B \tag{1}$$

It is computationally impractical to assess either ΔG_A or ΔG_B directly. Instead we note that G is a thermodynamic state function and therefore sums to zero around a cycle. This allows us to consider instead

$$\Delta\Delta G_{AB} = \Delta G_1 - \Delta G_2 \tag{2}$$

Next we assume that G is a continuous function of a parameter, λ,

$$\Delta G = \int_0^1 \frac{\partial G(\lambda)}{\partial \lambda} d\lambda \tag{3}$$

*Electronic address: philip.fowler@ucl.ac.uk
†Electronic address: p.v.coveney@ucl.ac.uk

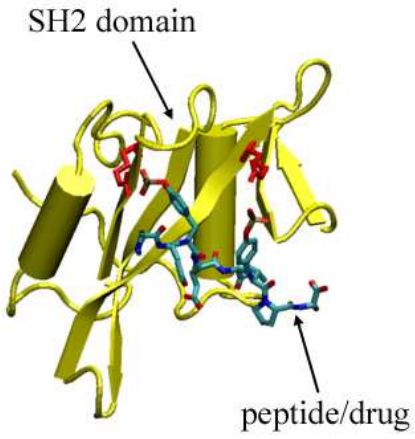

FIG. 1: A drug/peptide (cyan) bound to an SH2 domain (yellow)

.

where λ represents the extent the system has changed from A to B. Hence at intermediate values of λ, the system is in an unphysical state. The concept of changing one system into another leads to this procedure being referred to as an *alchemical* transformation. Fortunately, computing ΔG_1 and ΔG_2, whilst not trivial, is computationally tractable. Leach [9] contains a good description of the theory of thermodynamic integration which we will not go into detail here we will rather present a brief outline of the theory. We first substitute in the standard thermodynamic result

$$\Delta G(\lambda) = -kT \ln Q(\lambda) \tag{4}$$

where $Q(\lambda)$ is the partition function for our system. Further substituting for $Q(\lambda)$ in terms of the Hamiltonian and thereby the potential energy, U gives us the final result [9].

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U(\mathbf{r^N}, \lambda)}{\partial \lambda} \right\rangle d\lambda. \tag{5}$$

This integral is numerically evaluated over a number of equilibrated intermediate values of λ or windows. Typically, about ten simulations are required ($\lambda = 0.1, 0.2 \ldots$) for both ΔG_1 and ΔG_2. Further simulations are often needed to avoid the "end point catastrophe" as $\lambda \rightarrow 0$ or 1. Each simulation must be run for a sufficient length of time to ensure that it is equilibrated and the ensemble average of the differential of the internal energy with respect to λ (see Equation 5) has properly converged. A typical length of one such simulation is 1ns.

III. COMPUTATIONAL REQUIREMENT AND MOTIVATION

To calculate the computational requirement for our system, let us assume that we have at our disposal a 16 processor SGI Onxy2. Each 1ns run would consume either 380 CPU hours (for the domain+peptide system, ΔG_1) or 120 CPU hours (for the peptide-only system, ΔG_2). We estimate that a single difference in binding free energy would therefore require around 10,000 CPU hours on this machine. This is equilivalent to 26 days of continuous computation. This clearly illustrates the large quantity of computational resource required to compute a single value of $\Delta\Delta G_{AB}$.

Due to the small size of the system (14,000 atoms), twenty six days is fast for a calculation of this kind. Even this is a major obstacle to the application scientist interacting in any meaningful way with other scientists, notably experimental scientists in the present case. Making use of high performance computational grids, the UK National Grid Service [10] or the US TeraGrid [11], allows us to reduce the turnaround time to a manageable 48 hours. This exploits the ability of such a grid to deliver vast quantities of computational resource.

