JCTC Journal of Chemical Theory and Computation

Rapid, Accurate, and Precise Calculation of Relative Binding Affinities for the SH2 Domain Using a Computational Grid

Philip W. Fowler,[‡] Sebastien Geroult,[§] Shantenu Jha,[‡] Gabriel Waksman,[§,||] and Peter V. Coveney*[,‡]

Centre for Computational Science, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, U.K., School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, U.K., and Department of Biochemistry & Molecular Biology, University College London, Gower Street, London WC1E 6BT, U.K.

Received October 24, 2006

Abstract: We describe and apply a method that reduces the time taken to calculate binding free energies using thermodynamic integration. This method uses a stack of grid software, which we call STIMD, that allows the scientist to easily distribute the necessary simulations around a computational grid thereby accelerating the process. We use this method to study how a series of phosphopeptides binds to the Src SH2 domain. The binding of phosphopeptides to the Src SH2 domain is described by the "two-pronged plug two-holed socket" model, and we investigate this model by reducing the length of the aliphatic side chain that engages the second of the two sockets through two successive alchemical mutations. Seven different values of $\Delta\Delta G$ have been calculated, and we report good agreement with experiment. We then propose an extension to this model using the insights gained from a free energy component analysis.

1. Introduction

Measuring the free energy of binding, ΔG, is a vital step in studying the interaction between any two molecules. Although theoretically exact techniques exist for calculating values of ΔG (e.g., free energy perturbation and thermodynamic integration), they require large amounts of computer time, and the values they yield are not always accurate or precise. Many research groups are developing theoretical approaches which hopefully both use less computer time and are accurate, for example LIE,[1] Jarzynski's equality,[2,3] MM-PBSA,[4−6] λ-dynamics,[7] ABF,[8] and 4D-PMF.[9] Existing techniques continue to be improved, for example through the use of replica exchange algorithms[10−12] and Bennett's acceptance ratio[13,14] and through the development of new approaches for assessing when simulations have equilibrated.[15] We shall not discuss the different theories here but direct the reader to one of the recent excellent reviews.[16−18] Despite several comparative studies[11,19,20] it is not clear which techniques are the most accurate or even if such a general conclusion can be made. Whichever technique is chosen, for each value of ΔG one must typically run and analyze a number of similar molecular dynamics or Monte Carlo simulations. This not only requires access to a commodity cluster and preferably a supercomputer but also takes a long time and is a difficult and tedious task. For theoretically based calculations to complement experiment and be useful, one must be able to run a calculation in a reasonable time.[21] We use grid computing to accelerate a free energy calculation so that it takes less than a week, a period of time that we suggest is the upper limit for free energy calculations to be useful. The method (i.e., the underlying software and design) is general and may be adapted, if appropriate, to other free energy techniques.

* Corresponding author phone: +44 20 7679 4850; fax: ++44 20 7679 1501; e-mail: p.v.coveney@ucl.ac.uk.

‡ Department of Chemistry, University College London.

§ Birkbeck College, University of London.

|| Department of Biochemistry & Molecular Biology, University College London.

10.1021/ct6003017 CCC: $37.00
Published on Web 04/06/2007

In this paper we shall describe how we have reduced the time taken to run calculations using thermodynamic integration to less than a week by using high performance parallel and grid computing. This has allowed us to study in detail how phosphopeptides bind to the SH2 domain of the Src kinase, and we shall validate our results against published experimental data. Our method relies heavily on a collection of grid technologies, which we call STIMD (steered thermodynamic integration using molecular dynamics), that we have described in more detail elsewhere.[22]

We have chosen to use thermodynamic integration since it is an established and accepted free energy technique, the enthalpy ($\Delta\Delta H$) and entropy ($T\Delta\Delta S$) change can be calculated, a free energy component analysis can be performed, and last it requires many (10−50) simulations to be run for each value of $\Delta\Delta G$ and is therefore well-suited to deployment on a computational grid. The speedup achieved is simply due to the simultaneous running of all the λ-simulations; this can be described as a second tier of parallelization since each of the simulations is itself running on multiple computer processors. This is not novel since it must be done, for example, when using a replica exchange algorithm,[10−12] but we believe this is the first use of a computational grid to accurately and precisely calculate values of $\Delta\Delta G$.

There are two main benefits to the STIMD method: first it is easier for the scientist to run a calculation and second each calculation takes less time. Being able to quickly calculate a value of $\Delta\Delta G$ allows us either to repeat calculations to improve our confidence in the results or to study more than one mutation. Mistakes are less likely to be made during the calculation, and one can also investigate how changing the method affects the accuracy and precision of the calculations. These advantages will become apparent when we examine our results, but we shall first discuss why an understanding of the binding of phosphopeptides to the Src SH2 domain is both interesting and important.

2. Background

SH2 domains are small protein subunits. They bind to phosphorylated tyrosine residues and are key components of tyrosine kinase mediated signal cascades.[23] SH2 domains are considered by some to be the prototypical protein−protein interaction domain.[24] There are many different SH2 domains; for example more than 80 SH2 domains have been identified in the human genome.[25] Over the last 15 years there has been an intense academic and industrial effort aimed at developing small molecules that inhibit specific SH2 domains, including that of the Src kinase. The hope is that these drugs could be used to treat a diverse range of conditions, from cancer to osteoporosis to autoimmune diseases.[23,26,27] We shall refer to phosphorylated tyrosine residues using both the single (pY) and multiple (pTyr) amino acid residue notations, and the pTyr residue and the three C-terminal residues of the phosphopeptide are conventionally labeled +0, +1, +2, and +3. We shall also adopt the standard SH2 structural nomenclature established by Waksman et al.[28] to identify specific residues (e.g., Tyr βD5) or elements of secondary structure (e.g., the βD β-sheet).

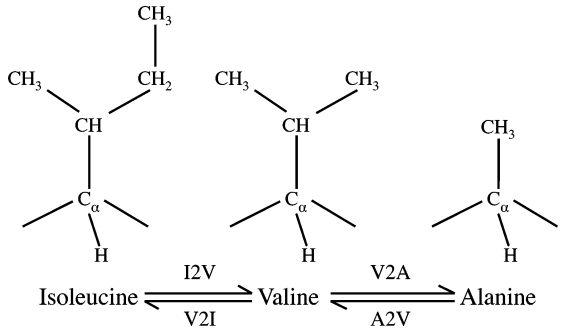

Figure 1. The simple hierarchy of aliphatic amino acid residues studied. The alchemical mutations are labeled using single-letter amino acid notation. For example the valine to alanine mutation is labeled V2A.

