Data integration

Tyler M. Earnest

July 19, 2018

Hands-On Workshop on Cell Scale Simulations, Urbana, IL

Introduction

- Not *ab initio*!
- Required data
  - Reactions
  - Rate parameters
  - Diffusion coefficients
  - Geometry

Common data sources

Reaction rate and diffusion coefficients

- Experiment
- Literature (measurements, published model parameters, etc.)
- Bionumbers
- BRENDA
- KEGG

Common data sources

Geometry

- Idealized
  - Experiment
  - Bionumbers
  - Literature
- Real
  - 3D optical microscopy
  - Cryo-electron tomography

Bionumbers

- `http://bionumbers.hms.harvard.edu/`
- Developed in 2007 by Ron Milo, Paul Jorgensen and Mike Springer[1]
- Database of biologically interesting numbers from the literature

[1] R. Milo *et al.*, *Nucleic Acids Research* **38**, D750–D753 (2009).

Bionumbers

- Each entry contains
  - Title
  - Value or range of values and units
  - Organism
  - Reference
  - Method
  - Bionumbers accession number

Bionumbers

Example

`http://bionumbers.hms.harvard.edu/bionumber.aspx?id=104324`

Bionumbers

- Generally trustworthy
- But, no programmatic access

BRENDA

- `https://www.brenda-enzymes.org/`
- Started in 1987 at the German National Research Centre for Biotechnology in Braunschweig (GBF), continued at the University of Cologne, and is now curated and hosted at the Technical University of Braunschweig, Institute of Biochemistry and Bioinformatics.[2]
- Database of enzymatic data indexed by EC number

[2] S. Placzek *et al.*, *Nucleic Acids Research* **45**, D380–D388 (2016).

BRENDA

- Available Data
  - Michaelis-Menton parameters: K_{M}, k_{cat}, etc.
  - Inhibitor parameters: K_{I}, IC_{50}, etc.
  - Temperature and pH ranges
  - Isoelectric point
- Parameters given for organism and substrate

BRENDA

Example

`https://www.brenda-enzymes.org/enzyme.php?ecno=2.2.1.1`

BRENDA

- Need to critically evaluate each parameter value (typos exist)
  - Check primary reference if given.
- Programmatic access available
  - SOAP $\rightarrow$ Use SOAPpy

Programmatic access: SOAP

```
from SOAPpy import SOAPProxy
import hashlib

brenda = SOAPProxy("http://www.brenda-enzymes.org/soap/brenda_server.php")
username = "the_username"
password = hashlib.sha256("the_password").hexdigest()

print(brenda.getKmValue("%s,%s,ecNumber*2.2.1.1#organism*Escherichia coli"
                        % (username=username, password=password)))
```

The result will be delimited by #, !, and *.

NOTE Only works with `SOAPpy` on Python 2.7. Other Python SOAP implementations do not work!

SABIO-RK

- `http://sabio.h-its.org/`
- SABIO-RK is a curated database that contains information about biochemical reactions, their kinetic rate equations with parameters and experimental conditions.[3]

[3] U. Wittig *et al.*, *Nucleic Acids Research* **40**, D790–D796 (2011).

SABIO-RK

Example

- `http://sabiork.h-its.org/newSearch?q=sabioreactionid:1113`

Programmatic access: REST

```
import requests

request = requests.get(
        'http://sabiork.h-its.org/sabioRestWebServices/searchKineticLaws/entryIDs',
        params={"q": 'ECNumber:"2.7.1.11"'
                     ' AND Organism:"Escherichia coli"'
                     ' AND Parametertype:"Vmax"',
                "format": 'txt'})
ids = [int(x) for x in request.text.strip().split('\n')]
request = requests.post(
        'http://sabiork.h-its.org/entry/exportToExcelCustomizable',
        params={'format': 'tsv',
                'fields[]': ['Parametertype', 'DateSubmitted',
                             'PubMedID', 'Parameter']},
        data={'entryIDs[]': ids})
print(request.text)
```

Bioservices

- http://bioservices.readthedocs.io/en/master/
- Programmatic access to over 30 online databases

```
from bioservices import KEGG
s = KEGG()
print(s.get("hsa:7535"))
```

Estimating parameters

Diffusion limited reactions

$$\mathrm{A} + \mathrm{B} \xrightarrow{k_{\mathrm{DL}}} \mathrm{C}$$

$$k_{\mathrm{DL}} \approx 4\pi(D_{\mathrm{A}} + D_{\mathrm{B}})(r_{\mathrm{A}} + r_{\mathrm{B}})N_{\mathrm{A}}$$

Rule of thumb

$$k_{\mathrm{DL}} \approx 10^{9}\ \mathrm{L} \cdot \mathrm{mol}^{-1} \cdot \mathrm{s}^{-1}$$

