

# Current and future applications of sampling algorithms in modelling biochemical networks

Ronan M.T. Fleming

School of Medicine, National University of Ireland, Galway, Ireland  
& Leiden Academic Centre for Drug Research, Leiden University, The Netherlands.

Sampling Algorithms and Geometries on Probability Distributions  
Simons Institute for the Theory of Computing  
29 Sept. 2021

# Outline

- ▶ Constraint-based modelling of biochemical networks

# Outline

- ▶ Constraint-based modelling of biochemical networks
- ▶ Uniform sampling of polyhedral convex constraint-based models

# Outline

- ▶ Constraint-based modelling of biochemical networks
- ▶ Uniform sampling of polyhedral convex constraint-based models
- ▶ Non-uniform sampling of polyhedral convex constraint-based models

# Outline

- ▶ Constraint-based modelling of biochemical networks
- ▶ Uniform sampling of polyhedral convex constraint-based models
- ▶ Non-uniform sampling of polyhedral convex constraint-based models
- ▶ Sampling non-convex feasible sets

# Outline

- ▶ Constraint-based modelling of biochemical networks
- ▶ Uniform sampling of polyhedral convex constraint-based models
- ▶ Non-uniform sampling of polyhedral convex constraint-based models
- ▶ Sampling non-convex feasible sets
- ▶ Entropy optimisation and constraint-based modelling

# Outline

- ▶ Constraint-based modelling of biochemical networks
- ▶ Uniform sampling of polyhedral convex constraint-based models
- ▶ Non-uniform sampling of polyhedral convex constraint-based models
- ▶ Sampling non-convex feasible sets
- ▶ Entropy optimisation and constraint-based modelling
- ▶ Sampling the intersection of polyhedral convex and convex cone constraints

# Outline

- ▶ Constraint-based modelling of biochemical networks
- ▶ Uniform sampling of polyhedral convex constraint-based models
- ▶ Non-uniform sampling of polyhedral convex constraint-based models
- ▶ Sampling non-convex feasible sets
- ▶ Entropy optimisation and constraint-based modelling
- ▶ Sampling the intersection of polyhedral convex and convex cone constraints

# Notation

- ▶ Unless specified otherwise, all variables are real valued.
- ▶ Householder notation:
  - ▶  $A$ , matrix;  $b$ , column vector;  $b_i$  is the  $i^{\text{th}}$  entry in a column vector
  - ▶  $\Omega$ , set;  $\omega$ , scalar.
  - ▶  $\phi(x)$  is a scalar valued function of a vector variable
  - ▶  $f(x)$  is a vector valued function of a vector variable
- ▶  $I$  is an identity matrix
- ▶  $\mathbf{1}$  is a vector of ones
- ▶  $[A, B]$  horizontal concatenation of two matrices
- ▶  $\log(x)$  is the component-wise logarithm of each element

# Generic versus mechanistic modelling

- ▶ **Generic modelling**
  - ▶ Mathematical modelling approaches that do not satisfy any mechanistic principles
    - ▶ e.g. network inference with no limitation on the class of network topology being inferred

# Generic versus mechanistic modelling

- ▶ **Generic modelling**

- ▶ Mathematical modelling approaches that do not satisfy any mechanistic principles
  - ▶ e.g. network inference with no limitation on the class of network topology being inferred

- ▶ **Mechanistic modelling**

- ▶ Mathematical model that satisfy certain mechanistic principles
  - ▶ e.g. network inference where the inferred biochemical network topology must satisfy mass conservation

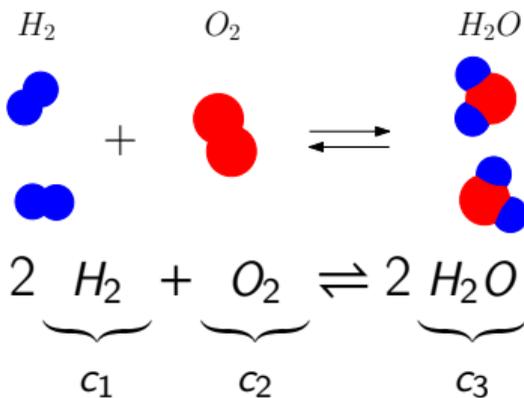
# Main mechanistic modelling approaches

- ▶ **Differential equation based modelling**
  - ▶ input: biochemical network topology, uniquely specified initial conditions and parameters
  - ▶ output: unique temporal trajectory

# Main mechanistic modelling approaches

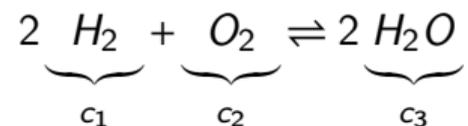
- ▶ **Differential equation based modelling**
  - ▶ input: biochemical network topology, uniquely specified initial conditions and parameters
  - ▶ output: unique temporal trajectory
- ▶ **Constraint-based modelling**
  - ▶ input: biochemical network topology, non-unique initial conditions and parameters
  - ▶ output: non-unique set of temporal trajectories

## A reversible elementary reaction



- ▶ An example elementary reaction
  - ▶  $c_1$ ,  $c_2$  and  $c_3$  denote the concentration of molecular species  $H_2$ ,  $O_2$  and  $H_2O$  respectively
- ▶ **Chemically** all reactions are reversible in principle, but in practice only one direction may be **biochemically** feasible.

## Dynamics under elementary reaction kinetics



- ▶ Assuming elementary kinetics, the forward and reverse elementary rate functions are  $v_f(c) := k_f c_1^2 c_2$  and  $v_r(c) := k_r c_3^2$ , where  $k_f, k_r$  are (elementary) kinetic parameters.
- ▶ The dynamical equations are

$$\begin{aligned}\frac{dc_1}{dt} &= -2k_f c_1^2 c_2 + 2k_r c_3^2 \\ \frac{dc_2}{dt} &= -k_f c_1^2 c_2 + k_r c_3^2 \\ \frac{dc_3}{dt} &= -2k_r c_3^2 + 2k_f c_1^2 c_2\end{aligned}$$

## Dynamics under elementary reaction kinetics

- ▶ The dynamical equations are

$$\frac{dc_1}{dt} = -2k_f c_1^2 c_2 + 2k_r c_3^2$$

$$\frac{dc_2}{dt} = -k_f c_1^2 c_2 + k_r c_3^2$$

$$\frac{dc_3}{dt} = -2k_r c_3^2 + 2k_f c_1^2 c_2$$

- ▶ Let  $N := \begin{bmatrix} -2 \\ -1 \\ 2 \end{bmatrix} \in \mathbb{Z}^{m \times n}$  denote a **stoichiometric matrix**, then given  $k_f, k_r$  we have

$$\frac{dc}{dt} = N(v_f(c) - v_r(c))$$

- ▶ One approach is to define  $v(c) := v_f(c) - v_r(c)$  as some composite of many elementary reactions.

## Advantages of a differential equation approach

- ▶ Correct representation of known chemical rate laws
  - ▶ reaction rates are explicit nonlinear functions of molecular species concentration

## Advantages of a differential equation approach

- ▶ Correct representation of known chemical rate laws
  - ▶ reaction rates are explicit nonlinear functions of molecular species concentration
- ▶ Steady-state and dynamic modelling are both possible
  - ▶ can be compared with both types of experimental data

## Advantages of a differential equation approach

- ▶ Correct representation of known chemical rate laws
  - ▶ reaction rates are explicit nonlinear functions of molecular species concentration
- ▶ Steady-state and dynamic modelling are both possible
  - ▶ can be compared with both types of experimental data
- ▶ Necessary for modelling certain biochemical processes
  - ▶ e.g. dynamics of signalling networks

## Advantages of a differential equation approach

- ▶ Correct representation of known chemical rate laws
  - ▶ reaction rates are explicit nonlinear functions of molecular species concentration
- ▶ Steady-state and dynamic modelling are both possible
  - ▶ can be compared with both types of experimental data
- ▶ Necessary for modelling certain biochemical processes
  - ▶ e.g. dynamics of signalling networks
- ▶ A rich variety of mathematical and computational tools are already in existence

