multiTFA: a Python package for multi-variate Thermodynamics-based Flux Analysis

Vishnuvardhan Mahamkali[a], Tim McCubbin[a], Moritz Emanuel Beber[b], Esteban Marcellin[a] and Lars Keld Nielsen[a,b]

[a]Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, 4072 Brisbane, Australia

[b]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

*Corresponding author: Prof. Lars Keld Nielsen; lars.nielsen@uq.edu.au; Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, 4072 Brisbane, Australia

Abstract

Summary

We achieve a significant improvement in thermodynamic-based flux analysis (TFA) by introducing multivariate treatment of thermodynamic variables and leveraging component contribution, the state-of-the-art implementation of the group contribution methodology. Overall, the method greatly reduces the uncertainty of thermodynamic variables.

Results

We present multiTFA, a Python implementation of our framework. We evaluated our application using the core *E. coli* model and achieved a median reduction of 6.8 kJ/mol in reaction Gibbs free energy ranges, while three out of 12 reactions in glycolysis changed from reversible to irreversible.

Availability and implementation

Our framework along with documentation is available on https://github.com/biosustain/multitfa.

1 Introduction

Constraint-based analysis of metabolic network models is used widely to explore metabolic phenotypes and guide metabolic designs (O'Brien *et al.*, 2015). Thermodynamic-based flux analysis (TFA) imposes thermodynamic constraints on constraint-based models, in order to obtain thermodynamically valid metabolic fluxes and metabolite concentration profiles (Henry *et al.*, 2007). TFA provides an ideal mechanism for incorporating metabolomics data into genome-scale modelling. TFA is also a critical pre-processing step when performing sampling based fitting and exploration of large kinetic models (Saa and Nielsen, 2017).

Thermodynamic constraints rely on the calculation of Gibbs free energies of compounds and reactions. The current best method for estimating standard Gibbs free energy of reaction ($\Delta_r G^{'}_{i}$) uses the component contribution method, which combines reactant and group contribution methods while maintaining thermodynamic consistency (Noor *et al.*, 2013; Flamholz *et al.*, 2012).

Accommodating the errors in the estimated $\Delta_r G'_i$ presents a challenge. We cannot introduce independent slack in each $\Delta_r G'_i$, since this would cause inconsistent thermodynamics with non-zero Gibbs energy loops. The original TFA implementation was based solely on the group contribution method (Henry *et al.*, 2007). It avoided inconsistency by computing $\Delta_r G'_i$ within the algorithm from 'groups' treated as independent variables allowed to vary within their individual 95% confidence intervals, i.e., approximately two standard deviations (*SD*) around the mean ($\bar{\mu}$) (Henry *et al.*, 2007). More recently, the loop issue has been addressed by using metabolite formation energies rather than reaction energies in pyTFA (Salvy *et al.*, 2019). These formation energies can be user defined or calculated as a linear combination of respective group Gibbs free energies (from a suitable database). The pyTFA algorithm also treats formation energies as independent variables which are allowed to vary in the range $\bar{\mu} \pm 2 \cdot SD$ (Salvy *et al.*, 2019).

It is not optimal to use the n-box formed from individual 95%-confidence intervals to capture the range of feasible values in a multivariate distribution, such as the full set of formation energies generated by the component contribution method. Firstly, the n-box does not define a 95%-confidence range for the mean vector of formation energies. Secondly, and more importantly, it does not capture the correlation in the distribution. This is particularly problematic using formation energies for substrates and products linked through a reaction, since they will tend to be highly correlated. For illustration consider the multivariate normal distribution estimates for ATP and ADP with the following mean vector and covariance (Σ) matrix (see details later):

$$\Delta_f G'^0 \epsilon N\left(\begin{pmatrix}-2814.24\\-1945.90\end{pmatrix},\begin{pmatrix}2.32 & 1.77\\1.77 & 1.57\end{pmatrix}\right)$$

Unsurprisingly, the estimates are highly correlated (0.928), which is reflected in the cigar shaped 95%-confidence ellipse (Fig. 1A, blue line). While the n-box (red box) almost captures the range, it fails to capture the correlation. Since ATP and ADP are commonly found on either side of a reaction, the most important value is the difference in free energy. The range for the difference is much smaller using the proper confidence region (2.9 kJ/mol) compared to using the n-box (10.9 kJ/mol). Using multivariate confidence regions effectively ensures that we cancel out common error contributions on either side of a reaction

(Haraldsdóttir *et al.*, 2012). We note that the original method from Henry *et al.* achieved error cancellation for common groups but did not address correlation between group estimates.

multiTFA is an internally consistent TFA framework with multivariate treatment of errors in formation energies. Constraining formation energies within the 95% confidence ellipsoid rather than the n-box more accurately captures the range of values, while narrowing the likely range of free energies of reaction and concentration values.

