Improving MCMC sampler performance for statistical models of biochemical thermodynamics

Introduction

Biochemical compounds' formation energies are typically only weakly identified by measurements of reactions' Gibbs energy changes, leading to slow performance of Markov Chain Monte Carlo samplers under naive statistical model parameterisations. This paper presents a new parameterisation that improves the sampler performance, making it feasible to fit Bayesian models of biochemical formation energies.

To demonstrate that the new parameterisation is effective, the paper includes two simulation studies: a simple toy example and a larger example based on publicly available data. In both cases the new parameterisation is shown to lead to better sampler performance.

Background

Thermodynamic parameters are important inputs to kinetic models of cellular metabolism. Much effort has been spent acquiring thermodynamic measurements of important biochemical reactions in order to quantify these parameters. Bayesian statistical analysis of this data is desirable it offers a natural way to accommodate problematic features such as missing measurements, heterogeneity and the need for uncertainty quantification. However, to our knowledge this has not previously been done successfully, likely because of the computational difficulties we identify below.

This section summarises previous work on the quantification of biochemical thermodynamic parameters, sets out the key issues and explains why Bayesian statistical models of biochemical thermodynamics are difficult to fit.

Previous work

Du, Zielinski, and Palsson (2018) is a recent review paper describing the state of the art in estimation of biochemical formation energies. It explains some of the main challenges, particularly incomplete data (particularly in the decomposition of compounds into chemical groups), heterogeneity in informativeness between reactions, and heterogeneity between different kinds of compounds in the accuracy of the group additivity assumption (see below).

A difficulty that Du, Zielinski, and Palsson (2018) do not address directly is that there are potential sources of variation and bias in the adjustment for experimental conditions. For example, the effect of ionic strength is governed by Debeye-Hükel constants that are not known precisely, and the effect of magnesium ion concentration depends on microspecies-specific dissociation constants that are difficult to measure precisely. A further challenge is that downstream applications, such as kinetic modelling of biochemical networks, require not just good point estimates of unknown formation energies but also good overall uncertainty assessments.

Mahamkali et al. (2020) presents a framework for analysing metabolic thermodynamics data by finding boundary conditions for XXXX. This is a better way of generating inputs for a kinetic model than XXX as it takes into account dependencies in the data.

Gollub, Kaltenbach, and Stelling (2020)

Thermodynamic relationships

The Gibbs energy change of a chemical reaction in aqueous solution, denoted $\Delta_r G$, is the amount of energy the reaction releases from its reactants or stores from its environment. Gibbs free energy changes associated with biochemical reactions are important properties of metabolic networks. In order to make inferences about a biochemical system - for example to predict whether overexpressing an enzyme would help or hinder an organism's metabolism - the Gibbs energy changes of the system's reactions must either be measured or inferred.

The Gibbs energy change of a reaction that involves multiple compounds is determined by the Gibbs energy change of each compound's formation reaction in standard conditions, denoted $\Delta_f G^o$ and generally referred to as 'standard condition formation energy', together with its stoichiometry - i.e. the number of molecules of each compound that the reaction creates or consumes - and experimental conditions like temperature, pH, ionic strength and metal ion concentration. Since the dependency on conditions and stoichiometries of most biochemical reactions are relatively well understood, the main challenge in modelling the thermodynamic properties of biochemical reactions is to estimate standard condition formation energies of their compounds.

Relationship between standard condition Gibbs energy change and formation energies

The relationship between a reaction's standard-condition Gibbs free energy change and the standard-condition formation energies of the reactants is as follows:

$$\Delta_r G_j = \sum_{i \ is \ a \ reactant} s_i \Delta_f G_i$$

where s_i is the stoichiometric coefficient of reactant i in reaction j.