## Advantages of a differential equation approach

- ▶ Correct representation of known chemical rate laws
  - ▶ reaction rates are explicit nonlinear functions of molecular species concentration
- ▶ Steady-state and dynamic modelling are both possible
  - ▶ can be compared with both types of experimental data
- ▶ Necessary for modelling certain biochemical processes
  - ▶ e.g. dynamics of signalling networks
- ▶ A rich variety of mathematical and computational tools are already in existence

## Disadvantages of a differential equation approach

- ▶  $c(t) = c_0 + \int \frac{dc}{dt}$
- ▶ Initial condition,  $c_0$ , is often not known

## Disadvantages of a differential equation approach

- ▶  $c(t) = c_0 + \int \frac{dc}{dt}$
- ▶ Initial condition,  $c_0$ , is often not known
- ▶ Most of the kinetic parameters  $k_f$  and  $k_r$  are also not known

## Disadvantages of a differential equation approach

- ▶  $c(t) = c_0 + \int \frac{dc}{dt}$
- ▶ Initial condition,  $c_0$ , is often not known
- ▶ Most of the kinetic parameters  $k_f$  and  $k_r$  are also not known
- ▶ Given  $k_f, k_r, c_0$ , the criteria for convergence to a non-equilibrium steady state are not known

## Disadvantages of a differential equation approach

- ▶  $c(t) = c_0 + \int \frac{dc}{dt}$
- ▶ Initial condition,  $c_0$ , is often not known
- ▶ Most of the kinetic parameters  $k_f$  and  $k_r$  are also not known
- ▶ Given  $k_f, k_r, c_0$ , the criteria for convergence to a non-equilibrium steady state are not known
- ▶ Difficult to fit the parameters  $k_f$  and  $k_r$  to (partial) experimental data on  $c$

## Disadvantages of a differential equation approach

- ▶  $c(t) = c_0 + \int \frac{dc}{dt}$
- ▶ Initial condition,  $c_0$ , is often not known
- ▶ Most of the kinetic parameters  $k_f$  and  $k_r$  are also not known
- ▶ Given  $k_f, k_r, c_0$ , the criteria for convergence to a non-equilibrium steady state are not known
- ▶ Difficult to fit the parameters  $k_f$  and  $k_r$  to (partial) experimental data on  $c$
- ▶ Reformulation into composite rate laws an application of generic parameter fitting approaches
  - ▶ lower dimensionality but still all end up in local minima even for small problems

## Disadvantages of a differential equation approach

- ▶  $c(t) = c_0 + \int \frac{dc}{dt}$
- ▶ Initial condition,  $c_0$ , is often not known
- ▶ Most of the kinetic parameters  $k_f$  and  $k_r$  are also not known
- ▶ Given  $k_f, k_r, c_0$ , the criteria for convergence to a non-equilibrium steady state are not known
- ▶ Difficult to fit the parameters  $k_f$  and  $k_r$  to (partial) experimental data on  $c$
- ▶ Reformulation into composite rate laws an application of generic parameter fitting approaches
  - ▶ lower dimensionality but still all end up in local minima even for small problems
- ▶ Currently computationally intractable at genome-scale (high dimensional systems)

# Principles of constraint-based modelling

- ▶ Eliminate infeasible biochemical network states with mathematically specified constraints
  - ▶ Physicochemical constraints, e.g., mass conservation

# Principles of constraint-based modelling

- ▶ Eliminate infeasible biochemical network states with mathematically specified constraints
  - ▶ Physicochemical constraints, e.g., mass conservation
  - ▶ Biochemical constraints, e.g., vitamin C is essential for human but not murine metabolism

# Principles of constraint-based modelling

- ▶ Eliminate infeasible biochemical network states with mathematically specified constraints
  - ▶ Physicochemical constraints, e.g., mass conservation
  - ▶ Biochemical constraints, e.g., vitamin C is essential for human but not murine metabolism
- ▶ Underdetermined systems of equations  $\rightarrow$  Non-unique predictions

# Mass conservation and steady-state

- ▶ Assume mass conservation and steady state
  - ▶ internal production + external input = internal consumption + external output

## Mass conservation and steady-state

- ▶ Assume mass conservation and steady state
  - ▶ internal production + external input = internal consumption + external output
- ▶ We know that each steady state is a solution to

$$\frac{dc}{dt} = N(v_f(c) - v_r(c)) =: 0.$$

## Mass conservation and steady-state

- ▶ Assume mass conservation and steady state
  - ▶ internal production + external input = internal consumption + external output
- ▶ We know that each steady state is a solution to

$$\frac{dc}{dt} = N(v_f(c) - v_r(c)) =: 0.$$

- ▶ Instead, we assume that the set of feasible steady states is defined implicitly by

$$Nv = 0$$

where  $v \in \mathbb{R}^n$  is a variable vector of net reaction rates (*fluxes*) and  $N \in \mathbb{Z}^{m \times n}$  is a given stoichiometric matrix, typically  $m < n < \text{rank}(N)$ .

## Flux balance analysis

- ▶ A prototypical constraint-based modelling approach (Orth, J.et. al. (2010) Nat Biotech 28, 245–248)

## Flux balance analysis

- ▶ A prototypical constraint-based modelling approach (Orth, J.et. al. (2010) Nat Biotech 28, 245–248)
- ▶ Hypothesise a particular linear objective coefficient vector  $d \in \mathbb{R}^n$ , then obtain bounds on net reaction rates  $l, u \in \mathbb{R}^n$  from, e.g., thermochemical data, then

$$\begin{aligned} & \underset{v \in \mathbb{R}^n}{\text{minimise}} && d^T v \\ & \text{s.t.} && Nv = 0 \\ & && l \leq v \leq u \end{aligned} \tag{FBA}$$

where  $N \in \mathbb{Z}^{m \times n}$  is a stoichiometric matrix, typically  $m < n < \text{rank}(N)$ .

## Advantages of flux balance analysis

- ▶ Given a linear objective coefficient vector  $c \in \mathbb{R}^n$ , a stoichiometric matrix  $N \in \mathbb{Z}^{m \times n}$  and bounds on net reaction rates  $l, u \in \mathbb{R}^n$

$$\begin{aligned} & \underset{v \in \mathbb{R}^n}{\text{minimise}} && d^T v \\ & \text{s.t.} && Nv = 0 \\ & && l \leq v \leq u \end{aligned} \tag{FBA}$$

- ▶ Convex (linear) optimisation problem
  - ▶ efficient optimisation software
- ▶ Applicable to genome-scale models (high dimensional)
- ▶ Methodology is accessible to a broad user base
  - ▶ wide variety of variations and applications

## A disadvantage of flux balance analysis

- ▶ We must first hypothesise a particular biochemical objective, i.e.,  $d \in \mathbb{R}^n$  in

$$\begin{aligned} & \underset{v \in \mathbb{R}^n}{\text{minimise}} && d^T v \\ & \text{s.t.} && Nv = 0 \\ & && l \leq v \leq u \end{aligned} \quad (\text{FBA})$$

- ▶ It is unknown what the biochemical objective function is for many systems

## Uniform sampling of polyhedral convex constraint-based models

- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .

## Uniform sampling of polyhedral convex constraint-based models

- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .
- ▶ Many applications, some early examples ...

# Uniform sampling of polyhedral convex constraint-based models

- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .
- ▶ Many applications, some early examples ...
- ▶ **Design of Experiments**
  - ▶ predict the most informative measurements ( $l$  &  $u$ ) to make (Savinell, J. M., and Palsson, B. (1992) J. Theor. Biol. 155, 201–214).