2 Materials and Methods

The constraints in TFA are (Salvy *et al.*, 2019)

$$S * v = 0 \tag{1}$$

$$0 \leq v_i \leq y_i * v_{max} \tag{2}$$

$$\Delta_r G_i' + K * y_i < K \tag{3}$$

$$\Delta_r G' = S^T \left(\Delta_f G' + RTln(x)\right) + \Delta_r G_{transport} \tag{4}$$

$$x \in \Omega_x \tag{5}$$

$$\Delta_f G' \in \Omega_G \tag{6}$$

where *S* and *v* are the stoichiometric matrix and flux vector, respectively. Reactions only progress in the forward direction (i.e., reversible reactions are split in two), and only if the binary coupling variable, y_i, is 1 (2), which can only happen when the Gibbs free energy of the reaction ($\Delta_r G'$) is negative (3) (K is a large positive constant).

$\Delta_r G'$ is calculated from the formation energies ($\Delta_f G'$) and concentrations (x) of the metabolites (4). For transporters, the Gibbs free energy of transport was calculated as detailed in (Jol *et al.*, 2010). Our implementation automatically detects transporters and predicts the species that is being transported based on the pKa value and the compartment pH. Users are also able to explicitly define the charged form of the transported metabolite and transportation mechanism. For the calculation of Gibbs free energies of reaction at non-standard conditions, users can input the range for each metabolite concentration, Ω_x(5). Where not specified, metabolites can assume pre-defined compartment specific bounds or otherwise adopt loose bounds ($10^{-5} – 10^{-2}$ M).

The formation energies are estimated using the component contribution method (Noor *et al.*, 2013) and adjusted for compartment specific pH and ionic strength (Alberty, 2005;

Haraldsdóttir *et al.*, 2012). It is assumed that the estimate follows a multivariate normal distribution, $\Delta_f G' \epsilon N(\mu, \Sigma)$. TFA allows for noise in the $\Delta_f G'$ estimate by defining a region, Ω_G (6). A common approach is to use as Ω_G the n-box defined by the individual 95%-confidence intervals for each formation Gibbs energy, i.e. $\bar{\mu} \pm u_{97.5\%}\left(diag(\bar{\Sigma})\right)^{1/2}$. This is not a true 95% confidence region for the multivariate estimate: it greatly underestimates the range of individual variables and ignores the correlation between related compounds such as ADP and ATP. A more appropriate region, Ω_G, would be the 95%-confidence ellipsoid defined by:

$$(\bar{\mu} - \mu)^T \bar{\Sigma}^{-1} (\bar{\mu} - \mu) \leq \chi^2_{n,95\%} \qquad (7)$$

Introducing this constraint converts the problem from a Mixed Integer Linear Problem (MILP) to a Mixed Integer Quadratic Constraint Problem (MIQCP).

In general, Σ does not have full rank and (7) cannot be used directly. $\Delta_f G'$ is calculated using the component contribution method as

$$\Delta_f G' = V_{cc} \Delta_{cc} G'^0 + \Delta\Delta G \qquad (8)$$

Where $V_{CC}$is the metabolite component composition vector, $\Delta_{cc} G'^0$ is a vector of component and group standard Gibbs energies, and $\Delta\Delta G$ is a (deterministic) adjustment for compartment pH, pI and Mg concentration. This estimate is the fit of the component contribution model to the thermodynamic reference data and is assumed to follow a multivariate normal distribution, $\Delta_{cc} G'^0 \epsilon N(\mu_{cc}, \Sigma_{cc})$. We can express this distribution as

$$\Delta_{cc} G'^0 = \mu_{cc} + Cu \qquad (9)$$

where $CC' = \Sigma_{cc}$ and $u \epsilon N(0, I)$. Allowing for Σ_{cc} not having full rank, we use LDL decomposition to achieve pivoted Cholesky decomposition finding $C_{full} = L\sqrt{D}$. C_{full} has $n = rank(\Sigma_{cc})$ non-zero columns and we obtain C by removing the remaining columns. Finally, we define a 95%-confidence circle for the n-dimensional standard normal distribution

$$u^T u \leq \chi^2_{n,95\%} \qquad (10)$$

Using (10), we still need to solve an MIQCP, but this is numerically more robust.