Let us consider how this process works in practice. The application scientist launches a NAMD2 job on a computational grid and monitors its progress using an appropriate steering client (desktop, hand-held PDA or browser-based) [4] . From this initial simulation, the scientist then spawns the required ten or so simulations. Each simulation is not quite independent: each is seeded from a checkpoint part way through the previous simulation leading to a "chaining" procedure. However, this procedure is rapid in terms of wallclock time. The scientist

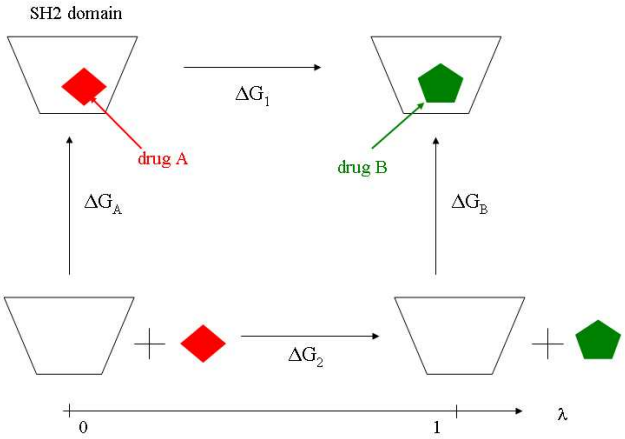

FIG. 2: Thermodynamic cycle for the difference in free energy of binding between drug A and B binding to an SH2 domain.

then monitors the convergence of the ensemble average and terminates or extends individual simulations as appropriate but can also launch jobs in this fashion. It is during this monitoring stage that the hand-held PDA client is of particular value as it allows the application scientist the freedom to leave his or her desk with the proviso that they remain within the range of an 802.11b wireless network [12]. Depending on the amount of computational resource available, the second set of chained simulations may be launched either at the same time as the first or subsequently. Finally, each simulation contributes a datum to a graph from which $\Delta\Delta G_{AB}$ is numerically integrated according to Equation 5. It is our aim that this entire process take no more than 48 hours as compared to weeks or months at present.

Gridded applications of this kind are widely applicable to other systems, for example the PGHS monotopic enzyme [13] or the MHC-peptide-TCR complex arising in molecular immunology [14]. We have extensive experience with the MHC-peptide-TCR systems and, due to their large size (at least 100,000 atoms), have previously used a wide range of off-grid HPC resources (HPCx, CSAR and the Pittsburgh Supercomputing Centre) to examine the bindings of different epitopes (short peptides) and the recognition of those peptides by the t-cell receptor (TCR). Within the present project we also have access to all these resources within the grid.

IV. GRID DEVELOPMENT AND DEPLOYMENT

To accomplish our aim of completing a binding free energy calculation in 48 hours we have integrated NAMD2 and VMD with the RealityGrid steering library [4], developed existing steering clients and created new ones. There are several advantages to integrating NAMD2 and VMD with the RealityGrid steering library. We gain flexibility by being able to run in a distributed manner on a Grid. The existing steering functionality is maintained, for example, the NAMD2 Interactive Steering (IMD) function which applies forces to parts of a molecule. But in addition we also gain the ability to steer parameters that previously were not steerable, for example, λ and various data collection parameters. We are also developing the ability to checkpoint simulations, rewind and restart simulations, migrate between computers and spawn separate simulations [4]. These generic capabilities were demonstrated within the TeraGyroid project at SC2003 which used another one of our codes (a lattice-Boltzmann algorithm, LB3D) also integrated with the steering library [15, 16].

To make these calculations easier for users to perform, a hand-held PDA .Net steering client has been created [12] and the existing steering client has been further developed to allow the monitoring and control of multiple simulations using tabs. The hand-held PDA steering client allows the application scientist to connect to a simulation via a Steering Grid Service (SGS) and then remotely steer and monitor the simulation. Real-time visualisation of the molecular system is less important than the requirement to monitor the convergence of many different simulations making a hand-held PDA steering client particularly well suited to this application.

Both the grids we are using, the National Grid Service (NGS) and the TeraGrid, are based on GT2 middleware. Whilst we have extensive experience in using GT2, OGSI::Lite [17] has proved essential in allowing rapid deployment and hosting of high-level services (such as the SGS registry) [18]. We plan to expose the NAMD2 application itself as a service through OGSI::Lite, which is necessary to be able to launch jobs from the PDA based .Net steering client. We note that the creation of the UK-

Light 10Gbps optical light-switched network provides a persistent link between the NGS and the TeraGrid.