# Uniform sampling of polyhedral convex constraint-based models

- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .
- ▶ Many applications, some early examples ...
- ▶ **Design of Experiments**
  - ▶ predict the most informative measurements ( $l$  &  $u$ ) to make (Savinell, J. M., and Palsson, B. (1992) J. Theor. Biol. 155, 201–214).
- ▶ **Volume computation**
  - ▶ genetic defects that decreased the volume of  $\Omega$  most significantly were more likely to have a clinical effect in vivo (Price, N.D. et. al.(2004) Biophys. J. 87, 2172–2186)

# Uniform sampling of polyhedral convex constraint-based models

- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .
- ▶ Many applications, some early examples ...
- ▶ **Design of Experiments**
  - ▶ predict the most informative measurements ( $l$  &  $u$ ) to make (Savinell, J. M., and Palsson, B. (1992) J. Theor. Biol. 155, 201–214).
- ▶ **Volume computation**
  - ▶ genetic defects that decreased the volume of  $\Omega$  most significantly were more likely to have a clinical effect in vivo (Price, N.D. et al.(2004) Biophys. J. 87, 2172–2186)
- ▶ **Interpretation of marginal distributions**, i.e. distribution of  $v_j$  given sample of  $\Omega$ 
  - ▶ identify set of reactions that are always required for growth of a microorganism in different conditions, same  $N$ , but different  $l$  &  $u$  (Almaas, et al (2004) Nature 427, 839–843)

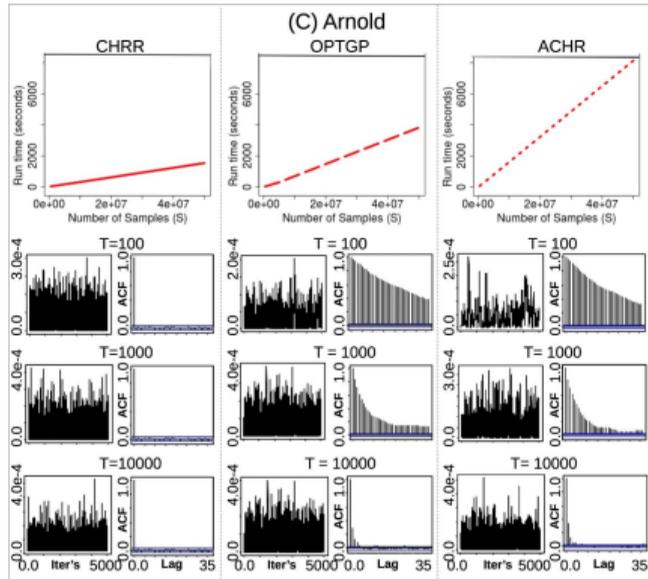
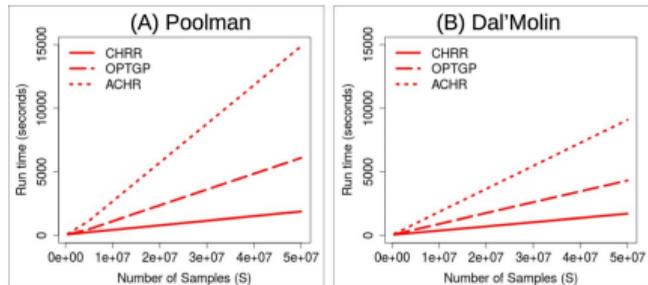
# Uniform sampling of polyhedral convex constraint-based models

- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .
- ▶ Many applications, some early examples ...
- ▶ **Design of Experiments**
  - ▶ predict the most informative measurements ( $l$  &  $u$ ) to make (Savinell, J. M., and Palsson, B. (1992) J. Theor. Biol. 155, 201–214).
- ▶ **Volume computation**
  - ▶ genetic defects that decreased the volume of  $\Omega$  most significantly were more likely to have a clinical effect in vivo (Price, N.D. et al.(2004) Biophys. J. 87, 2172–2186)
- ▶ **Interpretation of marginal distributions**, i.e. distribution of  $v_j$  given sample of  $\Omega$ 
  - ▶ identify set of reactions that are always required for growth of a microorganism in different conditions, same  $N$ , but different  $l$  &  $u$  (Almaas, et al (2004) Nature 427, 839–843)
  - ▶ identification of mass conservation constraints as the reason for changes in rates of human metabolic reactions in diabetes (Thiele, I.,et al (2005) J. Biol. Chem. 280, 11683–11695)

# Uniform sampling of polyhedral convex constraint-based models

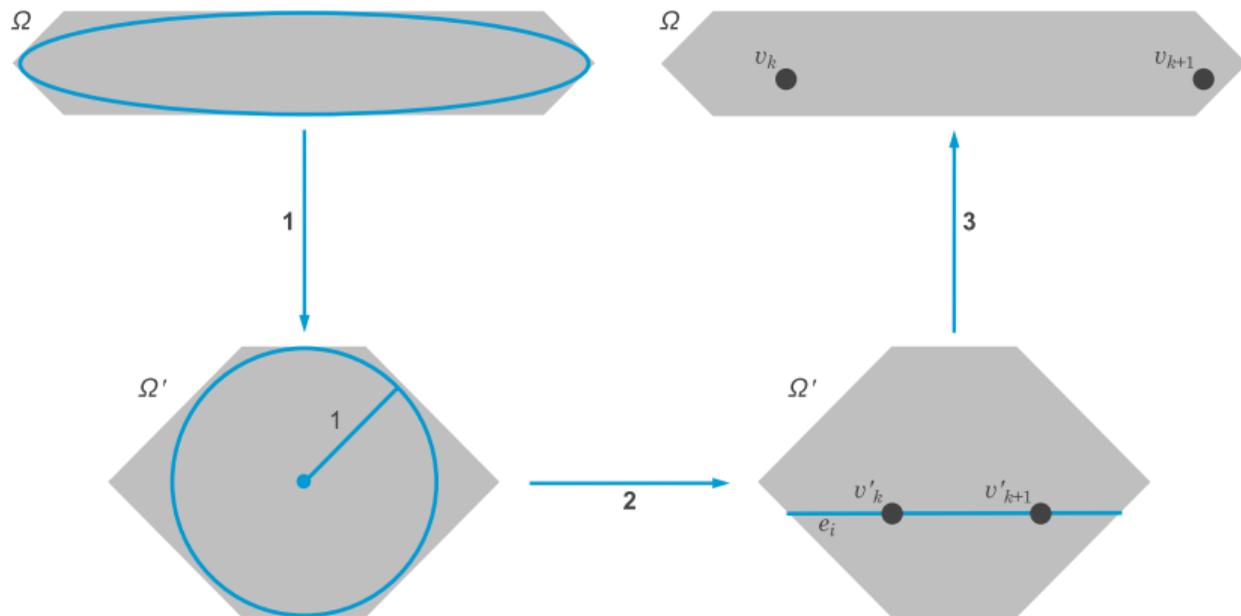
- ▶ Sample polyhedral convex sets of the form  $\Omega := \{Nv = 0, l \leq v \leq u\}$ .
- ▶ Many applications, some early examples ...
- ▶ **Design of Experiments**
  - ▶ predict the most informative measurements ( $l$  &  $u$ ) to make (Savinell, J. M., and Palsson, B. (1992) J. Theor. Biol. 155, 201–214).
- ▶ **Volume computation**
  - ▶ genetic defects that decreased the volume of  $\Omega$  most significantly were more likely to have a clinical effect in vivo (Price, N.D. et al.(2004) Biophys. J. 87, 2172–2186)
- ▶ **Interpretation of marginal distributions**, i.e. distribution of  $v_j$  given sample of  $\Omega$ 
  - ▶ identify set of reactions that are always required for growth of a microorganism in different conditions, same  $N$ , but different  $l$  &  $u$  (Almaas, et al (2004) Nature 427, 839–843)
  - ▶ identification of mass conservation constraints as the reason for changes in rates of human metabolic reactions in diabetes (Thiele, I.,et al (2005) J. Biol. Chem. 280, 11683–11695)

# Uniform sampling of polyhedral convex constraint-based models



- ▶ Herrmann, H.A., et al. **Flux sampling is a powerful tool to study metabolism under changing environmental conditions**. *npj Syst Biol Appl* 5, 32 (2019).
- ▶ **CHRR**: coordinate hit-and-run with rounding (MATLAB, Haraldsdottir, H. S., et al. *Bioinformatics* 33, 1741–1743 (2017))
- ▶ **OPTGP**: optimized general parallel sampler (Python; Megchelenbrink, W., et al. *PLoS ONE* 9, e86587 (2014).)
- ▶ **ACHR**: artificially centered hit-and-run (MATLAB, Python, Kaufman, D. E. et al. *Oper. Res.* 46, 1 (1998).)