For a system of reactions described by stoichiometric matrix S, the relationship can be described as follows:

$$\Delta_r G = S^T \Delta_f G$$

Relationship between standard condition and condition-specific Gibbs energy changes

The Gibbs energy change of a biochemical reaction at given experimental conditions is related to the standard-condition measurement by a complicated but well-understood system of relationships. See Alberty (2003) section 4, 'thermodynamics of pseudo-isomer groups at specified pH', Noor et al. (2013) and Du et al. (2018) for discussion of these relationships, and the supplementary material [SECTION?] for as summary.

In this paper we use a curated data-set of reaction measurements that have already been transformed to their theoretical values at standard conditions.

Approximate group additivity

The latest approaches to estimating biochemical formation energies, e.g. Noor et al. (2013) and Du et al. (2018), assume that compound formation energies are approximately group additive. This is the case when the formation energy of a compound is approximately equal to the sum of the formation energies of its chemical groups. In Noor et al. (2013) this assumption is incorporated by fitting two models - one assuming perfect group additivity ("group contribution") and one assuming no group additivity ("reactant contribution"). The two models are combined using a fall-back procedure, with the reactant contribution model preferred except for compounds where no reactant contribution estimate is available.

In the context of a Bayesian statistical model, approximate group additivity can naturally be taken into account with a multi-level parameter structure. In this

kind of model, compounds' standard condition formation energies are informed by latent parameters representing the standard condition formation energies of chemical groups. The relevance of group sum can be determined based on the data by a higher order parameter.

Computational issues for structured Bayesian models

The main feature that makes it difficult to apply structured Bayesian modelling to biochemical thermodynamics is that formation energies are typically only weakly identified by the available reaction measurements. This happens because a single reaction measurement is informative only about a linear combination of the formation energies of its compounds. For example, consider a reaction with the following stoichiometry:

$$A + B \leftrightarrow 2C$$

If this reaction's Gibbs free energy change were measured in standard conditions, this would provide information as to the value of $2\theta_C - \theta_A - \theta_B$. In order to fix the value of all three unknowns, three independent linear combinations would need to be measured or informed by prior information.

Unfortunately the available measurements typically do not identify enough linear combinations to determine the absolute values of all formation energies, even given perfect measurements. In addition, reaction measurements are typically very accurate relative to the accuracy of the other available information about formation energies.

Together these features of the formation energy estimation problem induce what Betancourt, Michael (2020) terms 'additive degeneracy' in the target posterior distribution. Instead of concentrating in one region of parameter space, posterior mass is spread out along comparatively diffuse surfaces, making the posterior distribution difficult to explore.

Degenerate posterior distributions can often be explored more efficiently when they are reparameterised. For example, Stan's adaptive HMC algorithm optionally transforms unknown parameters in order to take into account correlations. Unfortunately in the case of formation energy estimation the induced parameter correlations are too extreme for this or other generic re-parameterisation strategies to be effective. Thus a re-parameterisation strategy that is tailored to the problem of formation energy estimation seems to be needed.

Method

We made a strategy for parameterising Bayesian regression models of biochemical thermodynamics that addresses the typical additive degeneracy issue. To demonstrate that it works we made a representative regression model and compared its behaviour under a naive parameterisation and a re-parameterisation based on our strategy.

Modelling approach and reparameterisation strategy

Our full model is as follows:

$$
\begin{aligned}
y &\sim N(\Delta_r G, \sigma_r) \\
\Delta_r G &= S^T \theta \\
\theta &\sim Normal(G\gamma, \tau) \\
\theta &\sim Normal(\mu_\theta, \sigma_\theta) \\
\gamma &\sim Normal(\mu_\gamma, \sigma_\gamma) \\
\tau &\sim Half\text{-}Normal(\mu_\tau, \sigma_\tau)
\end{aligned}
$$

where

- γ and θ are vectors representing unknown group and compound formation energies respectively
- tau is an unknown positive number representing the inaccuracy of the group additivity assumption
- S and G are known matrices representing reaction stoichiometry and group incidence respectively
- μ_θ and μ_{gamma} are vectors of known prior means for compound and group formation energies
- σ_θ and σ_γ are vectors of known prior standard deviations for compound and group formation energies
- μ_τ and σ_τ are the known prior mean and standard deviation for the inaccuracy of the group additivity assumption

Naive parameterisation

We made two Stan programs implementing our model: one with a naive parameterisation and one with our new parameterisation.

