CMO-BIRS workshop report

Free-Energy Calculations. A Mathematical Perspective (15w5128)

Introduction

Statistical mechanics provides the link between macroscopic observables and the microscopic dynamics of molecular systems. Central to this is the notion of free energy differences and related concepts such as potentials of mean force. Mathematically, these objects can be expressed in terms of probabilities and probability density functions, and they can be formulated as computations of expectations in a high dimensional space of microstates. The aim of the workshop is to examine computational methods for various types of free energy calculations, using mathematics to analyze and improve them. Following are representative calculations: determining free energy differences between two different states of a system, comparing free energy differences for two different systems (relative free energy differences), determining an effective energy function (potential of mean force) for a coarse-grained description of a system, and the use of Markov chains (Markov chain Monte Carlo methods and chaotic/stochastic dynamics) to generate samples from the target distribution.

Scientific presentations

Overview of the program

The five-day workshop consisted each day of an hour-long pedagogical lecture, reviewing the current state of the art, followed by shorter, half-hour research talks. On Monday, Michael Gilson (UCSD) gave a didactic lecture on statistical thermodynamics and computational approaches for the determination of binding free energies, enthalpies and entropies. On Tuesday, Ron Elber (University of Texas, Austin) gave a one-hour lecture on sampling of pathways, trajectories, and trajectory fragments to estimate kinetics and free energies. On Wednesday, Jérome Hénin (IBPC, Paris) delivered a general pedagogical lecture entitled Free energy isn't free: From intuition to computation, and back again. On Thursday, Gersende Fort (Télécom, Paristech) gave the last didactic talk on the mathematical aspects of adaptive samplers, with applications to free energy calculations.

Instead of dichotomizing in independent blocks full-fledged mathematics lectures and applied, chemistry and biophysics-oriented ones, which would be detrimental to scientific exchanges, the talks were alternated. On Monday, James Gumbart (GeorgiaTech) discussed practical issues in free-energy calculations illustrated through the the energetics of deca-alanine folding. Manuel Athènes (CEA, Paris) followed with a talk on the use of Bayes formula in free energy methods. Florent Calvo (UJF, Grenoble) tackled then the question of equilibrium shapes from non-equilibrium path-integral free-energy simulations, illustrated with alkali clusters interacting with a helium droplet. David Mobley (UCI) presented the current successes and challenges in calculating binding free energies from molecular simulations. Thomas Simonson (École Polytechnique, Paris) then discussed electrostatic free energies in infinite molecular systems and the implication of conditional convergence, which was followed by the lecture of Mahmoud Moradi (University of Arkansas) on the thermodynamic characterization of large-scale structural transitions using an iterative approach of free-energy calculations and path-finding algorithms. The day was concluded with a talk by Sunhwan Jo (Argonne) on the quantification of protein-protein binding energy and entropy using molecular dynamics simulations.

After the pedagogical lecture, the second day was pursued by the talk of Gabriel Stoltz (École des Ponts, ParisTech) on error estimates for the computation of transport coefficients. Andrew Pohorille (NASA Ames) then discussed free energies from non-equilibrium simulations, illustrated in the case of transmembrane ion transport. Jonathan Weare (University of Chicago) presented a stratification strategy of Markov processes for rare event simulations. Frédéric Legoll (École des Ponts, ParisTech) continued with reduced models for computing the dynamics of reaction coordinates. Next, Daniel Zuckerman (University of Pittsburgh) discussed the connection between free energy landscapes and kinetics without bias in the context of exact and approximate non-Markovian analyses. Robert Skeel (Purdue) presented some recent work on transition paths, while David Aristoff (Colorado State) proposed in the last lecture of the day a mathematical framework for exact milestoning.

On the third day, Wei Yang (University of Florida) delivered a talk on orthogonal space sampling of hierarchical energy landscapes for free energy calculations. Next, Yuko Okamoto (University of Nagoya) applied generalized-ensemble algorithms to calculations of ligand affinity. James Dama (University of Chicago) presented quasi-equilibrium methods utilized in the convergence analysis and improvement of metadynamics. Fabian Paul (Freie Universität Berlin) wrapped the day with a talk on transition-based reweighting analysis methods.