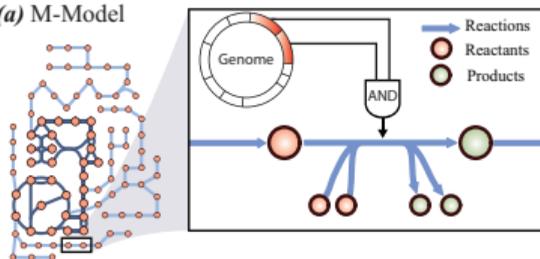
# CHRR: coordinate hit-and-run with rounding



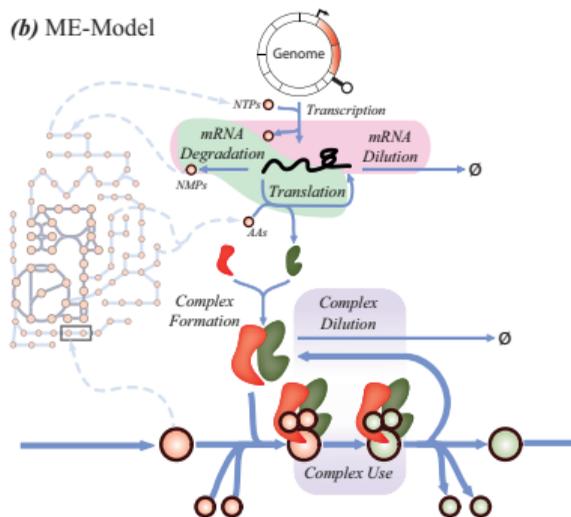
- ▶ Haraldsdóttir, H. S., Cousins, B., Thiele, I., Fleming, R. M. T. & Vempala, S. *Bioinformatics* 33, 1741–1743 (2017).
- ▶ Laddha, A. & Vempala, S. S. *Convergence of Gibbs Sampling: Coordinate Hit-And-Run Mixes Fast*. 12 (2021).
- ▶ Aditi Laddha, *Algorithms for Sampling Convex Bodies*, *Simons Institute Workshop on Sampling Algorithms and Geometries on Probability Distributions*, 2021

# Sampling challenge 1: anisotropy

(a) M-Model

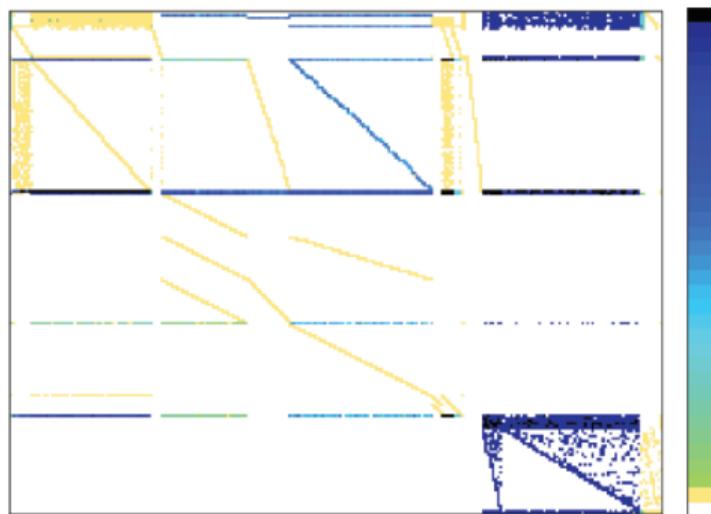


(b) ME-Model



- ▶ Multiscale constraint-based models
- ▶ Examples
  - ▶ **M**etabolism  $\pm$  integration with **E**xpression
  - ▶ Macro & micronutrients
  - ▶ Organ & cellular scales

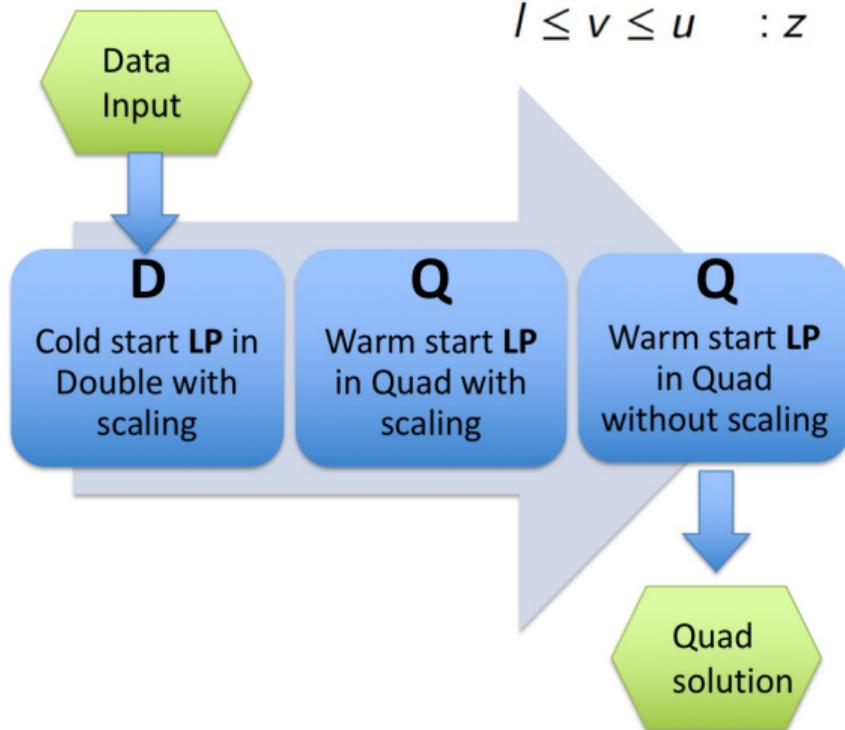
# A multiscale stoichiometric matrix



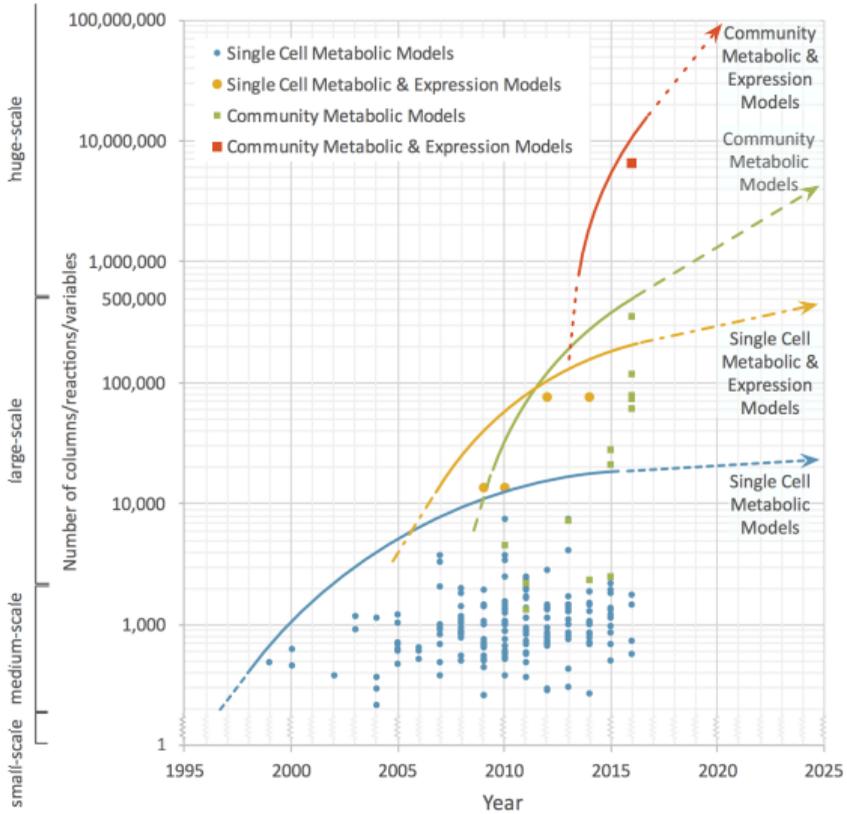
- ▶ **Stoichiometric matrix**,  $\dim(N) = 60,000 \times 80,000$ , of an integrated metabolic and macromolecular expression model in *E. coli* (Thiele et. al. 2012).
- ▶ **Coefficients** are sparse, but spread over 5 orders of magnitude.
- ▶ **Colorbar**: tiny absolute values are light orange, large magnitudes are black. In the midrange, the median of  $\log_{10}$  of the nonzero values,  $\pm 1$  one standard deviation, range from light green to deep blue.
- ▶ Flux bounds are also spread over multiple orders of magnitude (not shown)