See Carpenter et al. (2017) for a description of the Stan probabilistic programming language.

The naive model is as follows:

```
data {
  int<lower=1> N_rxn;
  int<lower=1> N_cpd;
  int<lower=1> N_grp;
  matrix[N_cpd, N_rxn] S;
  matrix[N_cpd, N_grp] G;
  vector[N_rxn] y;
  vector<lower=0>[N_rxn] sigma;
  vector[N_cpd] prior_theta[2];
  vector[N_grp] prior_gamma[2];
  real prior_tau[2];
}
parameters {
  real<lower=0> tau;
  vector[N_cpd] theta;
  vector[N_grp] gamma;
}
model {
  target += normal_lpdf(theta | prior_theta[1], prior_theta[2]);
  target += normal_lpdf(gamma | prior_gamma[1], prior_gamma[2]);
  target += normal_lpdf(tau | prior_tau[1], prior_tau[2]);
  target += normal_lpdf(theta | G * gamma, tau);  // approximate group additivity
  target += normal_lpdf(y | S' * theta, sigma);
}
generated quantities {
  vector[N_rxn] yrep;
  real log_lik = 0;
  {
    vector[N_rxn] yhat = S' * theta;
    for (n in 1:N_rxn){
      yrep[n] = normal_rng(yhat[n], sigma[n]);
      log_lik += normal_lpdf(y[n] | yhat[n], sigma[n]);
    }
  }
}
```

The re-parameterised model does not include a Jacobian adjustment, even though the target log probability is incremented by the probability density of the transformed parameter variable `theta`. This is because the transformation from `eta_cpd_z` to `theta` is linear - i.e. adding and multiplying by vectors of constants. Since, for a Bayesian statistical model it is only necessary to know the posterior density up to proportionality, the Jacobian adjustment can safely be ignored.

New parameterisation

Following Alberty (1991) we find an identifiable set of N_{rxn} parameters by multiplying the original parameters θ by the reduced row echelon form of the transpose of stoichiometric matrix, $rref(S^T)$. In order to incorporate pre-experimental information about the formation energies, we embed $rref(S^T)$ inside the $N_{cpd} \times N_{cpd}$ identity matrix to define a $N_{cpd} \times N_{cpd}$ transformation matrix R. The procedure for generating R is as follows.

1. Start with a $N_{cpd} \times N_{cpd}$ identity matrix called R
2. Find the $N_{cpd} \times N_{rxn}$ stoichiometric matrix S.
3. Find the reduced row echelon form of S^T, $rref(S^T)$
4. For each of $rref(S^T)$' leading ones, find its row index i and column index j, and replace row j of R with row i of $rref(S^T)$

The re-parameterised Stan program is as follows:

```
data {
  int<lower=1> N_rxn;
  int<lower=1> N_cpd;
  int<lower=1> N_grp;
  matrix[N_cpd, N_rxn] S;
  matrix[N_cpd, N_grp] G;
  matrix[N_cpd, N_cpd] R;
  matrix[N_cpd, N_cpd] R_inv;
  matrix[N_grp, N_grp] RG;
  matrix[N_grp, N_grp] RG_inv;
  vector[N_rxn] y;
  vector<lower=0>[N_rxn] sigma;
  vector[N_grp] prior_gamma[2];
  vector[N_cpd] prior_theta[2];
  real prior_tau[2];
}
parameters {
  real<lower=0> tau;  // controls group additivity accuracy
  vector[N_cpd] eta_cpd;
  vector[N_grp] eta_grp;
}
transformed parameters {
  vector[N_cpd] theta = R_inv * eta_cpd;
  vector[N_grp] gamma = RG_inv * eta_grp;
}
model {
  target += normal_lpdf(theta | prior_theta[1], prior_theta[2]);
  target += normal_lpdf(gamma | prior_gamma[1], prior_gamma[2]);
  target += normal_lpdf(theta | G * gamma, tau);
  target += normal_lpdf(tau | prior_tau[1], prior_tau[2]);
```

```
  target += normal_lpdf(y | S' * theta, sigma);
}
generated quantities {
  vector[N_rxn] yrep;
  real log_lik = 0;
  {
    vector[N_rxn] yhat = S' * theta;
    for (n in 1:N_rxn){
      yrep[n] = normal_rng(yhat[n], sigma[n]);
      log_lik += normal_lpdf(y[n] | yhat[n], sigma[n]);
    }
  }
}
```

