Statistical analysis of kinetic parameter data from the BRENDA database

This paper presents a range of multilevel Bayesian linear models describing kinetic parameter data from the online database BRENDA. We assess the models' qualitative fit to the available data and quantitatively compare their out-of-sample predictive performance. The best model is shown to compare favourably with the current state of the art, without needing to use costly external features such as chemical fingerprints. We discuss how the results of our analysis can be integrated into a kinetic modelling framework.

Methods

Software

All the software that we used to produce the results in this paper, as well as instructions for reproducing our analysis, can be found at https://github.com/biosustain/brenda_km.

Data Acquisition and filtering

We used the SOAP API provided by BRENDA to fetch a table of all available Km parameter measurements, as well as a table with information about natural substrates. We then discared rows where the organism, EC4 number, substrate were not recorded.

Statistical model

In order to aggregate the information about km parameters in the BRENDA database we use a multilevel Bayesian regression model with a nested structure. In this model the response variables are km values, which we assume are measured such that for enzyme j, substrate k and organism l

$$y_{jkl} = \ln km_{jkl} \sim T4(\hat{y_{jkl}}, \sigma) \tag{1}$$

where $\hat{y_{jkl}}$ is the estemated true km value on natural logarithmic scale and σ is an unknown positive number. $T4$ represents the student t distribution with four degrees of freedom.

We assume that the true log km has the following structural dependency on latent variables:

$$\hat{y}_{jkl} = \mu + \alpha_j^{EC3} + \alpha_j^{EC4} + \beta^{nat} * nat(j,k,l) + \epsilon_{jkl} \quad (2)$$

In this equation: - μ is a single number representing the global average - α^{EC3} is an EC3-specific effect - α^{EC4} is an EC4-specific effect - β^{nat} is a single number representing the effect of a substrate/organism/enzyme combination being natural - *nat* is a function indicating whether and substrate/organism/enzyme combination is natural - ϵ is a substrate/organism/enzyme-specific effect

The prior distributions for the latent parameters α^{EC3} and α^{EC4} have the following nested hierarchical structure:

$$\alpha^{EC3} \sim T4(0, \tau) \quad (3)$$
$$\alpha_j^{EC4} \sim T4(0, \tau_{EC3(j)}^{EC3} \quad (4)$$

where τ is a single number representing the variation of EC3 effects and τ^{EC3} is an EC3-specific number representing the variation of EC4 effects within each EC3 group.

The parameters ϵ have hierarchical priors at the organism level:

$$\epsilon_{jkl} \sim T4(0, \tau_l^{\epsilon}) \quad (5)$$

To complete our model we use informative priors for the remaining unknown model parameters:

Parameter	Distribution	1% prior quantile	99% prior quantile
σ	Log normal	0.4	3
μ	Normal	-3	1
τ	Log normal	0.4	2.5
τ^{EC3}	Log normal	0.02	4
β^{nat}	Normal	-1	0
τ^{ϵ}	Log normal	0.2	0.9

Reasoning behind modelling choices

The nested model design is motivated by our assessment of the available information. We judged that differences from the baseline in the km measurements are conditionally exchangeable given the same ec4 number, and that ec4 effects with the same ec3 number would be similar.

We used the student-t distribution with four degrees of freedom to model mea-

surement errors and for our group-level regression models because the BRENDA data is known to contain many outliers. Since the t4 distribution has heavy tails compared to the normal distribution, we expected that this distribution would make the model less sensitive to outliers, resulting in better predictive performance. Testing confirmed this: models that used the t4 distribution resulted in better out-of-sample predictive accuracy than an equivalent models with normal-distirbuted errors [refer to appendix].

The parameter β^{nat} was introduced to account for substrate-specific variation. It would have been preferable to include more information about each enzyme/substrate combination - for example based on the match between the two molecules' shape. However, it is not currently straightforward to match substrate identifiers from BRENDA with other identifiers for which detailed chemical information is available.

The parameter ϵ is intended to account for organism-specific variation. We judged that there are unlikely to be noticeable systematic organism-level effects (for example the Km parameters for one organism tending to be higher or lower than for another), but that these are nonetheless likely to be differences between the kinetic parameters of enzyme/substrate combinations depending on organism. We assume that there is little correlation between these deviations and the other parameters in our model. For example, if a certain EC3 number is associated with high Km values, we do not expect that this EC3 number will also tend to have higher or lower organism-specific variation.

Results

Discussion