The X-ray crystallographic structure of the Src SH2 domain revealed that it is a compact protein domain composed of a central β-sheet surrounded by two α-helices.[29] The regions between the α-helices and the central β-sheet form two sockets; the first engages the pTyr residue, while the second binds the amino acid residue at the +3 position. A "two-pronged plug two-holed socket" model has been proposed to explain the binding of phosphopeptides to the Src SH2 domain.[28] The second of these sockets preferentially binds hydrophobic amino acid residues and is dubbed the "+3 binding pocket". Other structural and biophysical studies have indicated that only the pTyr residue and the +1, +2, and +3 residues contribute significantly to the binding event.[28,30,31] Subsequent isothermal titration calorimetry (ITC) experiments demonstrated that the pTyr residue accounts for about half of the binding free energy,[32] and therefore SH2 domains bind a wide range of phosphopeptides.

In this paper we shall investigate how progressively shortening the length of the aliphatic side chain of the amino acid at the +3 position affects the binding of phosphopeptides to the Src SH2 domain. This will help us understand how phosphopeptides engage with the +3 binding pocket and therefore what form an artificial inhibitor might take. We shorten the side chain by mutating the phosphopeptide from PQpYEEIPI to PQpYEEVPI to PQpYEEAPI (see Figure 1). For each of the two mutations, isoleucine to valine (I2V) and valine to alanine (V2A), the differences in binding free energies, $\Delta\Delta G$, entropies, $T\Delta\Delta S$, and enthalpies, $\Delta\Delta H$, are calculated. Values are also calculated for the reverse mutations (A2V, V2I) to provide a further estimate of the error. These values are then compared to two sets of experimental values which have been measured previously using isothermal titration calorimetry.[33,34]

The magnitude of the estimated error must be small if we are to both distinguish between the two mutations and detect any effect on the precision of our calculations when the method is altered. The magnitudes of the errors for the binding free enthalpies and entropies are at least an order of magnitude larger than the error for the binding free energy[35,36] and hence requiring that the magnitude of the errors in the values of $\Delta\Delta H$ and $T\Delta\Delta S$ be reasonably small will ensure that the magnitude of the error in the value of $\Delta\Delta G$ is smaller still. These criteria ensure that, although the mutations are

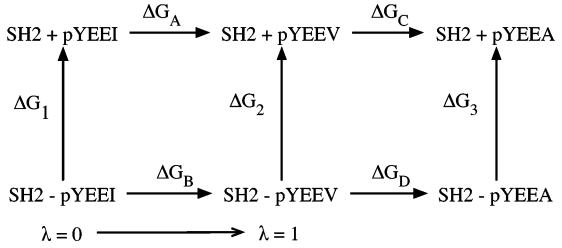

Figure 2. The combined thermodynamic cycle for the V2A and I2V mutations studied here. A '+' indicates that the two species are bound, and a '−' indicates that they are unbound. The progress parameter, λ, is shown for the cycle on the left.

themselves comparatively small, this is a challenging problem, and one that is a good test of our method.

SH2 domains have been studied using a range of different computational free energy approaches.[22,37−42] Woo and Roux[42] recently described an interesting approach to calculating ΔG using a combined potential of mean force and free energy perturbation approach. They applied it to the binding of a pYEEI phosphopeptide to the Lck SH2 domain and obtained reasonable agreement with experimental data, although the magnitude of the error is relatively large (2−3 kcal/mol). To minimize our error we shall limit ourselves to calculating how the binding free energy changes when a single amino acid is mutated (i.e., $\Delta\Delta G$); if this is successful, then we can consider calculating a value of ΔG for a phosphopeptide, either by using a conventional "double-decoupling" approach or by accelerating a newer free energy technique using our grid-based method.

3. Method

We shall first outline in section 3.1 the theory of thermodynamic integration and discuss the use of thermodynamic cycles. We shall describe in section 3.2 how we accelerate thermodynamic integration calculations. Finally, we shall explain in section 3.3 how the individual simulation models were constructed and what simulation protocols were used.

3.1. Thermodynamic Integration. We shall calculate the difference in the binding free energy, $\Delta\Delta G$, using the thermodynamic cycle drawn in Figure 2. The vertical paths denote the binding of the different phosphopeptides to the Src SH2 domain, while the horizontal paths denote the change in free energy when the amino acid residue at the +3 position is mutated from, for example, isoleucine to valine. We shall refer to these last free energies as *alchemical* since one molecule is transmuted, or mutated, into another. A parameter, λ, describes the progress between the physical states ($0 \le \lambda \le 1$): when $\lambda = 0$ the phosphopeptide is pYEEI and when $\lambda = 1$ it is pYEEV, but when $\lambda = 0.5$ it is in an intermediate state.

Since the free energy is a state function, we can write the difference in binding free energy, $\Delta\Delta G$, in terms of the alchemical free energies, $\Delta G'$, which converge more quickly. This leads to the following relationships

where ΔG_A and ΔG_C are the alchemical free energies when

$$\Delta\Delta G_{I2V} = \Delta G_A - \Delta G_B$$

$$\Delta\Delta G_{I2V} = \Delta G_C - \Delta G_D$$

the phosphopeptide is bound to the SH2 domain, and ΔG_B and ΔG_D are the alchemical free energies when the phosphopeptide is solvated (see Figure 2).

Thermodynamic integration[43] is "theoretically exact" in the classical approximation, although in practice approximations are introduced via the method, for example the use of a classical forcefield to describe the potential. The method also assumes that one can adequately sample phase space. The alchemical free energy, $\Delta G'$, is given by

$$\Delta G' = \int_0^1 \left\langle \frac{\partial U(\mathbf{p},\mathbf{r},\lambda)}{\partial\lambda} \right\rangle_\lambda d\lambda \qquad (1)$$

where $\mathbf{p}$ and $\mathbf{r}$ are the canonical momenta and positions of all the atoms, λ is the progress parameter as defined above, $U(\mathbf{p},\mathbf{r},\lambda)$ is the potential energy of the microstate, and the angled brackets indicate that an ensemble average is to be taken at the specified value of λ. We chose to generate the Boltzmann-factor weighted ensemble of structures using molecular dynamics (MD) since a Monte Carlo approach would be less efficient in this case due to the flexibility of the unbound phosphopeptide. The alchemical entropy, $T\Delta S'$, is calculated via

$$T\Delta S' = \frac{1}{k_B T}\int \left(\left\langle \frac{\partial U}{\partial\lambda} \right\rangle_\lambda \langle U \rangle_\lambda - \left\langle \frac{\partial U}{\partial\lambda} U \right\rangle_\lambda \right) d\lambda \qquad (2)$$

where T is the absolute temperature, and k_B is Boltzmann's constant.[35,44] The presence of the potential energy terms, U, in eq 2 leads to the estimated error for $T\Delta S'$ being 1−2 orders of magnitude larger[35,36] than that for $\Delta G'$. The alchemical enthalpy, $\Delta H'$, is then inferred via $\Delta G = \Delta H - T\Delta S$.

