Analysing enzyme databases with Stan and Arviz

Teddy Groves (Joint work with Areti Tsigkinopoulou)

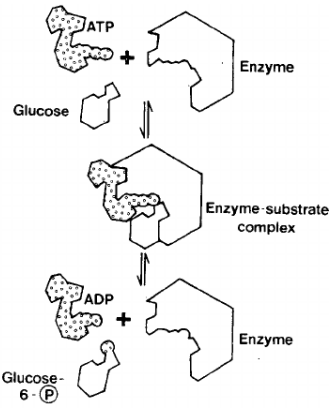
- Introduction
- What we did
- Results
- Final thoughts
- Thanks for listening!
- References

Introduction

The good news

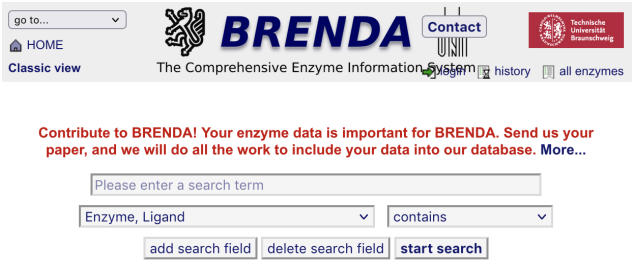
Since ~1920 we can parameterise enzyme-catalysed reactions.

$$v = E \cdot \frac{k_+^{cat} \frac{GLC}{k_{GLC}^M} \frac{ATP}{k_{ATP}^M} - k_-^{cat} \frac{G6P}{k_{G6P}^M} \frac{ADP}{k_{ADP}^M}}{(1 + \frac{GLC}{k_{GLC}^M})(1 + \frac{ATP}{k_{ATP}^M}) + (1 + \frac{G6P}{k_{G6P}^M})(1 + \frac{ADP}{k_{ADP}^M}) - 1}$$



More good news

Kinetic parameter measurements are available online.



The bad news

- Disagreement between studies
- Relevant information can be inaccessible
- Many parameters have not been measured

Result: kinetic modelling step zero

1. Manually trawl the database
2. Read the actual papers
3. Decide how to resolve disagreements

This is annoying!

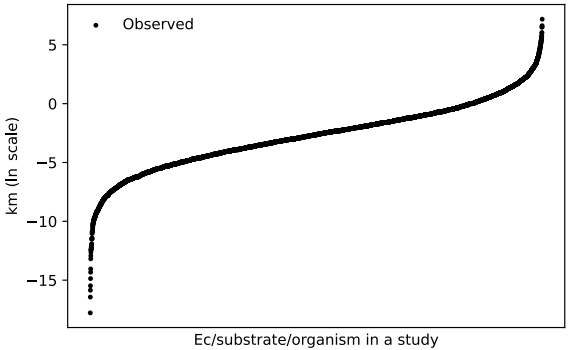
Our plan

1. Summarise the km information in BRENDA with a statistical model.
2. Make the results easy to access

What we did

Fetch and process some data

- One ’observation’ per study
- Three organisms (Human, yeast, E. coli)
- Natural enzyme/substrate combinations only



Try out some models

We settled on this measurement model from Borger, Liebermeister, and Klipp (2006):

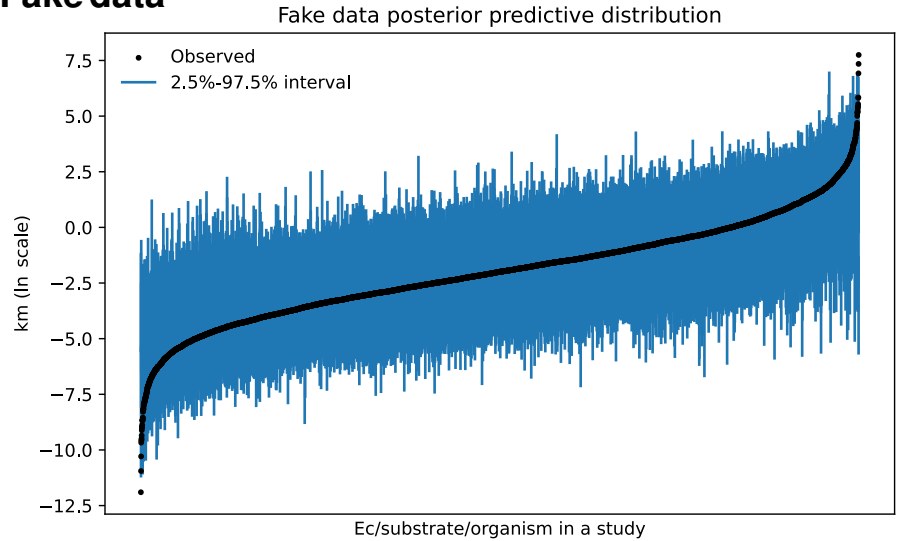
$$\ln \hat{km}_{ijk} = \mu + \alpha_k^{sub} + \alpha_{jk}^{org:sub} + \alpha_{ik}^{ec:sub}$$