# Adapting to anisotropy with novel linear optimisation methods

$$\begin{aligned} & \underset{v \in \mathbb{R}^n}{\text{minimise}} && d^T v \\ & \text{s.t.} && Nv = b \quad : y \\ & && l \leq v \leq u \quad : z \end{aligned}$$

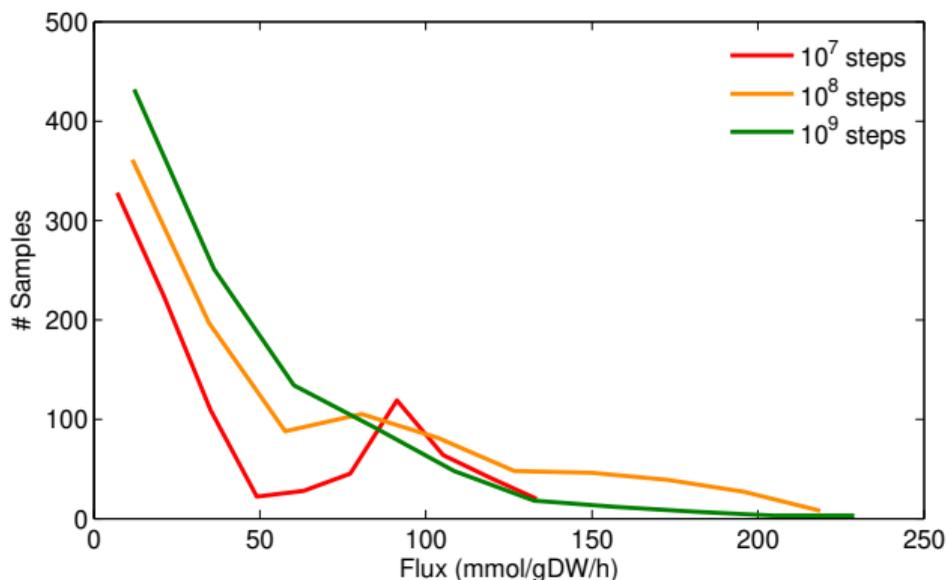


# Sampling challenge 2: dimensionality



► Models will continue to grow in size

## Sampling challenge 3: convergence criteria



- ▶ **Convergence to a stationary sampling distribution, but when is it uniform?**
- ▶ Pre- (red and orange) and post- (green) convergence marginal sampling distributions of **thioredoxin reductase** flux samples obtained with **CHRR** in a constraint-based model of human metabolism (**Recon 2**,  $\dim(\Omega) = 2,430$ ).



## The COBRA Toolbox

- Home
- Installation
- Functions
- Tutorials
- How to contribute
- How to cite
- Support
- FAQ
- Contributors
- Funding
- Development Plan
- Contact

🏠 » Home Page

## The COntstraint-Based Reconstruction and Analysis Toolbox

View The COBRA Toolbox source code on [GitHub](#).



The COntstraint-Based Reconstruction and Analysis Toolbox is a MATLAB software suite for quantitative prediction of cellular and multicellular biochemical networks with constraint-based modelling. It implements a comprehensive collection of basic and advanced modelling methods, including reconstruction and model generation as well as biased and unbiased model-driven analysis methods.

It is widely used for modelling, analysing and predicting a variety of metabolic phenotypes using genome-scale biochemical networks.

The **COBRA Toolbox** is a MATLAB software suite for quantitative prediction of cellular and multicellular biochemical networks with constraint-based modelling.

# COBRA Toolbox

github.com/opencobra/cobratoolbox/

opencobra / cobratoolbox Public

Unwatch 31 Unstar 168 Fork 246

<> Code Issues 62 Pull requests 1 Discussions Actions Security Insights Settings

master 4 branches 14 tags

Go to file Add file Code

rmtfleming Merge pull request #1815 from opencobra/develop ✓ 95f2582 5 days ago 8,151 commits

.artanolis	openAndConvert() in matlab.internal.liveeditor package (not matlab.i...	4 months ago
.github	remove broken link	3 years ago
binary @ e2c9bc5	submodules point to master	12 days ago
deprecated	More readable default font sizes in drawCbMap	2 months ago
docs	Merge pull request #1789 from Gpreciat/chemDB	2 months ago
external	switched to master2	12 days ago
papers @ 92e4835	switched to master	12 days ago
src	Merge pull request #1806 from almut-heinken/updateDemeter	5 days ago
test	Adapted functions to newer versions of MATLAB	6 days ago
tutorials @ 0236f4e	switched to master3	12 days ago
.artanolis.yml	fix for path of tutorial pdfs	3 years ago

About

The COntstraint-Based Reconstruction and Analysis Toolbox. Documentation:

opencobra.github.io/cobratoolbox

tutorial metabolomics reconstruction transcriptomics cobra metabolic-models strain-engineering metabolic-reconstruction constraint-based-modeling microbiome-analysis metabolic-engineering gap-filling cobra-toolbox omics-data-integration human-metabolism

Readme

GPL-3.0 License

- ▶ Heirendt, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nature Protocols 14, 639 (2019).
- ▶ 421 citations (Google scholar, 28/9/21), ~1000 website visits/month, ~120 git clones/month.

# Log-concave sampling of polyhedral convex constraint-based models

- ▶ Uniform sampling of a polyhedral convex set

$$\Omega := \{Sv = 0, l \leq v \leq u\}.$$

# Log-concave sampling of polyhedral convex constraint-based models

- ▶ Uniform sampling of a polyhedral convex set

$$\Omega := \{Sv = 0, l \leq v \leq u\}.$$

- ▶ Sampling of certain log-concave functions over a polyhedral convex set

$$\mathcal{K} := \left\{ Sv = 0, \right. \\ \left. l \leq v \leq u, \right. \\ \left. v_j \propto \exp\left(-\sum f_j(v_i)\right) \right\},$$

where  $f_i(v_i)$  is a convex function.

# Log-concave sampling of polyhedral convex constraint-based models

- ▶ Uniform sampling of a polyhedral convex set

$$\Omega := \{Sv = 0, l \leq v \leq u\}.$$

- ▶ Sampling of certain log-concave functions over a polyhedral convex set

$$\mathcal{K} := \left\{ Sv = 0, \right. \\ \left. l \leq v \leq u, \right. \\ \left. v_j \propto \exp\left(-\sum f_j(v_i)\right) \right\},$$

where  $f_i(v_i)$  is a convex function.

- ▶ Think of a log-concave function as roughly equivalent to a unimodal function, e.g., probability density of a normal distribution

## Log-concave sampling of polyhedral convex constraint-based models

- ▶ Example: Biochemical reaction flux constraints from experimental data: mean  $\bar{v} \in \mathbb{R}^n$  ± covariance matrix  $\Sigma \in \mathbb{R}^{n \times n}$

$$\mathcal{K} := \left\{ S v = 0, \right. \\ \left. l \leq v \leq u, \right. \\ \left. v \propto \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(v - \bar{v})\Sigma^{-1}(v - \bar{v})\right)\right\},$$

# Riemannian Hamiltonian Monte Carlo sampling

ConstrainedSampler / PolytopeSamplerMatlab Public

<> Code Issues 6 Pull requests Actions Projects Wiki Security Insights

83e69fb652 2 branches 0 tags

File	Commit	Time
bin	update mex	2 months ago
code	typo	last month
coverage	bug fixes	2 months ago
.gitignore	bug fixes	2 months ago
LICENSE	Initial Commit	2 years ago
README.md	Update README.md	3 months ago
demo.m	Update demo.m	last month
InitSampler.m	fixing the program	3 months ago

## PolytopeSampler

PolytopeSampler is a `Matlab` implementation of constrained Hamiltonian Monte Carlo for sampling from high dimensional distributions on polytopes. It is able to sample efficiently from sets and distributions of more than 100K dimensions.