If the potential energy function, U, in eq 1 is pairwise additive, then we can rewrite U as a sum, $U_1 + U_2$..., where we have divided the atoms into groups. This allows us to rewrite eq 1 as a sum of integrals

$$\Delta G' = \int_0^1 \left\langle \frac{\partial U_1(\mathbf{p},\mathbf{r},\lambda)}{\partial\lambda} \right\rangle_\lambda d\lambda + \int_0^1 \left\langle \frac{\partial U_2(\mathbf{p},\mathbf{r},\lambda)}{\partial\lambda} \right\rangle_\lambda d\lambda + \ldots$$

and therefore we can also write the free energy as a sum, $\Delta G = \Delta G_1 + \Delta G_2$.... These components are not state functions, and hence their values are dependent upon the path taken.[45] This approach is known as free energy component analysis and is a useful method for estimating which groups of atoms contribute significantly to a binding event.[46,47] The results, however, must be carefully interpreted due to their path dependence.[48]

3.2. The STIMD Method. To calculate an alchemical free energy using eq 1 one must measure the ensemble average of $(\partial U/\partial\lambda)$ at several different values of λ and then numerically integrate the function to yield $\Delta G'$. We shall refer to this set of simulations used to compute the ensemble averages as "λ-simulations". It is customary to run these λ-simulations sequentially, i.e., the last conformation of the nth λ simulation is used as the initial conformation of the $(n+1)$th λ simulation. We speed up the calculation by simply running all of the λ simulations simultaneously. It is usually difficult to apply this approach to the study of protein mutations using MD for two reasons. First, the number of computer proces-

sors required to efficiently and quickly run an entire calculation is very large even for a small protein like the SH2 domain. Second, it is usually assumed that the accuracy of the calculation is improved by allowing the simulation unit cell to equilibrate at each value of λ before starting the next λ simulation. The first problem is addressed by our use of high performance and grid computing; we shall test the second assumption later. We also discuss and define grid computing[49] in detail elsewhere.[50,51]

The key step in our method was the interfacing of our chosen classical molecular dynamics package, NAMD2.5,[52,53] with the RealityGrid steering library.[54] This piece of grid middleware was developed as part of the U.K. e-Science pilot project RealityGrid and is freely available for download via that project's Web site.[55] Once interfaced with the steering library, an application can be remotely monitored and restarted, checkpoints can be taken, and variables can be steered.[54] The remote launching functionality is implemented using the Globus Toolkit 2[56] (GT2). This grid infrastructure is described in more detail by Fowler et al.[22]

The scientist uses the launcher, a local application, to launch each λ simulation onto the computational grid. This grid may consist of several high performance computers or it may consist of only one depending on circumstance or demand. Another local application, the steerer, which also exists in personal digital assistant[57] and Web-based forms, is then used to remotely monitor and control one or more λ simulations. If the simulations are not being run simultaneously, then a checkpoint is taken at the appropriate time and used to seed the next λ simulation. If this new simulation is on a different computer, then the relevant simulation files are automatically copied to the new machine. In summary, the scientist only interacts with two local applications, and they do not need to log in to the remote machines that form the computational grid. Of course, the same simulations may still be launched in the conventional way, but it is our experience that it is easy to make mistakes while keeping track of tens of simulations running on several different computers. Several STIMD calculations have been performed which combined the U.S. TeraGrid[58] and the U.K. National Grid Service[59] to create a single large federated intercontinental computational grid. We used six of the supercomputers in this computational grid, and we note and welcome the recent trend toward providing interactive access to high performance computers as this is necessary for methods like ours.[60]

One major disadvantage of this method is that one must install and maintain several different pieces of software, including GT2, on the local computer from where all the simulations are launched and controlled. At the time of writing we have adopted the Application Hosting Environment[61] (AHE), a set of tools that has been developed in our group, which only requires a local installation of Java 1.4.2. This makes it much easier for new users to start launching jobs onto a computational grid. The AHE is freely available for download from the U.K. Open Middleware Institute.[62] It does not, at present, allow a user to steer the variables of a remote simulation, but this is not necessary for running a thermodynamic integration on a computational grid. Unlike our original method the AHE automatically retrieves the simulation files from the remote computational resource and stores them in a local folder. One can launch and monitor simulations either through a graphical user interface or using command line tools. The former method is easier, but the latter is more appropriate when one is launching several jobs at once. Neither the RealityGrid steering library-based method or the AHE helps the scientist create the initial files required for a free energy simulation. This unfortunately remains a tedious and often error-prone task.

Table 1. Sizes of the Bound and Unbound Src SH2 Protein Complexes Studied[a]

mutation	system	waters	ions	atoms
Ile to Val	bound	6621	4 Cl^-	21 676
	unbound	2333	3 Na^+	7153
Val to Ala	bound	6621	4 Cl^-	21 673
	unbound	2286	3 Na^+	7009

[a] The reverse mutations are identical in size to the forward mutations.

3.3. Details of the Molecular Dynamics Simulations. We shall now describe how the simulation cells were constructed and how individual molecular dynamics simulations were run. The free energy calculations were performed with NAMD2.5[52,53] using the dual topology method to construct the alchemical atoms.[63] Only those atoms with different forcefield parameters were duplicated in the dual topology alchemical residues. This means that for the V2A alchemical residue the side chain is split at the C_α atom, whereas for the I2V alchemical residue it is split at the C_β atom. The Alchemify program was used to turn off the nonbonded interactions between vanishing and appearing atoms.[64]

All the initial structures were derived from the X-ray crystallographic structure of the octapeptide PQpYEEIPI bound to the Src SH2 domain.[28] No crystallographic waters were included. We chose to use explicit water since both the alchemical residue in the unbound system and the phosphopeptide-protein interface are exposed to solvent. Each structure was neutralized with NaCl counterions and was placed in a box of TIP3P water[65] such that the minimum distance from the protein or peptide to the surface of the box was 12 Å. The CHARMM27 forcefield,[66] which includes parameters for the pTyr residue,[67] was used. The sizes of the unbound and bound systems are listed in Table 1.

