Bayesian inference of conformational ensembles from limited structural data

Wojciech Potrzebowski

ESS, Lund, Sweden

The small-angle scattering (SAS) from proteins in solution samples the ensemble average of the randomly oriented structures, and ensemble modelling for proteins with flexible regions against SAS data is increasingly popular. However, the smooth SAS profile can typically be defined by as few as 10-15 points, and the ensemble model has many more degrees of freedom. Typically, a very large ensemble (10,000 or more) is generated within some constrained set, and a population weighted sub-set of structures is identified that predicts a profile that best-fits the data. Representative structures are selected based on clustering analysis to aid in visualizing the nature of the ensemble, but their accuracy and what minimal set is justified by the data are outstanding questions.

In this study we use a model evidence to select ensembles with an optimal number of members. Model evidence (ME) is widely used in Bayesian model comparison and provides an automatic Occam's razor effect by balancing between fit to data and model complexity, thereby providing a rigorous approach to combat overfitting. However, ME is a multidimensional integral that can be very difficult to evaluate and this serves as a significant barrier to its use in ensemble selection. Our ensemble method is based on an approximate variational Bayesian inference (VBI) method for model selection. The VBI approach has two major benefits. First, it is significantly faster than complete Bayesian inference, which enables the use of large structural libraries. Second, VBI implicitly leads to maximization of ME without the need for evaluation of a multidimensional integral. However, a downside of the VBI approach is the approximation of the posterior inherent in method. Hence, after arriving at the optimal ensemble with VBI we carry out a complete Bayesian inference of weights. This enables quantification of uncertainties in the ensemble model and population weights.

A significant benefit of Bayesian methods is that multiple experimental observations can be rigorously combined in both model selection and weight inference to gain insight into the underlying ensemble. It is also possible to combine experimental data with information from simulations and force fields. This is exemplified in this study where we demonstrate how data from SAXS, NMR and structural energy values of individual conformers can be combined into one probabilistic model. The inference machinery is applied study the conformational ensembles of two-domain proteins studied by SAXS. We demonstrate the feasibility of full Bayesian inference from large structural libraries from detailed all atom simulations and show by simulations that the method is capable of accurate recovery of population weights and ensemble sizes. We also investigate how noise in the experimental data impacts the accuracy of ensemble inference and show that information encoded in energy functions can compensate for noisy SAXS data. The method is applied to two systems with experimental data from SAXS and NMR: calmodulin and cardiac myosin binding protein C. This analysis demonstrated how the combination of data from SAXS, NMR and structural energies from conformational sampling simulations can be used in synergy for improved ensemble inference.