### Quick Tutorial

PolytopeSampler samples from distributions of the form  $\exp(-f(x))$ , for a convex function  $f$ , subject to constraints  $A_{\text{ineq}} * x \leq b_{\text{ineq}}$ ,  $A_{\text{eq}} * x = b_{\text{eq}}$  and  $\text{lb} \leq x \leq \text{ub}$ .

About: A matlab implementation for sampling log-concave distributions with polytope constraints. License: GPL-3.0 License.

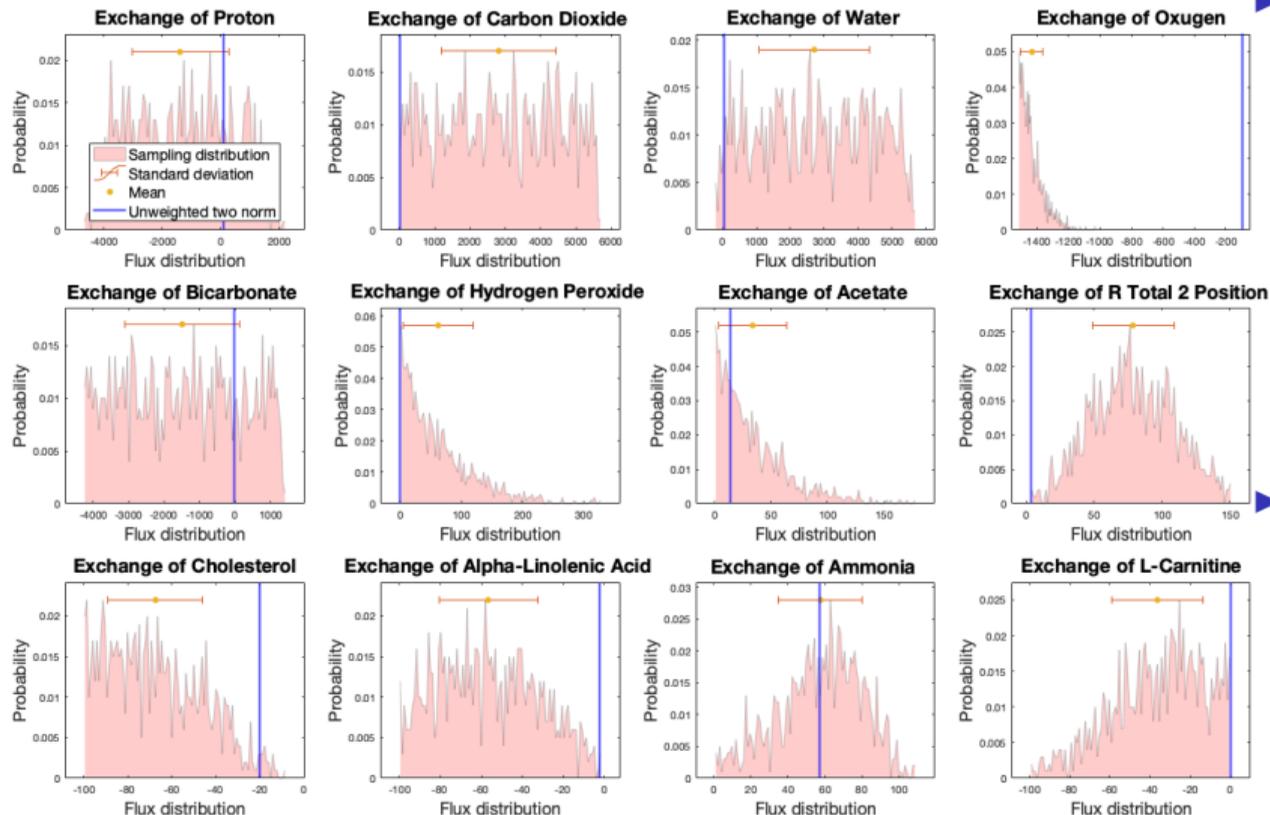
Contributors: YinTat, santoshv Santosh Vempala, ConstrainedSampler, ruoqishen Ruoqi Shen.

► Implementation: Ruoqi Shen, Yin Tat Lee, Santosh Vempala

► Lee, Y. T. & Vempala, S. S. Convergence Rate of Riemannian Hamiltonian Monte Carlo and Faster Polytope Volume Computation. arXiv:1710.06261 (2017).

► Santosh Vempala Sampling Convex Bodies: A Status Report, Simons Institute Workshop on Sampling Algorithms and Geometries on Probability Distributions, 2021

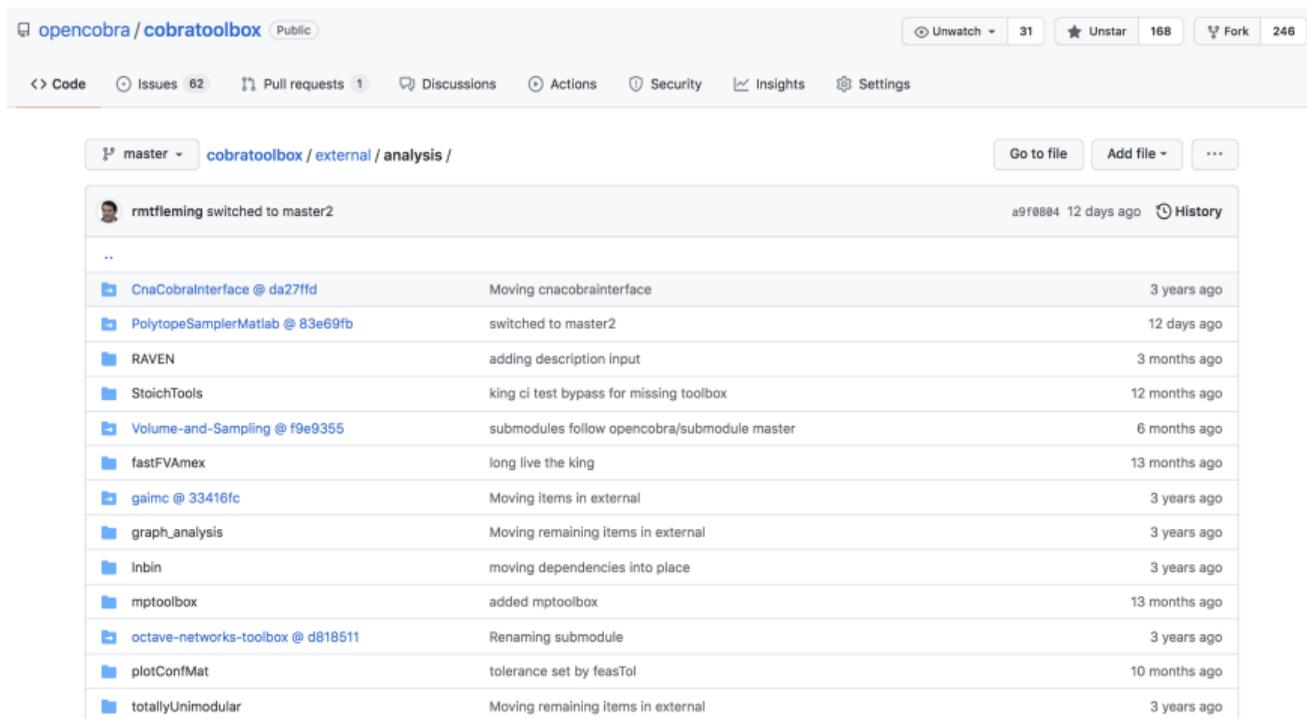
# Riemannian Hamiltonian Monte Carlo sampling



▶ Riemannian Hamiltonian Monte Carlo sample predictions compared with experimentally measured metabolite uptake and secretion.