The potential energy of the system was first minimized for 500 steps. The positions of the protein heavy atoms were then fixed, and the hydrogen atoms and water molecules were allowed to evolve for 200 ps at a temperature of 298 K. This is the same temperature as the isothermal titration calorimetry experiments.[33,34] The positions of the backbone atoms of the protein and the phosphopeptide were then restrained by a small harmonic potential with a spring constant of 2 kcal mol^{-1} $Å^{-2}$. Next the entire system was warmed from 108 to 298 K in steps of 10 K over 200 ps. The magnitude of the restraining spring constant was then reduced to 0.5 kcal mol^{-1} $Å^{-2}$, and the system was evolved for 10 ps with the temperature held at 298 K by a Langevin thermostat. The restraints were then removed, and the system was evolved for a further 10 ps. Finally, a Berendsen barostat was applied to maintain the pressure at 1.01 kPa, and the simulation was

Table 2. The Seven Calculations Performed[a]

mutation	name	Δt (fs)	α	direction
Val to Ala	V2Aa	1	0.17	forward
	V2Ab	2	0.17	forward
	V2Ac	2	0	forward
	V2Ad	2	1	forward
	A2V	2	0.17	reverse
Ile to Val	I2V	2	0.17	forward
	V2I	2	0.17	reverse

[a] We shall refer to each calculation by the name given here (e.g., V2Aa). To allow an integration time step, Δt, of 2 fs the length of all bonds that include a hydrogen atom are constrained using the SHAKE[68] and SETTLE[69] algorithms. No constraints are required for $\Delta t = 1$ fs. α is defined to be the proportion of the nth λ simulation that has elapsed before the (n+1)th λ simulation is started. For example $\alpha = 1$ corresponds to running the λ simulations sequentially and $\alpha = 0$ to launching all the λ simulations simultaneously.

equilibrated for at least 2 ns in the isothermal−isobaric (NPT) ensemble at the smallest value of λ.

SHAKE[68] and SETTLE[69] algorithms were applied to allow a 2 fs integration time step. The particle mesh Ewald approach[70] was used to compute the electrostatic forces, and the van der Waals forces were cut off at 12 Å with a switching distance of 10 Å. NAMD does not have a softcore or separation-shifted van der Waals potential,[71] and therefore we used the conventional Lennard-Jones 6−12 van der Waals potential. This leads to the so-called "end-point catastrophe" as atoms vanish. To minimize this effect the 17 values of λ were nonuniformly spaced: $\lambda = 0.02$, 0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.84, 0.88, 0.91, 0.93, 0.94, ..., 0.99. We shall discuss the problems introduced by the use of a hardcore van der Waals potential later.

Each λ simulation was run for a duration of 2.5 ns making a total of 85 ns per calculation requiring 11.5k CPU hours to run. Due to the lack of continuity in the Lennard-Jones 6−12 form of the van der Waals potential it is not possible to run simulations at the endpoints, $\lambda = 0$ and $\lambda = 1$. These values of $(\partial U/\partial\lambda)$ were extrapolated by fitting functions to the electrostatic and van der Waals components as described elsewhere.[46,47]

The next step was to estimate the correlation and equilibration times. Computing the statistical inefficiency[15] provides a method of determining the correlation time, τ. The statistical inefficiency of both the $\Delta G'$ and λ data sets indicated that $\tau \sim 4-10$ ps (data not shown) and therefore the data sets were split into 250 bins of width $\tau = 10$ ps. A single value of $\Delta\Delta G$ or $T\Delta\Delta S$ was calculated from each bin, and errors were estimated using the batch averaging method.[72] In an attempt to determine the equilibration time, we applied a recently suggested method of testing for normality temporally reversed $\Delta G'$ or λ data sets.[15] This was not conclusive (data not shown), most likely because of the complexity of the mutations[22] compared to the simpler models the technique was validated on.

4. Results

We performed seven different calculations; these are listed in Table 2. We shall first determine if the calculations are equilibrated and converged before comparing the results to experimental data. Then we shall evaluate the effect of altering the length of the integration time step and the value of α (see Table 2 for a definition). Last we shall investigate which parts of the system are important in the binding of the pYEEI phosphopeptide to the Src SH2 domain using a free energy component analysis.

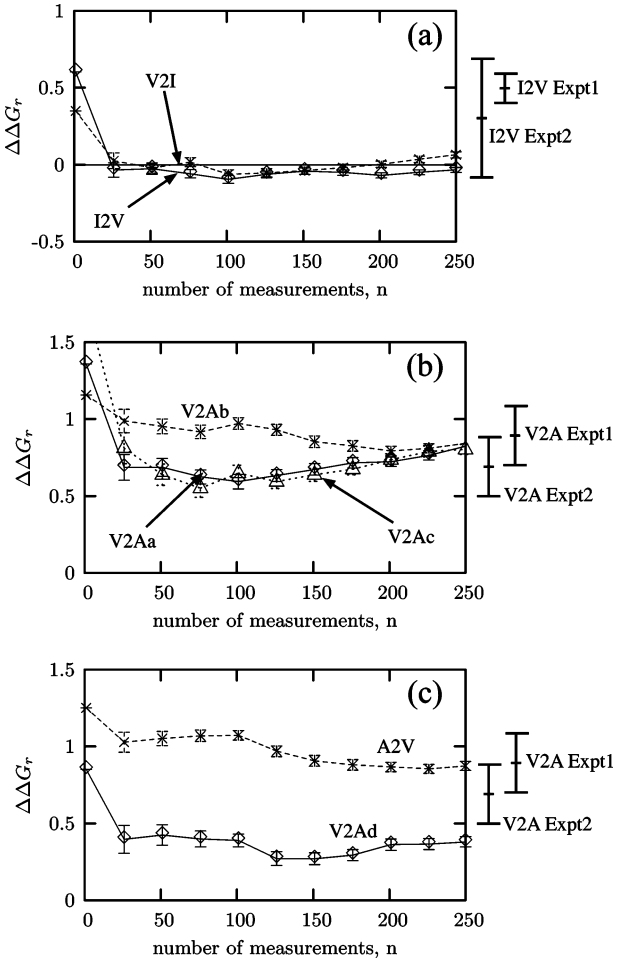

Figure 3. The differences in the binding free energies are not sensitive to the amount of data included in their calculation. This is shown by the reverse cumulative averages, $\Delta\Delta G_r$, of the isoleucine to valine (I2V) and valine to alanine (V2A) mutations. The results from the reversed mutations, e.g., valine to alanine (A2V), have been multiplied by (−1) to allow direct comparison with the forward mutations. The experimental values, including their error ranges, are also plotted. All energies are in kcal/mol.