The two programs are almost the same, with the only difference being that the re-parameterised model uses the known matrix `R_inv` to define the transformed parameter `theta`.

Results

Example 1: simple toy case

This section illustrates our method with a worked example.

The stoichiometric matrix S and $rref(S^T)$, the reduced row echelon form of its transpose are as follows:

$$S = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0 \\ 1 & 1 & 2 & 0 \end{bmatrix} \; rref(S^T) = \begin{bmatrix} 1 & 1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

In this the re-parameterisation matrix R is

$$R = \begin{bmatrix} 1 & 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

with rows 1, 4, 5 and 6 taken from $rref(ST)$ because its leading ones are at these columns.

If the original parameters were

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{bmatrix}$$

then the transformed parameters would be

$$\gamma = R\theta = \begin{bmatrix} \theta_a - \theta_b + 2\theta_c \\ \theta_b \\ \theta_c \\ \theta_d \\ \theta_e \\ \theta_f \end{bmatrix}$$

Given γ, the interpretable parameters θ can be recovered using the following relationship:

$$\theta = R^{-1}\gamma = \begin{bmatrix} \gamma_a + \gamma_b - 2\gamma_c \\ \gamma_b \\ \gamma_c \\ \gamma_d \\ \gamma_e \\ \gamma_f \end{bmatrix}$$

Simulation study

We ran a simulation study to compare the performance of the naive and reparameterised models using the stoichiometry specified above and a simulated data generating process matching the model, with the following true parameters (all units are kJ/mol):

- τ: 40
- σ_r: [1, 20, 20, 20]
- \$: [-900, -700, -700, -1910]

We generated values for *theta* and then simulated measurements y using these inputs and our model assumptions.

We then fit our model under both parameterisations, using the following priors:

- μ_θ: [-900, -700, -600, -2000, -1400, -2700]
- σ_θ: [400, 400, 400, 400, 400, 400]

- μ_γ: [-1000, -700, -600, -2000]
- σ_γ: [400, 400, 400, 400]
- μ_τ: 50
- σ_τ: 20

All of these values were chosen to broadly reflect a realistic measurement and prior information setup.

For each parameterisation we ran 4 Markov chains with 2000 warmup and 2000 sampling iterations per chain. We configured Stan's sampler to use a binary tree with depth of at most 13 (the default cap is 10) and otherwise used Stan's default configuration.

Both programs converged satisfactorily. and approximately recovered the parameters of the true data generating process, as can be seen from the following graph:

Here are a pair of histograms showing the number of leapfrog steps per iteration that Stan's sampler took while running 4 chains with 1000 iterations per chain under both parameterisations:

The sampler tended to take far more leapfrog steps under the new parameterisation compared to the naive parameterisation, indicating that our method improved performance in this case.

Example 2: component contribution data

To test whether the performance improvement extends to more a realistic dataset we ran a second simulation study with a larger dataset.

We took the stoichiometry for XXX reaction and XXX compounds from the python library and website equilibrator Flamholz et al. (2012). We assumed that equilibrator's group formation energy estimates were true, then used these values to generate compound formation energies, under the assumption that compound formation energies are approximately group additive with standard deviation 50 kJ/mol.

We assumed known measurement error standard deviation of +/-5 kJ/mol for measurements of non-formation reactions, and +/-200 kJ/mol for formation reactions.