▶ 1000 samples from ~900 dimensional model

# Riemannian Hamiltonian Monte Carlo sampling in COBRA Toolbox



opencobra / cobratoolbox Public

Unwatch 31 Unstar 168 Fork 246

Code Issues 62 Pull requests 1 Discussions Actions Security Insights Settings

master cobratoolbox / external / analysis /

Go to file Add file ...

rmtfleming switched to master2 a9f884 12 days ago History

..		
CnaCobrainterface @ da27ffd	Moving cnacobrainterface	3 years ago
PolytopeSamplerMatlab @ 83e69fb	switched to master2	12 days ago
RAVEN	adding description input	3 months ago
StoichTools	king ci test bypass for missing toolbox	12 months ago
Volume-and-Sampling @ f9e9355	submodules follow opencobra/submodule master	6 months ago
fastFVAmex	long live the king	13 months ago
gaimc @ 33416fc	Moving items in external	3 years ago
graph_analysis	Moving remaining items in external	3 years ago
Inbin	moving dependencies into place	3 years ago
mptoolbox	added mptoolbox	13 months ago
octave-networks-toolbox @ d818511	Renaming submodule	3 years ago
plotConfMat	tolerance set by feasTol	10 months ago
totallyUnimodular	Moving remaining items in external	3 years ago

- ▶ Lee, Y. T. & Vempala, S. S. Convergence Rate of Riemannian Hamiltonian Monte Carlo and Faster Polytope Volume Computation. arXiv:1710.06261 (2017).
- ▶ Santosh Vempala: Sampling Convex Bodies: A Status Report, Simons Institute Workshop on Sampling Algorithms and Geometries on Probability Distributions, 2021.

# COBRA Toolbox: shared interface to multiple sampling algorithms

opencobra / cobratoolbox Public

Unwatch 31 Unstar 168 Fork 246

<> Code Issues 62 Pull requests 1 Discussions Actions Security Insights Settings

Implementation:  
Ben Cousins,  
Hulda  
Haraldsdottir,  
Ruoqi Shen,  
Yin Tat Lee,  
Santosh  
Vempala,  
German  
Preciat, Ronan  
Fleming &  
others.

master cobratoolbox / src / analysis / sampling / sampleCbModel.m

YinTat Update sampleCbModel.m Latest commit f9b4ca9 on 7 Jul History

4 contributors

278 lines (250 sloc) | 9.99 KB

```
1 function [modelSampling,samples,volume] = sampleCbModel(model, sampleFile, samplerName, options, modelSampling)
2 % Samples the solution-space of a constraint-based model
3 %
4 % USAGE:
5 %
6 % [modelSampling, samples] = sampleCbModel(model, sampleFile, samplerName, options, modelSampling)
7 %
8 % INPUTS:
9 % model: COBRA model structure with fields
10 %     * .S - Stoichiometric matrix
11 %     * .b - Right hand side vector
12 %     * .lb - Lower bounds
13 %     * .ub - Upper bounds
14 %     * .C - 'k x n' matrix of additional inequality constraints
15 %     * .d - 'k x 1' rhs of the above constraints
16 %     * .dsense - 'k x 1' the sense of the above constraints ('L' or 'G')
17 %
18 % OPTIONAL INPUTS:
19 % sampleFile: File names for sampling output files (only implemented for ACHR)
20 % samplerName: {'CHRR'}, 'ACHR', 'RHMC' Name of the sampler to be used to
21 % sample the solution.
```

## Sampling challenge 4: FAIR software

- ▶ Findability, Accessibility, Interoperability, and Reuse (FAIR) of sampling software is essential to increase the impact of theoretical and computational sampling research. <https://www.go-fair.org/fair-principles/>

## Sampling challenge 4: FAIR software

- ▶ Findability, Accessibility, Interoperability, and Reuse (FAIR) of sampling software is essential to increase the impact of theoretical and computational sampling research. <https://www.go-fair.org/fair-principles/>
- ▶ Increasing the impact of theoretical and computational sampling research will facilitate greater investment in fundamental and applied research in this area.

## A disadvantage of flux balance analysis

- ▶ Even if we know the biochemical objective, i.e.,  $d \in \mathbb{R}^n$  in

$$\begin{aligned} & \underset{v \in \mathbb{R}^n}{\text{minimise}} && d^T v \\ & \text{s.t.} && Nv = b \\ & && l \leq v \leq u \end{aligned} \tag{FBA}$$

- ▶ Important constraints are missing, e.g., energy conservation, the second law of thermodynamics, etc.

## Biochemistry as a linear resistive network

- ▶ Assume that we are given a vector of resistances  $r \in \mathbb{R}_{++}^n$  and that Maxwell's minimum heat theorem is the variational principle underlying this network, that is

$$\begin{aligned} & \underset{v}{\text{minimise}} && \frac{1}{2} v^T \text{diag}(r) v \\ & \text{s.t.} && Nv = b \quad : y \end{aligned}$$

## Biochemistry as a linear resistive network

- ▶ Assume that we are given a vector of resistances  $r \in \mathbb{R}_{++}^n$  and that Maxwell's minimum heat theorem is the variational principle underlying this network, that is

$$\begin{aligned} & \underset{v}{\text{minimise}} && \frac{1}{2} v^T \text{diag}(r) v \\ & \text{s.t.} && Nv = b \quad : y \end{aligned} \tag{QP}$$

- ▶ The optimality conditions are

$$\begin{aligned} \text{diag}(1/r) N^T y^* &= v^* \\ Nv^* &= b \end{aligned}$$

## Biochemistry as a linear resistive network

- ▶ Assume that we are given a vector of resistances  $r \in \mathbb{R}_{++}^n$  and that Maxwell's minimum heat theorem is the variational principle underlying this network, that is

$$\begin{aligned} & \underset{v}{\text{minimise}} && \frac{1}{2} v^T \text{diag}(r) v \\ & \text{s.t.} && Nv = b \quad : y \end{aligned} \tag{QP}$$

- ▶ The optimality conditions are

$$\begin{aligned} \text{diag}(1/r) N^T y^* &= v^* \\ Nv^* &= b \end{aligned}$$

$$\Rightarrow N \text{diag}(1/r) N^T y^* = b$$

- ▶ Kirchhoff's current & voltage laws. Ohm's law: current linearly proportional to change in electrical potential.

## Biochemistry as a linear resistive network

- ▶ Assume that we are given a vector of resistances  $r \in \mathbb{R}_{++}^n$  and that Maxwell's minimum heat theorem is the variational principle underlying this network, that is

$$\begin{aligned} & \underset{v}{\text{minimise}} && \frac{1}{2} v^T \text{diag}(r) v \\ & \text{s.t.} && Nv = b \quad : y \end{aligned} \tag{QP}$$

- ▶ The optimality conditions are

$$\begin{aligned} \text{diag}(1/r) N^T y^* &= v^* \\ Nv^* &= b \end{aligned}$$

$$\Rightarrow N \text{diag}(1/r) N^T y^* = b$$

- ▶ Kirchhoff's current & voltage laws. Ohm's law: current linearly proportional to change in electrical potential.
- ▶ However,  $r$  is unknown, motivating efforts to sample the non-convex set of optimal solutions to **QP**.

## Sampling challenge 5: Interdisciplinary communication

- ▶ Several papers have been published in the biochemical literature that report algorithms and software for sampling non-convex feasible sets that are **broadly** of the form

$$\mathcal{J} := \{r \in \mathbb{R}_{++}^n, y \in \mathbb{R}^m \mid N \text{diag}(1/r) N^T y = b\}.$$

## Sampling challenge 5: Interdisciplinary communication

- ▶ Several papers have been published in the biochemical literature that report algorithms and software for sampling non-convex feasible sets that are **broadly** of the form

$$\mathcal{J} := \{r \in \mathbb{R}_{++}^n, y \in \mathbb{R}^m \mid N \text{diag}(1/r) N^T y = b\}.$$

- ▶ What is the relationship between these algorithms in the computational biology literature and the theoretical and computational results of the mathematics and computer science community?