The λ simulations will probably contain some unequilibrated data, which, if included in our calculations, will reduce the accuracy and precision of the calculated value of $\Delta\Delta G$. It is therefore important that we assess whether all the λ simulations are both equilibrated and converged. Here we adopt a suggestion made by Yang et al.[15] and compute the reverse cumulative average of $\Delta\Delta G$. This is plotted in Figure 3. The values of $\Delta\Delta G$ are not very sensitive to the amount of data included in their calculation, and furthermore all values (with the exception of V2Ad) converge during the first 0.5 ns to within the error of the experimental values. This indicates that, for these mutations, either the unequilibrated regions have little effect on the calculated values of $\Delta\Delta G$ or the equilibration time is much longer than 2.5 ns,

Table 3. Calculated and Experimentally Measured Thermodynamic Variables for the V2A and I2V Mutations[a]

calculation	$\Delta\Delta G$	$\Delta\Delta H$	$T\Delta\Delta S$
I2V Expt 1	0.5 ± 0.1	2.3 ± 0.4	1.8 ± 0.4
I2V Expt 2	0.3 ± 0.4	-1.1 ± 0.6	-1.4 ± 0.7
I2V	0.0 ± 0.1	1.5 ± 0.9	1.5 ± 0.9
V2I	0.1 ± 0.1	-0.7 ± 1.0	-0.7 ± 1.0
V2A Expt 1	0.9 ± 0.2	0.3 ± 0.4	-0.6 ± 0.4
V2A Expt 2	0.7 ± 0.2	0.9 ± 0.4	-0.2 ± 0.4
V2Aa	0.8 ± 0.1	-1.0 ± 1.3	-1.8 ± 1.3
V2Ab	0.8 ± 0.1	-1.2 ± 1.7	-2.0 ± 1.7
V2Ac	0.8 ± 0.1	0.3 ± 1.4	-0.5 ± 1.4
V2Ad	0.4 ± 0.1	0.2 ± 1.3	-0.2 ± 1.3
A2V	-0.9 ± 0.1	-2.0 ± 1.3	-1.1 ± 1.3

[a] The experimental results are taken from two separate studies, which we shall label Expt1[33] and Expt2.[34] See Table 2 for an explanation of the different mutations. All values are quoted with a precision of 0.1 kcal/mol, and all energies are in kcal/mol.

and hence we are unable to resolve any time-dependent behavior. We shall assume the former is true and will include all the data in our calculations. Interestingly, the I2V calculations converge more rapidly which we would expect given they have fewer alchemical degrees of freedom. Likewise, examining the reverse cumulative averages of the alchemical free energies shows that $\Delta G'$ for the bound systems converges more slowly than $\Delta G'$ for the unbound system (data not shown). This reflects the relatively slow rearrangements of the phosphopeptide and the protein around the alchemical residue compared to the faster rearrangements of the water and the phosphopeptide.

Values of $\Delta\Delta G$, $\Delta\Delta H$, and $T\Delta\Delta S$ were calculated using the entire trajectories for each of the seven calculations; these are given in Table 3 and are plotted in Figure 4. For comparison two sets of results from isothermal titration calorimetry experiments[33,34] are also given. The errors estimated for the calculated $\Delta\Delta G$ values (see the Method section) are smaller than the estimated experimental errors, and there is also good agreement between the forward and reverse mutations. All the calculated $\Delta\Delta G$ values agree with the experimental results, to within experimental error. As expected, the estimated errors for our calculated values of $T\Delta\Delta S$ and $\Delta\Delta H$ are more than an order of magnitude larger than those of $\Delta\Delta G$, and therefore while our calculated values of $T\Delta\Delta S$ and $\Delta\Delta H$ are accurate, they are not precise enough to be useful. We estimate that reducing the estimated error to under 0.5 kcal/mol would require all the λ simulations to be extended in duration by an order of magnitude. This demonstrates the importance of developing new techniques that improve sampling.

We shall now compare the different calculated values of $\Delta\Delta G$ for the V2A mutation to determine the effect of altering our method as described in Table 2. Since there is no measurable difference between the values of $\Delta\Delta G$ for the V2Aa and V2Ab mutations, we conclude that using SHAKE and SETTLE to restrain the lengths of the hydrogen to heavy atom bonds has no effect on the accuracy or precision of our calculations. This is useful as it allows us to speed up our calculations by using an integration time step of 2 fs. Our method fundamentally relies on there being no effect

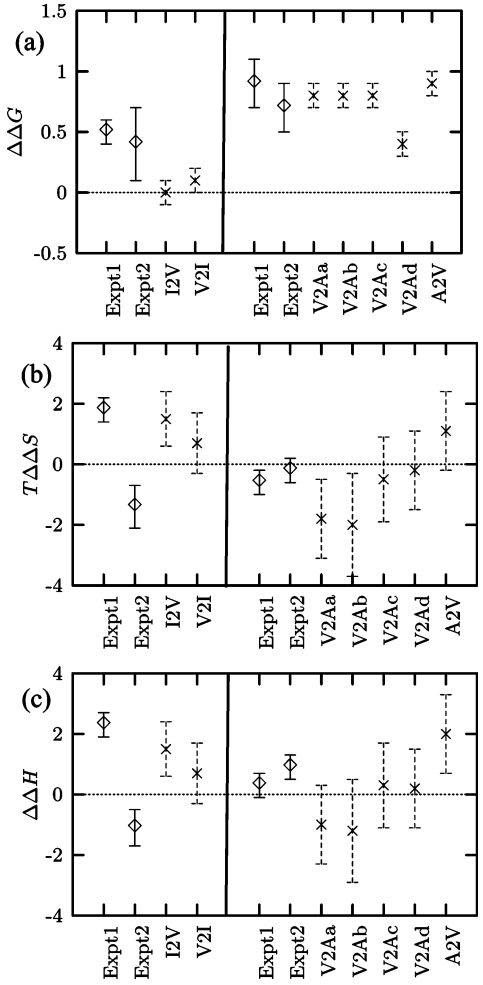

Figure 4. There is good agreement between our calculated values of $\Delta\Delta G$ and the experimental values (a). There is reasonable agreement between our calculated values of $T\Delta\Delta S$ (b) and $\Delta\Delta H$ (c) and the experimental values. The results from the reversed mutations, e.g., valine to alanine (A2V), have been multiplied by (-1) to allow direct comparison with the forward mutations. All energies are in kcal/mol.

on the accuracy of our calculations if we launch all the λ simulations simultaneously ($\alpha = 0$) or nearly simultaneously (e.g., $\alpha = 0.17$) compared to launching the λ simulations consecutively. To test this assumption we ran three free energy calculations (V2Ac,b,d) with values of $\alpha = 0$, 0.17, and 1.