On the fourth day, Benjamin Jourdain (École des Ponts ParisTech) detailed an analysis of discrete space versions of the self-healing umbrella sampling and well-tempered metadynamics algorithms. This talk was followed by that of Antonietta Mira (Università della Svizzera Italiana) on reduced variance Monte Carlo for doubly intractable problems exploiting multi-core architectures. In his lecture, Ludovic Goudenège (ENS Cachan) presented numerical simulations for a generalization of adaptive multilevel splitting. Next, Ben Leimkuhler (University of Edinburgh) delivered a talk on enhanced sampling using extended stochastic dynamics, followed by a discussion of Alessandro Laio (SISSA) on multidimensional free energy landscapes from bias-exchange metadynamics. The question of multilevel splitting was explored again by Arnaud Guyader (Université Pierre et Marie Curie, Paris) in his talk ? About stochastic waves and adaptive multilevel splitting. Michael Shirts (University of Colorado) closed the session with a lecture on reweighting from the mixture distribution as a unifying formalism for carrying out and analyzing free energy calculations.

On the last day, Alessandro Rodriguez (Sissa) discussed the mapping of complex free energy landscapes by fast search-and-find of density peaks. Next, Jeffrey Comer (Kansas State University) delved into the topic of enhanced sampling through atomistic-coarse coupling. In a didactic lecture, Tony Lelièvre (École des Ponts ParisTech) presented the athematical foundations of accelerated dynamics algorithms. The session was wrapped by Chris Chipot (University of Illinois, Université de Lorraine) and his talk on the determination of membrane permeabilities from first principles.

We provide hereafter the summary of selected lectures presented during the week.

Successes and challenges in calculating binding free energies from molecular simulations

Drug discovery is a time-consuming and expensive process, and inflation-adjusted discovery costs appear to be growing roughly exponentially over time for a number of reasons. We seek to develop accurate tools for predicting binding affinities of small-molecule ligands to proteins. These would have a number of important applications, but a particular focus for us is early stage pharmaceutical drug discovery. While computational tools such as docking are already widely used in this area, these are limited in their ability to predict binding affinity. Previously, we found that tools with RMS errors in the 0.5-2 kcal/mol range could provide substantial benefits in early stage drug discovery, so one of our major goals has been to get free energy calculations based on molecular simulations to the point where they can routinely achieve this level of accuracy. To this end we have studied a series of model binding sites of increasing complexity, predicting binding free energies and then comparing with experiments done separately, typically achieving accuracies in the 1-2 kcal/mol range. In the process, however, we have also discovered particular sampling challenges relating to (even small) protein conformational changes and slow transitions between possible ligand binding modes. While we have been able to deal with these challenges via careful treatment of the relevant degrees of freedom, better solutions using enhanced sampling methods are still needed (though some progress has already been made). Still, alchemical free energy techniques such as those we employ already appear to have enough accuracy that they can aid drug discovery applications in suitable circumstances, so we have invested substantial effort in automating these techniques, helping to enable the first large-scale tests of these techniques on pharmaceutically-relevant targets.

Analysis of a discrete version of the Self-Healing Umbrella Sampling algorithm

The Self-Healing Umbrella Sampling (SHUS) algorithm is an adaptive biasing potential algorithm, which has been proposed in order to efficiently sample a multimodal probability measure. We analyze a discrete time and discrete space version of this algorithm. We show that this method can be seen as a variant of the well-known Wang-Landau algorithm. Adapting results on the convergence of the Wang-Landau algorithm obtained recently for a deterministic step-size sequence, we prove the convergence of the SHUS algorithm. We also compare the two methods in terms of efficiency. We finally propose a modification of the SHUS algorithm in order to increase its efficiency.