We compared multiTFA with the n-box approach using an *E. coli* core model (e_coli_core) (Orth *et al.*, 2010). We determined the Gibbs free energy ranges using either (a) the "conventional" $\bar{\mu} \pm 1.96 \cdot SD$ n-box or (b) a multivariate treatment of the errors in the formation energy estimate used in multiTFA. Despite the n-box covering far less than the 95%-confidence range for individual formation energies, the estimated ranges for the Gibbs free energy of reaction were broader than using the confidence ellipsoid. The median reduction in Gibbs free energy ranges was 6.8 kJ/mol (Figure 1b; Supplementary Data), highlighting the significant potential for error cancellation between compounds captured in the correlation matrix. The reduction in Gibbs free energy ranges is reflected in a reduction in the reaction flux ranges (Supplementary Data).

2.1 Usage & Implementation

The Python package comes with example scripts to demonstrate the usage of different functionalities. The software takes a typical COBRA model as input and generates a MILP Optlang object (Jensen *et al.*, 2017) for the n-box approach that can be directly solved with COBRApy (Ebrahim *et al.*, 2013). If the user has Gurobi or CPLEX solver installed, the software will generate solver specific MIQCP objects to solve the multiTFA problem. For users without Gurobi or CPLEX, an alternative implementation of multiTFA is provided that uses random sampling of the surface of the confidence ellipsoid and a MILP solver to determine the maximum range. The exit criterion of the sampler can be chosen as either (1) the number of samples since last improvement or (2) a fixed number of samples followed by use of a generalized extreme value distribution to infer the maximum value.

The implementation is available at https://github.com/biosustain/multitfa. The framework is currently compatible with models that use different identifiers (SEED, KEGG, BIGG among others) for matching metabolite information against the thermodynamic database. We use the eQuilibrator API to retrieve data matrices for calculating the formation energies and covariance matrix (Noor *et al.*, 2013).

3 Conclusion

Using a multivariate confidence ellipsoid to describe the feasible range in the Gibbs free energy of formation estimate, multiTFA is able to account for a more realistic (and broader) range in individual estimates of formation energy, while simultaneously using correlation to reduce the ranges for the derived Gibbs free energies of reactions.

References

Alberty,R.A. (2005) Thermodynamics of Biochemical Reactions at Specified pH. *Thermodyn. Biochem. React.*, 57–88.

Ebrahim,A. *et al.* (2013) COBRApy: Constraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.*, **7**, 74.

Flamholz,A. *et al.* (2012) eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res.*, **40**, D770–D775.

Haraldsdóttir,H.S. *et al.* (2012) Quantitative assignment of reaction directionality in a multicompartmental human metabolic reconstruction. *Biophys. J.*, **102**, 1703–1711.

Henry,C.S. *et al.* (2007) Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.*, **92**, 1792–1805.

Jensen,K. *et al.* (2017) Optlang: An algebraic modeling language for mathematical optimization. *J. Open Source Softw.*, **2**, 139.

Jol,S.J. *et al.* (2010) Thermodynamic calculations for biochemical transport and reaction processes in metabolic networks. *Biophys. J.*, **99**, 3139–3144.

Noor,E. *et al.* (2013) Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol.*, **9**.

O'Brien,E.J. *et al.* (2015) Using genome-scale models to predict biological capabilities. *Cell*, **161**, 971–987.

Orth,J.D. *et al.* (2010) Reconstruction and Use of Microbial Metabolic Networks: the Core *Escherichia coli* Metabolic Model as an Educational Guide. *EcoSal Plus*, **4**.

Saa,P.A. and Nielsen,L.K. (2017) Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. *Biotechnol. Adv.*, 0–1.

Salvy,P. *et al.* (2019) pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. *Bioinformatics*, **35**, 167–169.

Figure Legends

Figure 1. A) Comparing the 95% confidence ellipse (blue) to the n-box (orange) for the eQuilibrator estimates of ATP and ADP formation energies. B) Comparison of the Gibbs free energy of reaction ranges across the glycolytic pathway estimated using the n-box (univariate) and the multiTFA (multivariate) methods. Three reactions change from reversible to irreversible when using a multivariate treatment (ENO, GAPD, and PGM).