## Sampling challenge 5: Interdisciplinary communication

- ▶ Several papers have been published in the biochemical literature that report algorithms and software for sampling non-convex feasible sets that are **broadly** of the form

$$\mathcal{J} := \{r \in \mathbb{R}_{++}^n, y \in \mathbb{R}^m \mid N \text{diag}(1/r) N^T y = b\}.$$

- ▶ What is the relationship between these algorithms in the computational biology literature and the theoretical and computational results of the mathematics and computer science community?
- ▶ For example, the following papers:
  - ▶ Gollub, M.G., Kaltenbach H.M., Stelling, J. **Probabilistic thermodynamic analysis of metabolic networks**, Bioinformatics, 2021, btab194.
  - ▶ Saldida, J. et al. **Unbiased metabolic flux inference through combined thermodynamic and  $^{13}\text{C}$  flux analysis** bioRxiv 2020.06.29.177063.
  - ▶ Pedro A. Saa, Lars K. Nielsen, **II-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks**, Bioinformatics, Volume 32, Issue 15, 1 August 2016, Pages 2330–2337.

## Sampling challenge 6: relationship to optimal solution sets?

- ▶ Every constraint-based modelling problem may be formulated as an optimisation problem
- ▶ Development of constraint-based modelling requires consideration of more general optimisation problems, e.g.,

$$\begin{aligned} & \underset{z \in \mathbb{R}^m}{\text{minimise}} && \phi(z) \\ & \text{s.t.} && f(z) = 0 \\ & && g(z) \leq 0 \end{aligned} \tag{1}$$

where  $\phi$  is a scalar valued continuous and convex function, and where  $f$  and  $g$  are vector valued functions.

- ▶ How do the established sampling algorithms map onto the sets of solutions to different classes of optimisation problems?

## Nonlinear resistive network

- ▶ Introduce unidirectional fluxes  $v_f, v_r \in \mathbb{R}_{\geq 0}^n$  such that  $v_f - v_r =: v$
- ▶ Maximise weighted linear sum of forward and reverse flux  $c^T(v_f + v_r)$
- ▶ Maximise entropy of unidirectional fluxes (Fleming, et. al. J. Theor. Biol. 292, 71–77 (2011)).

$$\begin{aligned} \underset{v_f, v_r > 0}{\text{minimise}} \quad & v_f^T \log(v_f) + v_r^T \log(v_r) + c^T(v_f + v_r) \\ \text{s.t.} \quad & Nv_f - Nv_r = b \\ & l \leq v_f - v_r \leq u \end{aligned}$$

## Nonlinear resistive network

- ▶ Introduce unidirectional fluxes  $v_f, v_r \in \mathbb{R}_{\geq 0}^n$  such that  $v_f - v_r =: v$
- ▶ Maximise weighted linear sum of forward and reverse flux  $c^T(v_f + v_r)$
- ▶ Maximise entropy of unidirectional fluxes (Fleming, et. al. J. Theor. Biol. 292, 71–77 (2011)).

$$\begin{aligned} & \underset{v_f, v_r > 0}{\text{minimise}} && v_f^T \log(v_f) + v_r^T \log(v_r) + c^T(v_f + v_r) \\ & \text{s.t.} && Nv_f - Nv_r = b \\ & && l \leq v_f - v_r \leq u \end{aligned} \tag{EP}$$

- ▶ Energy conservation, 2nd law of thermodynamics hold at optimal solution.

## Nonlinear resistive network

- ▶ Introduce unidirectional fluxes  $v_f, v_r \in \mathbb{R}_{\geq 0}^n$  such that  $v_f - v_r =: v$
- ▶ Maximise weighted linear sum of forward and reverse flux  $c^T(v_f + v_r)$
- ▶ Maximise entropy of unidirectional fluxes (Fleming, et. al. J. Theor. Biol. 292, 71–77 (2011)).

$$\begin{aligned} & \underset{v_f, v_r > 0}{\text{minimise}} && v_f^T \log(v_f) + v_r^T \log(v_r) + c^T(v_f + v_r) \\ & \text{s.t.} && Nv_f - Nv_r = b \\ & && l \leq v_f - v_r \leq u \end{aligned} \tag{EP}$$

- ▶ Energy conservation, 2nd law of thermodynamics hold at optimal solution.
- ▶ Information theory interpretation as the least biased prediction, given the data.

## Nonlinear resistive network

- ▶ Introduce unidirectional fluxes  $v_f, v_r \in \mathbb{R}_{\geq 0}^n$  such that  $v_f - v_r =: v$
- ▶ Maximise weighted linear sum of forward and reverse flux  $c^T(v_f + v_r)$
- ▶ Maximise entropy of unidirectional fluxes (Fleming, et. al. J. Theor. Biol. 292, 71–77 (2011)).

$$\begin{aligned} & \underset{v_f, v_r > 0}{\text{minimise}} && v_f^T \log(v_f) + v_r^T \log(v_r) + c^T(v_f + v_r) \\ & \text{s.t.} && Nv_f - Nv_r = b \\ & && l \leq v_f - v_r \leq u \end{aligned} \tag{EP}$$

- ▶ Energy conservation, 2nd law of thermodynamics hold at optimal solution.
- ▶ Information theory interpretation as the least biased prediction, given the data.
- ▶ However, again,  $c \in \mathbb{R}^n$  is a vector of free parameters. How to sample the set of optimal solutions?

# Sampling and entropy optimisation

Entropy maximisation

$$\begin{array}{ll} \underset{x>0}{\text{minimise}} & c^T x + x^T \log(x) \\ \text{s.t.} & Ax = b \end{array}$$

# Sampling and entropy optimisation

Entropy maximisation

$$\begin{array}{ll} \underset{x>0}{\text{minimise}} & c^T x + x^T \log(x) \\ \text{s.t.} & Ax = b \end{array} \quad (\text{EP})$$

may be reformulated as a linear maximisation problem, subject to an intersection of linear and exponential cone constraints

$$\begin{array}{ll} \underset{x>0}{\text{minimise}} & c^T x - 1^T t \\ \text{s.t.} & Ax = b \\ & \begin{bmatrix} 1 \\ x \\ t \end{bmatrix} \in \mathcal{K}_{\text{exp}} \end{array}$$

# Sampling and entropy optimisation

Entropy maximisation

$$\begin{array}{ll} \underset{x>0}{\text{minimise}} & c^T x + x^T \log(x) \\ \text{s.t.} & Ax = b \end{array} \quad (\text{EP})$$

may be reformulated as a linear maximisation problem, subject to an intersection of linear and exponential cone constraints

$$\begin{array}{ll} \underset{x>0}{\text{minimise}} & c^T x - 1^T t \\ \text{s.t.} & Ax = b \\ & \begin{bmatrix} 1 \\ x \\ t \end{bmatrix} \in \mathcal{K}_{exp} \end{array} \quad (\text{EXP})$$

where  $\mathcal{K}_{exp}$  denotes a set of  $n$  exponential cones. For  $p, q, r \in \mathbb{R}$

$\mathcal{K}_{exp} := \{p, q, r \mid p \geq q \exp\left(\frac{r}{q}\right), q > 0\}$ . Can the set of solutions to this optimisation problem, as a function of a convex and compact set of parameters  $c$ , be sampled?

# Summary

- ▶ Constraint-based modelling of biochemical networks provides a strong demand for sampling algorithms and a host of challenges
  1. Intrinsic anisotropy
  2. High dimensionality
  3. Convergence criteria
  4. FAIR software
  5. Interdisciplinary communication
  6. Feasible sampling problems and parametric solution sets of optimisation problems

# Acknowledgements

- ▶ Santosh Vempala, Ben Cousins, Aditi Laddha, Georgia Institute of Technology.
- ▶ Yin Tat Lee, University of Washington.
- ▶ Ines Thiele, Molecular Systems Physiology Group, National University of Ireland, Galway.
- ▶ Michael Saunders, Stanford Systems Optimization Laboratory.
- ▶ Hulda Haraldsdottir, German Preciat, Systems Biochemistry Group.