Diffusion coefficients

Diffusion slower in cytosol

- Small molecules: $\frac{D_{\mathrm{cyt}}}{D_{\mathrm{H_2O}}} \approx 0.3$
- Average protein: $\frac{D_{\mathrm{cyt}}}{D_{\mathrm{H_2O}}} \approx 0.03$

Diffusion coefficients

Estimate for *E. coli*:[4]

$$\ln \frac{D_{\mathrm{H_2O}}}{D_{\mathrm{cyt}}} = \ln \frac{\eta_{\mathrm{cyt}}}{\eta_{\mathrm{H_2O}}} = \left(\frac{\xi^2}{{R_{\mathrm{H}}}^2 + {r_{\mathrm{HR}}}^2} \right)^{-a/2}$$

Fit parameters:

$$\xi = 0.51 \ \pm 0.09 \ \mathrm{nm}$$
$$R_{\mathrm{h}} = 42 \ \pm 9 \ \mathrm{nm}$$
$$a = 0.53 \ \pm 0.04$$

[4] T. Kalwarczyk *et al.*, *Bioinformatics* **28**, 2971–2978 (2012).

Hydrodynamic radii

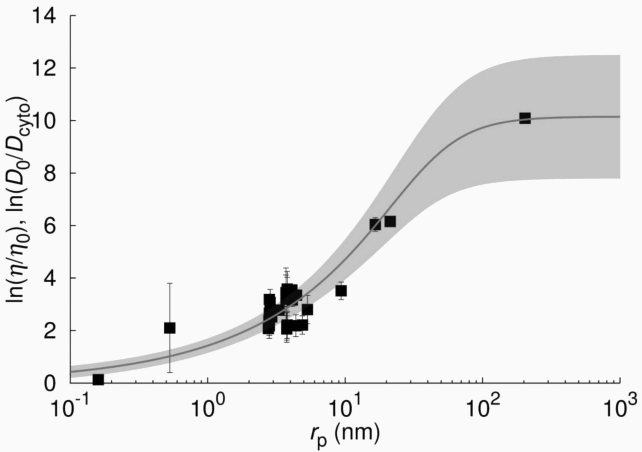
$$r_{\mathrm{HR}} \approx A \left(\frac{M_{\mathrm{W}}}{\mathrm{Da}} \right)^{\alpha}$$

$$D_{\mathrm{H_2O}} \approx \frac{k_{\mathrm{B}} T}{6\pi\, \eta_{\mathrm{H_2O}}\, r_{\mathrm{HD}}}$$

Type	A/nm	α
Protein[5]	0.0515	0.392
RNA[5]	0.0566	0.38
DNA (linear)[6]	0.024	0.57
DNA (circular)[6]	0.0125	0.59

[5] K. A. Dill *et al.*, *Proceedings of the National Academy of Sciences* **108**, 17876–17882 (2011).
[6] R. M. Robertson *et al.*, *Proceedings of the National Academy of Sciences* **103**, 7310–7314 (2006).

Diffusion coefficients



[7] T. Kalwarczyk *et al.*, *Bioinformatics* **28**, 2971–2978 (2012).

Fitting

Rate coefficient data can be estimated by fitting your model to experimental data

In many cases, an acceptable estimate can be made by fitting to a deterministic, well-stirred model.

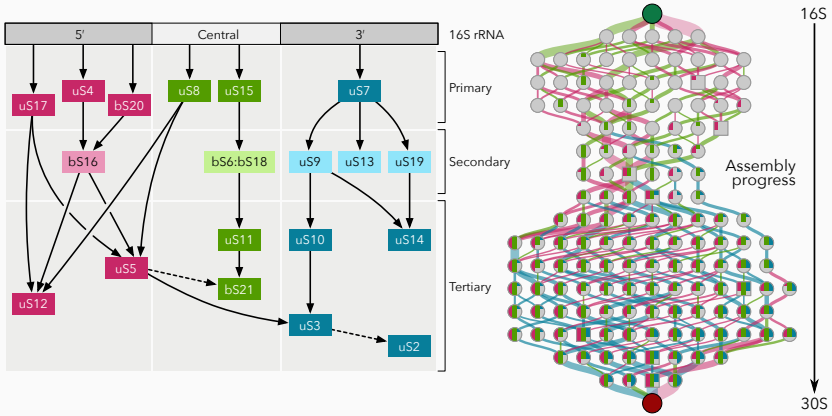
Fitting

The experimental data does not have to be concentration vs. time

- Any quantity predicted by the model can be used to construct an objective function
- Ill-posed problems, regularization

Example 1

Assembly of the ribosomal small subunit[7]



[8]T. M. Earnest *et al.*, *Biophysical Journal* **109**, 1117–1135 (2015).