Write the model in Stan

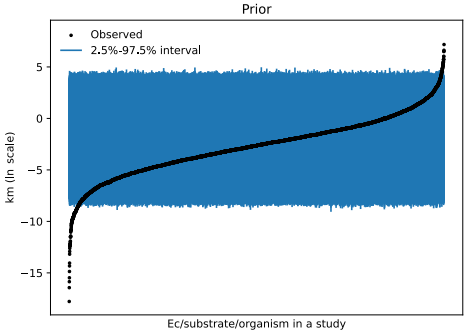
```
model {
  if (likelihood){
    y_train ~ student_t(nu, log_km[biology_train], sigma);
  }
  nu ~ gamma(2, 0.1);
  sigma ~ normal(0, 2);
  mu ~ normal(-2, 1);
  a_sub ~ normal(0, tau_sub);
  a_ec_sub ~ normal(0, tau_ec_sub);
  a_org_sub ~ normal(0, tau_org_sub);
  tau_org_sub ~ normal(0, 1);
  tau_ec_sub ~ normal(0, 1);
  tau_sub ~ normal(0, 1);
}
```

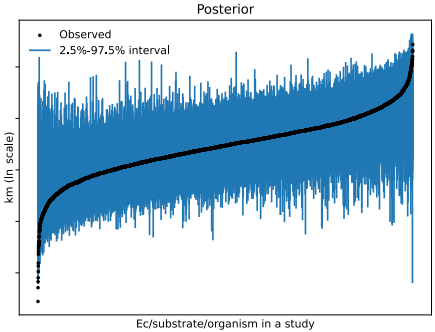
Results

Fake data


Fake data posterior predictive distribution
Observed
2.5%-97.5% interval
km (ln scale)
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
Ec/substrate/organism in a study

Real data


Prior
Observed
2.5%-97.5% interval
5
0
−5
−10
−15
km (ln scale)
Ec/substrate/organism in a study


Posterior
Observed
2.5%-97.5% interval
km (ln scale)
Ec/substrate/organism in a study

Approximate cross-validation

```
         Estimate       SE
elpd_loo -12549.28    86.78
p_loo     2681.82        -

There has been a warning during the calculation. Please check the res
------

Pareto k diagnostic values:
                         Count   Pct.
(-Inf, 0.5]   (good)     5914   92.6%
 (0.5, 0.7]   (ok)        407    6.4%
   (0.7, 1]   (bad)        65    1.0%
   (1, Inf)   (very bad)    3    0.0%
```

Exact cross-validation

```
: calculating out of sample log likelihoods for model blk...
:   split_4: -1477.2729990849998
:   split_3: -1481.00882687375
:   split_2: -1441.2448831225
:   split_5: -1477.1304594875
:   split_0: -1467.7910366725
:   split_7: -1480.53716923375
:   split_9: -1504.9884546787503
:   split_8: -1502.4060516825002
:   split_6: -1483.4119273775
:   split_1: -1515.77739092
:   total out of sample log likelihood: -14831.569199133752
```

Webapp

Choose which marginal posterior distribution you'd like to see!

Organism

Escherichia coli

EC4

1.1.1.169

Substrate

2-dehydropantoate

Download a csv table of model results.

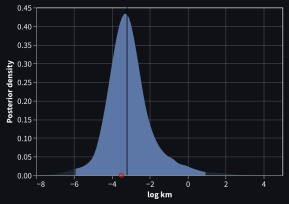
Download a csv table of measurements.

See here for the model's source code and for instructions to reproduce the full posterior distribution.

What does BRENDA say?

This webapp shows marginal posterior distributions for Km parameters from a model we trained on data from the BRENDA database.

The graph shows the posterior distribution (blue) and experimental measurements (red).



Marginal posterior summary:

```
Mean: -3.0908038874325

Standard deviation: 1.23

1% quantile: -5.99

Median: -3.20807

99% quantile: 1.0
```

Below is a table of all the measurements.

	reference	km	log km
0	687675	0.03	-3.506558

Final thoughts

Final thoughts

- This data is tricky to model: plenty of room for improvement.
- File-based cmdstanpy workflow is great!
- `llik` + `yrep` = long io
- cookiecutter-cmdstanpy: not yet but maybe some day

Thanks for listening!

Thanks for listening!

Our emails:

- tedgro AT biosustain.dtu.dk
- aretsi AT biosustain.dtu.dk

The project on github: https://github.com/biosustain/brenda_km

novo nordisk fonden

References

References I

.

Borger, Simon, Wolfram Liebermeister, and Edda Klipp. 2006. “Prediction of Enzyme Kinetic Parameters Based on Statistical Learning.” *Genome Informatics* 17 (1): 80–87.