The values of $\Delta\Delta G$ when $\alpha = 0$ and 0.17 (V2Ab,c) agree with experiment. This suggests that, for this mutation at least, there is no effect on the accuracy of the calculation if we launch all the λ simulations simultaneously. Our computed value of $\Delta\Delta G$ when $\alpha = 1$ (V2Ad) does not agree with the experimental values, although this is not conclusive given the magnitude of the error bars. Launching all the λ simulations consecutively may therefore affect the accuracy of the calculation. This could be due to error accumulating in the λ simulations and dominating any reduction in the equilibration time achieved by launching the λ simulations consecutively.

We examined the relative free energy profiles of all the mutations (see the Supporting Information) to investigate these effects in more detail. There are no large changes with

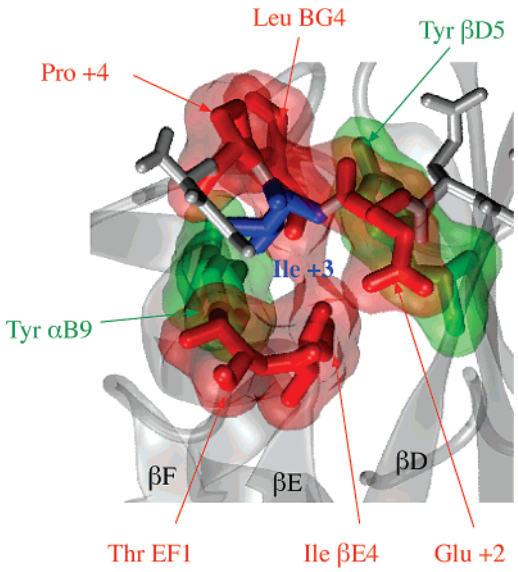

Figure 5. A schematic representation showing which amino acid residues of the Src SH2 domain favor (in green) or hinder (in red) the binding of isoleucine over alanine in the +3 binding pocket. Not indicated are the effects of the remainder of the protein, the solvent, and the alchemical residue itself which all favor the binding of isoleucine. The alchemical residue at the +3 position is drawn in dark blue, and the remainder of the phosphopeptide and the protein domain are drawn in gray.

λ in the relative free energy profiles for all of the V2A mutations. This indicates that, to a first approximation, all of the λ windows contribute a similar amount to the final value of $\Delta\Delta G$. In addition the shapes of all the profiles for the V2A mutations are similar with any differences being gradually accumulated, and therefore we cannot attribute the loss of accuracy for the V2Ad mutation to a single λ window.

Finally, we shall present the results of a free energy component analysis for all the calculations with $\alpha = 0.17$ and a time step of 2 fs (I2V, V2I, V2Ab, and A2V). All the atoms were assigned to one of 23 groups. The majority of these were amino acid residues, but larger groups, for example the entire solvent, were also defined (see the Supporting Information). The values of the free energy components are path-dependent and therefore should not be regarded as quantitative. They may, however, indicate whether or not a group of atoms, for example a residue, favors one of the mutated side chains over the other. The free energy components for the forward and reverse calculations were first averaged, and the results for the I2V and V2A mutations were added to create a single data set for the isoleucine to alanine (I2A) mutation. We assume that any group with a nonzero free energy component plays a role in the binding of the phosphopeptide. The results are shown qualitatively in Figure 5.

According to our analysis the solvent, the alchemical residue itself, the remainder of the Src SH2 domain, and Tyr βD5 and Tyr αB9 all favor the binding of the longer side chain of isoleucine. Three of the residues that surround the +3 binding pocket, Ile βE4, Thr EF1, and Leu BG4, in addition to the amino acid residues in the +2 and +4 positions, all favor the binding of the shorter side chain of alanine. The strength with which the aliphatic side chain engages the +3 binding pocket is therefore found to be a balance between favorable interactions with the two tyrosines deep within the pocket and repulsive steric clashes with the adjacent amino acid residues on the phosphopeptide and those surrounding the pocket.

5. Discussion

We have rapidly, accurately, and precisely calculated how the binding free energy changes as we mutate the +3 residue of the pYEEI phosphopeptide from isoleucine to valine to alanine. This was done using a grid-based method that we call STIMD. The development of such methods is necessary if we are both to subject the growing number of competing free energy theories to broader comparative studies and to use these techniques in a more routine way. We have also calculated values of $\Delta\Delta H$ and $T\Delta\Delta S$ for both mutations; however, while the values agreed with known experimental values, the calculated errors were too large to be useful. Being able to run thermodynamic integration calculations quickly and with relative ease allowed us to run forward and reverse calculations, test and optimize our method, and examine two different mutations. We found that either accelerating the calculation by using an integration time step of 2 fs or launching all the λ simulations simultaneously did not affect the accuracy or precision of $\Delta\Delta G$. This result is crucial as our method relies on being able to launch all the λ simulations simultaneously, and hence in future we will be able to more quickly calculate changes in binding free energy for this system. Finally we presented the results of a free energy component analysis.

The exact information determined by a free energy component analysis cannot be determined experimentally. The strength of an interaction between two amino acid residues may be assessed by calculating the coupling free energy, ΔG_c, using double mutant cycles.[45,47] Comparing these results to existing experimental data provides a further route for validating our method, and where there is no experimental data or it is inconclusive, we can then make predictions about the influence individual amino acid residues have on the binding of the phosphopeptide. Coupling free energy studies confirm that the individual contributions of the studied amino acid residues are weak.[33,73] The magnitude of the experimental errors, however, are similar to the magnitude to the coupling free energies which makes comparison difficult. The experimental results suggest that Tyr βD5, Leu βG4, and Glu +2 favor the binding of isoleucine, whereas our findings indicate that Tyr βD5 favors, and both Leu βG4 and Glu +2 hinder, the binding of isoleucine. These conclusions are tentative given the magnitudes of the experimental errors, the lack of an exact correspondence between free energy components and ΔG_c, and theoretical concerns over the interpretation of free energy components. The effect of mutating Tyr αB9 has not, however, been studied experimentally, and we therefore predict that mutating it to alanine will reduce the specificity of the Src SH2 domain for the canonical phosphopeptide pYEEI. This would confirm the importance of the pocket defined by the aromatic rings of Tyr αB9 and Tyr βD5 in binding pYEEI.