Electrostatic aspects in free energy calculations

Free energy simulations for electrostatic and charging processes in complex molecular systems encounter specific difficulties owing to the long-range, $1/r$ Coulomb interaction. To calculate the solvation free energy of a simple ion, it is essential to take into account the polarization of nearby solvent but also the electrostatic potential drop across the liquid-gas boundary, however distant. The latter does not exist in a simulation model based on periodic boundary conditions because there is no physical boundary to the system. An important consequence is that the reference value of the electrostatic potential is not an ion in vacuum. Also, in an infinite system, the electrostatic potential felt by a perturbing charge is conditionally convergent and dependent on the choice of computational conventions. Furthermore, with Ewald lattice summation and tinfoil conducting boundary conditions, the charges experience a spurious shift in the potential that depends on the details of the simulation system such as the volume fraction occupied by the solvent. All the aspects above can be handled with established computational protocols, as reviewed here and illustrated for several small ions and three solvated proteins. Importantly, for processes that conserve the total charge, like charge transfer from one molecule to another in solution, the computational methods commonly used to sum the potential series give the same result. This was shown explicitly for the small ions and a general proof given. Several important points were raised in the discussion, which will be addressed in another, upcoming review article. A connection to the Orthogonal Space Random Walk free energy method will be explicited.

Computation of transport coefficients

The laws of statistical physics give expressions for various transport coefficients, such as the mobility, thermal conductivity, the shear viscosity, etc. There are two main approaches to compute these coefficients: either approximate the Green-Kubo formula which expresses the transport coefficient as some integrated time correlation function, or compute the linear response of nonequilibrium systems under appropriate perturbations. These two approaches are equivalent at the theoretical level, but not from a practical/numerical viewpoint when the dynamics is integrated with a finite time-step. More precisely, (i) error estimates on the coefficients obtained with the linear response of nonequilibrium dynamics are of the same order as the errors on the invariant measure for equilibrium dynamics, which can be determined by rephrasing Talay-Tubaro estimates in an appropriate functional setting; (ii) the errors on the discretization of the Green-Kubo formula are determined by the error on the invariant measure upon slightly changing the functions appearing in the integrated correlation function. Taking the mobility as paradigmatic example, it is also possible to quantify the (extra) errors on the transport coefficients induced by the stabilization of a given numerical scheme by a Metropolis-Hastings procedure. In this situation, some moves are rejected, so that the time-correlation functions do not decay sufficiently fast. In fact, the errors can be reduced by considering appropriate numerical schemes in the proposal step of the Metropolis-Hastings algorithm, and possibly changing the Metropolis into a Barker rule, as proposed in recently. This allows in the most favorable cases to decrease the bias from

Δt to Δt^2, although the statistical error roughly increases by a factor of 2 when the Barker rule is used.

Conclusion

In conclusion, the workshop was an opportunity to bring a new community into the field of free energy calculations, namely researchers in computational statistics and specialists in Markov Chain Monte Carlo methods. Besides this, here are a examples of topics raised during the round table discussions: (i) the definition of a transition path: various approaches have been proposed (max flux, minimum free energy path, etc...) and it is still unclear what is a good definition depending on the objective (biasing the dynamics for numerical purposes or extracting important qualitative quantities such as the transition state); (ii) efficiency of adaptive biasing methods: adaptive biasing force methods based on interacting replicas are well understood, but other techniques are still challenging to analyze (for example the convergence and efficiency of metadynamics with a fixed Gaussian height and width)-this actually fostered a new research project following discussions between A. Laio and researchers from CERMICS ; (iii) optimization of force fields: how to use efficient optimization techniques to make the force fields more reliable (iv) molecular dynamics as qualitative or quantitative tool: do we believe that MD can be used to understand and predict mechanisms, or can we go beyond and measure precisely quantitative parameters (reaction rates, etc...) that can be compared to experimental data; (v) diffusion models on reduced corrdinates: there is a lack of understanding of the theoretical underpinning for deriving such models, and to assess the error associated with this coarse-graining procedure. Moreover, broadly speaking, the participants are now convinced that there is much to gain by intearcting more with computational statistics, in particular in the following fields: enhanced sampling techniques (non reversible perturbations, better time discretization techniques, etc....), rare event techniques and interacting replicas methods and machine learning techniques for the construction of empirical force fields.