Example 1

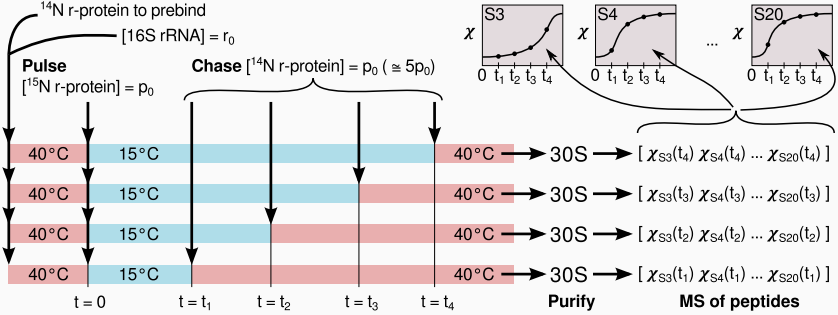
Assembly reactions

$$\mathrm{P}_i + \mathrm{I}_a \xrightarrow{k_i} \mathrm{I}_b$$

17 SSU protein types, one rate coefficient per protein

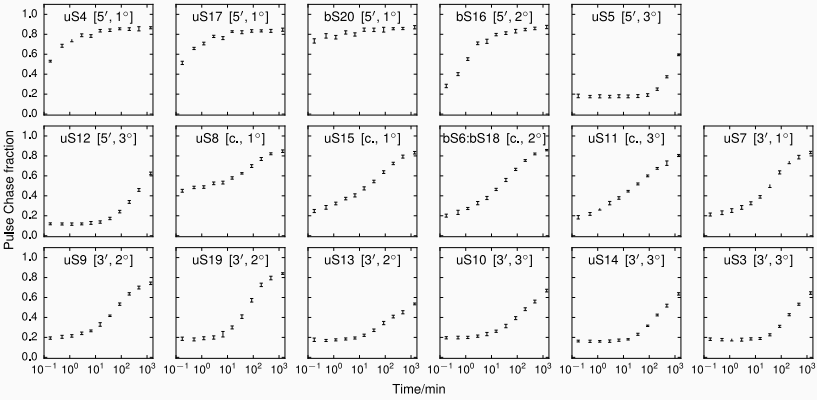
Example 1

Experimental data



Example 1

Experimental data



Example 1

How is this data related to the abundance of intermediates predicted by the model?

Is it simply:

$$\chi_i = \frac{\sum (\text{conc. of intermediates with protein } i)}{\sum (\text{conc. of all intermediates})}$$

Example 1

No: it is a more complicated function which must account for the exact details of the experiment

$$\chi_i(t) = \frac{p_i^{\mathrm{P}}}{p_i^{\mathrm{C}} + p_i^{\mathrm{P}}} + \frac{p_i^{\mathrm{C}}(p_i^{\mathrm{C}} - r + p_i^{\mathrm{P}})}{r(p_i^{\mathrm{C}} + p_i^{\mathrm{P}})} \left(\frac{p_i^{\mathrm{P}} - p_i(t)}{p_i^{\mathrm{C}} + p_i(t)} \right),$$

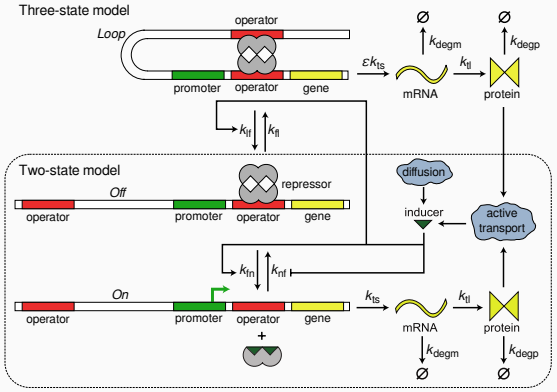
- r – Initial concentration of ribosomal RNA
- p_i^{P} – Initial concentration of labeled protein (pulse)
- p_i^{C} – Initial oncentration of unlabeled protein (chase)

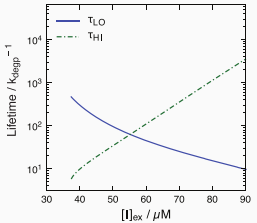
This function is what should be used to fit the data: minimize the squared deviation

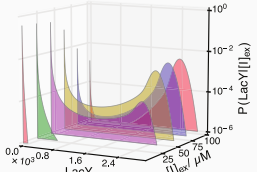
$$f(\{k_i\}) = \sum_i \sum_j \left[\chi_i(t_j) - \chi_{ij}^{\mathrm{expt}} \right]^2$$

Example 2

Three-state bistable switch[8]







[9]T. M. Earnest *et al.*, *Physical Biology* **10**, 026002 (2013).

Example 2

- 5 free parameters
- 17 parameters from experiment
- Behavior of interest is *stochastic*!
  - Simulation execution time is slow
- Experimental data: Bistability range
  - Only two numbers

Example 2

No fitting

- Fitting doesn't make sense
- Instead explore parameter space
  - Randomly sample parameters from a uniform distribution
  - Accept parameters which recover the range of bistability

Example 2

Sensitivity analysis