Our free energy component analysis also predicted that the solvent favors the binding of the longer side chain of isoleucine. It is difficult for experiment to investigate the role of water in protein−protein interactions, and this, therefore, is an example of where the computational approach is potentially useful. We hypothesize that the complex network of water-mediated hydrogen bonds that forms between the backbone of the alchemical atom and Lys βD6 and Arg βD′1 (see the Supporting Information) is disrupted by the mutation. Interestingly, despite including no crystallographic waters in our simulations, a water molecule is observed hydrogen bonding to the backbone of the +3 amino acid residue for the entire duration of several λ simulations. This water molecule is also resolved in several X-ray crystal structures,[28,74] while similar water-mediated hydrogen-bonded networks have been observed by experiment[75] and computational study.[39] We hypothesize that, through these bridging waters, Lys βD6 and Arg βD′1 contribute to the binding of aliphatic side chains in the +3 binding pocket. This is supported by a bioinformatics study[76] which indicates that both these residues contribute to the selectivity of Src SH2 domains despite being ∼6 Å away from the phosphopeptide. We predict that mutating these residues will also reduce the specificity of the Src SH2 domain for pYEEI.

Using a free energy component analysis we have identified which residues hinder and which encourage the binding of the longer side chain of isoleucine in the +3 binding pocket. The "two-pronged plug two-holed socket" model simply characterizes the +3 binding pocket as hydrophobic and therefore implies that all the residues lining the pocket encourage the binding of hydrophobic side chains. Our results therefore constitute an improvement on this model and may be tested by combined mutagenesis and isothermal titration calorimetry experiments. Improving our understanding of how SH2 domains bind phosphopeptides will help resolve whether specific SH2 domains can be inhibited by druglike molecules[27] and whether these will be useful, given the potential for crosstalk between different SH2 domains within tyrosine kinase mediated signal cascades.[24]

It should be recognized that there are limitations in a study of this nature. First, we used a standard Lennard Jones 6−12, or "hardcore", van der Waals potential, and therefore, due to the lack of continuity in $(\partial U/\partial\lambda)$, we could not run simulations with $\lambda = 0$ or 1 but instead had to extrapolate the values of $(\partial U/\partial\lambda)$ at these points. This is a source of error. Furthermore, if we had been able to use a softcore or "separation-shifted" van der Waals potential,[71] then not only would we have avoided this problem but also the λ simulations would have converged significantly faster. Although there are molecular dynamics codes that have softcore van der Waals potentials implemented (e.g., CHARMM), these do not currently scale very well on parallel computers. Instead we chose to use NAMD which scales very well on parallel computers but does not have a softcore potential. Using NAMD allowed us to run simulations of comparatively long durations in a short space of time, and we hope that studies of this nature will encourage the implementation of a softcore potential in codes such as NAMD. Finally, the accuracy of free energy calculations also relies on the phase space being adequately sampled. Applying a replica exchange algorithm[10−12] to all the λ simulations running on the computational grid would improve the sampling.

Acknowledgment. We thank Dr. Shunzhou Wan and our other RealityGrid collaborators, in particular the Supercomputing, Visualisation and e-Science group at Manchester University. We are also grateful for the computational time provided by the U.K. National Grid Service. This work was supported by EPSRC via RealityGrid grant GR/R67699. Access to the U.S. TeraGrid was provided by the National Science Foundation under NRAC grant MCA04N014. Use of UKLight was supported by the Exploitation of Switched Lightpaths for E-Science Applications PPARC/EPSRC/MRC project, grant number GR/T04465/01. EPSRC also provided a Ph.D. studentship for P.W.F.

Supporting Information Available: Relative free energy profiles for the V2A and I2V mutations (Figure 1), which groups were used for the free energy component analysis (Figure 2), an example of the water-mediated hydrogen-bonded network (Figure 3), and raw numerical results. This material is available free of charge via the Internet at http://pubs.acs.org

References

(1) Åqvist, J.; Medina, C.; Samuelsson, J.-E. *Protein. Eng.* **1994**, *7*, 385−391.

(2) Jarzynski, C. *Phys. Rev. E* **1997**, *56*, 5018−5035.

(3) Ytreberg, F. M.; Zuckerman, D. M. *J. Chem. Phys.* **2004**, *120*, 10876−10879.

(4) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9409.

(5) Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. *Proteins* **1998**, *32*, 399−413.

(6) Lee, M. S.; Olson, M. A. *Biophys. J.* **2006**, *90*, 864−877.

(7) Damodaran, K. V.; Banba, S.; Brooks, C. L., III *J. Phys. Chem. B* **2001**, *105*, 9316−9322.

(8) Hénin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904−3004.

(9) Rodinger, T.; Howell, P. L.; Pomès, R. *J. Chem. Phys.* **2005**, *123*, 034104.

(10) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703−13710.

(11) Rodinger, T.; Howell, P. L.; Pomès, R. *J. Chem. Theor. Comput.* **2006**, *2*, 725−731.

(12) Rick, S. W. *J. Chem. Theor. Comput.* **2006**, *2*, 939−946.

(13) Lu, N.; Singh, J. K.; Kofke, D. A. *J. Chem. Phys.* **2003**, *118*, 2977−2984.

(14) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 144107.

(15) Yang, W.; Bitetti-Putzer, R.; Karplus, M. *J. Chem. Phys.* **2004**, *120*, 2618−2628.

(16) Rodinger, T.; Pomès, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164−170.

(17) Kofke, D. A. *Fluid Phase Equilib.* **2005**, *228−229*, 41−48.

(18) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589−1615.

(19) Rodriguez-Gomez, D.; Garve, E.; Pohorille, A. *J. Chem. Phys.* **2004**, *120*, 2563−2578.

(20) Ytreberg, F. M.; Swendsen, R. H.; Zuckerman, D. M. *J. Chem. Phys.* **2006**, *125*, 184114.

(21) Chipot, C.; Pearlman, D. A. *Mol. Simul.* **2002**, *28*, 1−12.

(22) Fowler, P. W.; Jha, S.; Coveney, P. V. *Philos. Trans. R. Soc. London, Ser. A* **2005**, *363*, 1999−2015.