V. OUR EXPERIENCE

Integrating the RealityGrid steering library with both NAMD2 and VMD took over a year and proved more challenging than our previous experiences of integrating the steering library with other scientific applications [4]. This was partly due to the requirement to maintain the existing NAMD2/VMD steering capability and partly due to adapting a third-party code rather than a locally-developed code. The main obstacle was adapting the original NAMD2/VMD communication model to use Globus sockets. In 2004, RealityGrid decided to mimimise or remove all dependencies upon Globus which helped solve this problem as we then were able to use TCP/IP sockets.

Deployment of the application (NAMD2) on each compute node within the grid(s) remains a laborious process, with specific scripts required for each platform [16]. Previously, we have worked closely with colleagues at Manchester Computing in setting up compiled versions of the code on each machine; it is now becoming increasingly necessary to transfer that technical knowledge to the application scientists in a lasting way via detailed documentation.

VI. SUMMARY

In summary, the project aims to calculate the difference in binding free energy between two peptides for the SH2 system during the All Hands Meeting 2004 and to report on its progress over the four days of the conference. The computed binding energy will be compared to the experimental value determined by our collaborators [8]. The simulations will make use of the compute nodes of the UK National Grid Service and the US TeraGrid. This new approach represents a revamp of the technique of thermodynamic integration, decreasing turnaround time and improving the ease of use. In the present paper we have discussed both the development of the necessary codes and our experience to date in deploying the relevant components.

VII. ACKNOWLEDGEMENTS

We thank our RealityGrid collaborators, in particular the SVE group at Manchester University, including Stephen Pickles, Andrew Porter and Robin Pinning, for developing the Steering Library and associated steering clients. The hand-held PDA client was developed by Roy Kalawsky and Simon Nee at Loughborough University. The PhD studentship of PWF is funded by EPSRC. SW is funded by the Edward Jenner Institute for Vaccine Research. The Edward Jenner Institute for Vaccine Research thanks its sponsors: BBSRC, MRC, GlaxoSmithKline and the UK Department of Health. This project is funded under EPSRC RealityGrid grant GR/R67699 and an NSF NRAC grant for 2004/5.

[1] http://www.realitygrid.org.
[2] J. Chin, J. Harting, S. Jha, P. Coveney, A. Porter, and S. Pickles, Contemporary Physics **44**, 417 (2003).
[3] http://www.sve.man.ac.uk/Research/AtoZ/RealityGrid/.
[4] S. Pickles, R. Haines, R. Pinning, and A. Porter, *Practical Tools for Computational Steering* (2004), submitted, AHM Conference Proceedings.
[5] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten, Journal of Computational Physics **151**, 283 (1999).
[6] W. Humphrey, A. Dalke, and K. Schulten, J. Molec. Graphics **14**, 33 (1996).
[7] W. C. Shakespeare, Curr Opin in Chem Biol **5**, 409 (2001).
[8] O. Y. Lubman and G. Waksman, J. Mol. Biol. **328**, 655 (2003).
[9] A. R. Leach, *Molecular Modelling. Principles and Applications* (Pearson Education Ltd., 2001), 2nd ed.
[10] http://www.ngs.ac.uk.
[11] http://www.teragrid.org.
[12] S. P. Nee and R. S. Kalawsky, *Developing a Roaming PDA-Based Interface for a Steering Client for OGSI::Lite using .Net: Practical Lessons Learned* (2004), submitted, AHM Conference Proceedings.
[13] W. L. Smith, D. L. DeWitt, and R. M. Garavito, Annu. Rev. Biochem. **69**, 145 (2000).
[14] S. Wan, P. Coveney, and D. Flower, *Large scale molecular dynamics simulations of HLA-A*0201 complexed with a tumour-specific antigenic peptide: Can the a3 and b2m domains be neglected?* (2004), in press, J. Comp. Chem.
[15] S. Pickles, R. Blake, B. Boghosian, J. Brooke, J. Chin, P. Clarke, P. Coveney, N. González-Segredo, R. Haines, J. Harting, et al., in *Proceedings of GGF10* (2004).
[16] J. Chin, J. Harting, and P. Coveney, *The TeraGyroid project - collaborative steering and visualization in an HPC Grid for modelling complex fluids* (2004), submitted, AHM Conference Proceedings.
[17] http://www.sve.man.ac.uk/Research/AtoZ/ILCT.
[18] J. Chin and P. Coveney, *Towards tractable toolkits for the grid: a plea for lightweight, usable middleware*, UK e-Science Technical Report, number UKeS-2004-01.