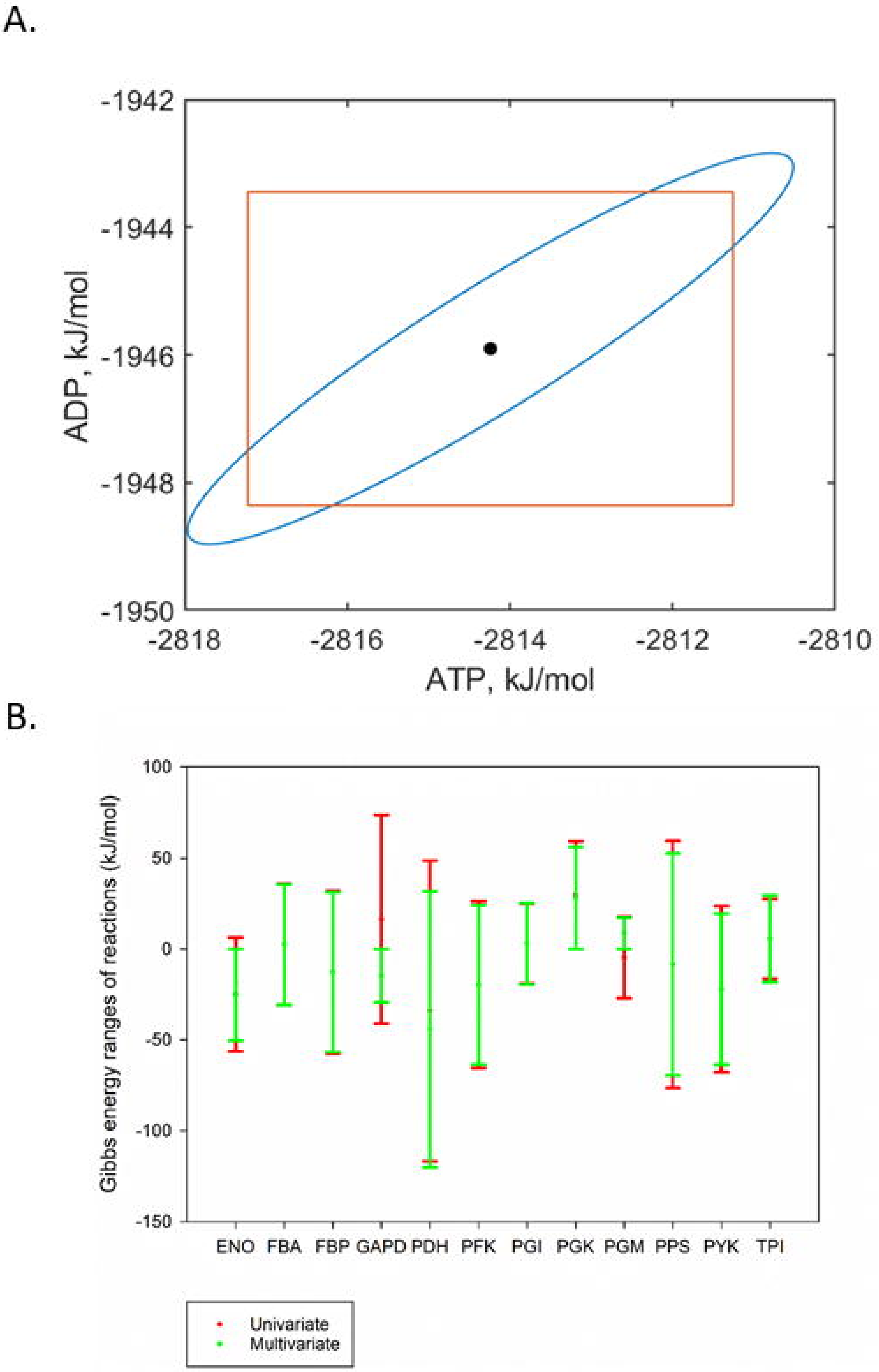

Figure 1. A) Comparing the 95% confidence ellipse (blue) to the n-box (orange) for the eQuilibrator estimates of ATP and ADP formation energies. B) Comparison of the Gibbs free energy of reaction ranges across the glycolytic pathway estimated using the n-box (univariate) and the multiTFA (multivariate) methods. Three reactions change from reversible to irreversible when using a multivariate treatment (ENO, GAPD, and PGM).