We generated one measurement per reaction.

To make the test shorter and remove variation due to different time spent finding the typical set we initialised all parameters at their true values and set Stan's initial adaptation window to zero.

Results

Figure 1 shows the marginal posterior predictive distributions for the model's training measurements:

Marginal posterior predictive distributions

Very few reaction measurements fall outside the 90% intervals, suggesting a degree of under-fitting. There also seems to be some systematic bias in the predictions for measurements of particularly low formation energies, which tend to be a bit higher than the measurements.

The marginal distributions of compound and group formation energies are shown in figure 2.

Marginal posterior distributions for compound and group formation energies

The component contribution estimates are shown in red. By and large the two models agree as to the best estimates, though there are some systematic differences for compounds and groups with particularly high and low estimated formation energy.

Here is a comparison of leapfrog steps taken by the naive and novel samplers:

Although the novel parameterisation resulted in substantially fewer leapfrog steps, the difference is not as marked as in the simple toy case study. We think this is because of the larger role played in the second case study by group contributions, which were not reparameterised.

Discussion

The simulation study shows that the reparameterisation works in a case where the true data generating process matches our model assumptions, recovering the input while reducing the computational burden compared to the naive parameterisation. The case study using real data shows that performance improves with a realistic model and dataset.

Acknowledgements

This work was funded by the Novo Nordisk Foundation Center for Biosustainability.

The authors are very grateful to Elaad Noor, Daniel Zielinski, Aki Vehtari,

References

Alberty, R A. 1991. "Equilibrium Compositions of Solutions of Biochemical Species and Heats of Biochemical Reactions." *Proceedings of the National Academy of Sciences of the United States of America* 88 (8): 3268–71.

Alberty, Robert A. 2003. *Thermodynamics of Biochemical Reactions*. Hoboken, N.J: Wiley-Interscience.

Betancourt, Michael. 2020. "Identity Crisis." https://betanalpha.github.io/assets/case_studies/identifiability.h

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32. https://doi.org/10.18637/jss.v076.i01.

Du, Bin, Zhen Zhang, Sharon Grubner, James T. Yurkovich, Bernhard O. Palsson, and Daniel C. Zielinski. 2018. "Temperature-Dependent Estimation of Gibbs Energies Using an Updated Group-Contribution Method." *Biophysical Journal* 114 (11): 2691–2702. https://doi.org/10.1016/j.bpj.2018.04.030.

Du, Bin, Daniel C. Zielinski, and Bernhard O. Palsson. 2018. "Estimating Metabolic Equilibrium Constants: Progress and Future Challenges." *Trends in Biochemical Sciences* 43 (12): 960–69. https://doi.org/10.1016/j.tibs.2018.09.009.

Flamholz, Avi, Elad Noor, Arren Bar-Even, and Ron Milo. 2012. "eQuilibratorthe Biochemical Thermodynamics Calculator." *Nucleic Acids Research* 40 (Database issue): D770–D775. https://doi.org/10.1093/nar/gkr874.

Gollub, Mattia G., Hans-Michael Kaltenbach, and Jörg Stelling. 2020. "Probabilistic Thermodynamic Analysis of Metabolic Networks." *bioRxiv*, August, 2020.08.14.250845. https://doi.org/10.1101/2020.08.14.250845.

Mahamkali, Vishnuvardhan, Tim McCubbin, Moritz Beber, Esteban Marcellin, and Lars K. Nielsen. 2020. "multiTFA: A Python Package for Multi-Variate Thermodynamics-Based Flux Analysis." *bioRxiv*, December, 2020.12.01.407387. https://doi.org/10.1101/2020.12.01.407387.

Noor, Elad, Hulda S. Haraldsdóttir, Ron Milo, and Ronan M. T. Fleming. 2013. "Consistent Estimation of Gibbs Energy Using Component Contributions." Edited by Daniel A. Beard. *PLoS Computational Biology* 9 (7): e1003098. https://doi.org/10.1371/journal.pcbi.1003098.