(23) Machida, K.; Mayer, B. J. *Biochim. Biophys. Acta* **2005**, *1747*, 1−25.

(24) Pawson, T.; Nash, P. *Science* **2003**, *300*, 445−452.

(25) Venter, J. C. et al. *Science* **2001**, *291*, 1304−1352.

(26) Waksman, G.; Kumaran, S.; Lubman, O. *Expert Rev. Mol. Med.* **2004**, *6*, 1−18.

(27) Shakespeare, W. C. *Curr. Opin. Chem. Biol.* **2001**, *5*, 409−415.

(28) Waksman, G.; Shoelson, S. E.; Pant, N.; Cowburn, D.; Kuriyan, J. *Cell* **1993**, *72*, 779−790.

(29) Waksman, G.; Kominos, D.; Robertson, S. C.; Pant, N.; Baltimore, D.; Birge, R. B.; Cowburn, D.; Hanafusa, H.; Mayer, B. J.; Overduin, M.; Resh, M. D.; Rios, C. B.; Silverman, L.; Kuriyan, J. *Nature* **1992**, *358*, 646−653.

(30) Songyang, Z. et al. *Cell* **1993**, *72*, 767−778.

(31) Eck, M. J.; Shoelson, S. E.; Harrison, S. C. *Nature* **1993**, *362*, 87−91.

(32) Bradshaw, J. M.; Mitaxov, V.; Waksman, G. *J. Mol. Biol.* **1999**, *293*, 971−985.

(33) Bradshaw, J. M.; Waksman, G. *Biochemistry* **1999**, *38*, 5147−5154.

(34) Bradshaw, J. M.; Grucza, R. A.; Ladbury, J. E.; Waksman, G. *Biochemistry* **1998**, *37*, 9083−9090.

(35) Wan, S.; Stote, R.; Karplus, M. *J. Chem. Phys.* **2004**, *121*, 9539.

(36) Pearlman, D. A.; Rao, B. G. Free energy calculations: Methods and Applications. In *Encyclopedia of Computational Chemistry;* Schleyer, P. v. R., Ed.; Wiley: Chichester, 1999; Vol. 2, pp 1036−1061.

(37) Henriques, D. A.; Ladbury, J. E.; Jackson, R. M. *Protein Sci.* **2000**, *9*, 1975−1985.

(38) Geroult, S.; Virdee, S.; Waksman, G. *Chem. Biol. Drug Des.* **2006**, *67*, 38−45.

(39) Price, D. L.; Jorgensen, W. L. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 681−695.

(40) Suenaga, A.; Hatakeyama, M.; Ichikawa, M.; Yu, X.; Futatsugi, N.; Narumi, T.; Fukui, K.; Terada, T.; Taiji, M.; Shirouzu, M.; Yokoyama, S.; Konagaya, A. *Biochemistry* **2003**, *42*, 5195−5200.

(41) Donnini, S.; Juffer, A. *J. Comput. Chem.* **2003**, *25*, 393−411.

(42) Woo, H.-J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825−6830.

(43) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300−313.

(44) Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. *J. Chem. Phys.* **2004**, *120*, 2652−2661.

(45) Mark, A. E.; van Gunsteren, W. F. *J. Mol. Biol.* **1994**, *240*, 167−176.

(46) Wan, S.; Coveney, P. V.; Flower, D. R. *J. Immunology* **2005**, *175*, 1715−1723.

(47) Michielin, O.; Karplus, M. *J. Mol. Biol.* **2002**, *324*, 547−569.

(48) Boresch, S.; Archontis, G.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 25−33.

(49) Foster, I.; Kesselman, C.; Tuecke, S. *Int. J. Supercomput. Appl.* **2001**, *15*, 200−224.

(50) Coveney, P. V.; Fowler, P. W. *J. R. Soc. Interface* **2005**, *2*, 267−280.

(51) Coveney, P. V., Ed.; Scientific Grid Computing. *Philos. Trans. R. Soc. London, Ser. A* **2005**, *363*(1833), 1705−2053.

(52) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadara, jan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283−312.

(53) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.

(54) Pickles, S. M.; Haines, R.; Pinning, R. L.; Porter, A. R. *Philos. Trans. R. Soc. London, Ser. A* **2005**, *363*, 1843−1853.

(55) RealityGrid. http://www.realitygrid.org (accessed Feb 23, 2007).

(56) The Globus Alliance. http://www.globus.org (accessed Feb 23, 2007).

(57) Kalawsky, R. S.; Nee, S. P.; Holmes, I.; Coveney, P. V. *Philos. Trans. R. Soc. London, Ser. A* **2005**, *363*, 1885−1894.

(58) The TeraGrid. http://www.teragrid.org (accessed Feb 23, 2007).

(59) The National Grid Service. http://www.ngs.ac.uk (accessed Feb 23, 2007).

(60) Coveney, P. V.; Jha, S.; Pickles, S. *CSAR Focus* **2005**, *14*, 16−18.

(61) Coveney, P. V.; Saksena, R. S.; Zasada, S. J.; McKeown, M.; Pickles, S. *Comput. Phys. Commun.* **2007**, *176*, 406−418.

(62) The Open Middleware Institute. http://www.omii.ac.uk (accessed Feb 23, 2007).

(63) Dixit, S. B.; Chipot, C. *J. Phys. Chem. A* **2001**, *105*, 9795−9799.

(64) Hénin, J. Alchemify, an X-PLOR PSF post-processor for alchemical free energy calculations in NAMD. http://www.edam.uhp-nancy.fr/Alchemify/ (accessed Feb 23, 2007).

(65) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(66) MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(67) Feng, M.-H.; Philippopoulos, M.; MacKerell, A. D. J.; Lim, C. *J. Am. Chem. Soc.* **1996**, *118*, 11265−11277.

(68) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(69) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952−962.

(70) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(71) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025−9031.

(72) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: London, 2002; p 529.

(73) Bradshaw, J. M.; Mitaxov, V.; Waksman, G. *J. Mol. Biol.* **2000**, *299*, 521−535.

(74) Lubman, O. Y.; Waksman, G. *J. Mol. Biol.* **2003**, *328*, 655−668.

(75) Chung, E.; Henriques, D. A.; Renozi, D.; Zvelebil, M.; Waksman, G.; Robinson, C. V.; Ladbury, J. E. *Structure* **1998**, *6*, 1141−1151.

(76) Sheinerman, Felix, B.; Al-Lazikani, B.; Honig, B. *J. Mol. Biol.* **2003**, *334*, 823−841.

CT6003017