MP3 - Using Big Data to Predict the Behaviour of Enzymes

Supervised by Geraint Thomas & Kevin Bryson

Word count: 4853

Ben K. Margetts

Abstract

Parameter inference through the use of bayesian prediction methods, is highly dependent on the quality of the prior distributions that describe the dataset. Using enzyme kinetic parameter data collected from the Sabio-RK database, a variety of methods were trialled for grouping and learning these data, with the goal being to produce high quality prior distributions for any enzyme of interest. Each level of the enzyme commission hierarchy was investigated, using a variety of statistical tests to determine which level provided the best trade off between information loss, and accuracy of kinetic parameter prediction. From our analyses, we have shown that this appears to be at the second level of the hierarchy (n.n.*.*). Our analyses also suggested that there will be very little difference in accuracy between the various levels of the hierarchy. The use of a k-nearest neighbours classifier was also trialled for predicting kinetic parameters from the database. This approach to machine-learning was ineffective, demonstrating an interesting lack of association between sub-levels of the hierarchy and their respective parameters, suggesting that sub-subgroups of the hierarchy are poor predictors of key enzyme kinetic parameters.

1 Introduction

Throughout biology, enzymes are considered to be essential components of metabolic systems [1, 2, 3]. They are biological catalysts; accelerating the rate at which molecules, know as substrates, are converted into variants, known as products, via chemical reactions [4].

Enzymes are a type of protein that catalyse reactions in a multitude of ways, with the specific goal of these methods being to lower the activation energy required to initialise the reaction [5]. For this to occur, an enzyme first needs to bind the substrate. Enzymes are usually highly specific for particular substrates through the utilisation of unique binding sites known as active sites. These active sites enforce specificity on bound substrates through shape, charge, and hydrophillic/hydrophobic preference. It is in these active sites in-which the chemical reaction takes place, forming the product [4]. These reactions can be represented by the following general equation:

$$E + S \rightleftharpoons ES \rightarrow EP \rightleftharpoons E + P$$

where E represents an enzyme, S represents the substrate, and P represents the product.

Each specific type of enzyme has a unique three-dimensional structure associated with it. This structure is determined by three key variables: the chain of amino acids that make-up the protein, the structures (α-helices and β-sheets) formed by amide (NH_2) and carboxyl (COOH) group interactions within this chain, and foldings that occur from α-helix and β-sheet interactions. This structure is essential for informing the function of the enzyme [6], and determining an enzyme's affinity for its substrate, along with its rate of catalysis. Given this, it is therefore inviting to use enzymatic structural data as a predictor of functional parameters. Three of these key values can immediately be suggested as targets for this prediction: V_{max}, which gives the maximal velocity of a reaction at a point where the substrate concentration is high enough to saturate all active sites, the Michaelis Constant or K_m, which describes substrate concentration at which half of the enzyme's active sites are occupied by substrate, and the turnover rate or k_{cat}, which gives the number of substrate molecules each active site converts to product per unit of time.

The efficiency of these enzymes can be highly variable when comparing one specific type of enzyme to the next and is largely dependent on the specific enzyme's structure and optimal working-conditions[7]. This variation in efficiency is often misrepresented within the literature base, with papers focusing on highly efficient enzymes over the

unrepresented 'average' enzyme [8].
When classifying enzymes, enzyme commission (EC) numbers are the foremost system used within biosciences. This nomenclature scheme was first published in 1961[9], and is now on the 6th version. The basic structure of an EC number takes on the form *.*.*.*, where * represents all values in a category, and n represents a specific value. Within the EC hierarchy, n.*.*.* donates the uppermost group of enzymes, n.n.*.* donates a subcategory (i.e. what bonds the enzyme acts on), n.n.n.* donates a further sub-subcategory of the enzymes, and n.n.n.n gives the final subclassifications of the enzymes and refers to a specific reaction that is catalysed[10]. All enzymes that catalyse a specific reaction are returned by the EC number, resulting in an effective method for grouping enzymes based upon their function.

Parametrising the aforementioned key-values represents a current challenge within biology. Measuring these parameters experimentally for an unknown enzyme is hindered by several limitations. It is limited by the equipment available to the researcher, by the time that it takes to complete the necessary experiments, and by the cost of doing so. This prohibitive environment surrounding parametrisation often leads to investigators searching the literature for these parameters, which can result in introducing other unexpected errors. For enzymes where this information exists, it is often parametrised in an unrepresentative environment. For example, if a human enzyme is parametrised *ex vivo*, it is often under ideal conditions which can result in unrepresentative parameter estimates, and for many enzymes, this information is not yet available.

Where these parameters are correctly estimated, they allow us to model the behaviour of enzymes through a selection of well-established models, such as the Michaelis-Menten kinetics model. This commonly-used equation considers the following irreversible reaction system:

$$E+S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} P+S \Longrightarrow v = \frac{d[P]}{dt} = \frac{V_{max}[S]}{K_m + [S]}$$

where:

$$K_m = \frac{(k_{cat} + k_{-1})}{k_1}$$
$$V_{max} = k_{cat}.e_0$$

Here k_1 and k_{-1} give their respective association/dissociation constants, and e_0 gives the total concentration of enzyme in the system.

The key problem with modelling enzymatic behaviour is, as eluded to earlier, the reliance on accurate parameter estimation for descriptors of enzyme function, i.e. K_m. When first investigating an unknown enzyme, accurate information on the enzyme's complete three-dimensional structure may be unavailable, or it may be difficult to experimentally measure parameters for it. One solution for this issue is to utilise bioinformatics tools for predicting these parameters when access to enzyme-specific kinetic data is limited. These predictors are often Bayesian in nature and utilise an incarnation of Bayes' theorem relating conditional probability of a given parameter value θ given data D:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

where $p(\theta|D)$ represents the posterior distribution, $p(D|\theta)$ the likelihood function, $p(\theta)$ the prior distribution, and $p(D)$ the evidence.

As can be seen above, predictors that utilise this variation on Baye's theorem rely on accurate prior distributions for producing representative posterior distributions, yet the choice of these priors is too often arbitrary. Uniform distributions across the range of potential values are commonly used as substitutes for accurate and descriptive prior distributions. One of the most promising predictive methodologies, approximate Bayesian computation (ABC)[11], is reliant on these priors, yet there is currently a lack of tools to approach this issue with.

ABC refers to a set of computational methodologies that utilise Bayesian statistics to predict values within complex systems, with ABC rejection sampling being the most commonly used of these methodologies[12]. This algorithm works by sampling points from a prior distribution [13], where given a sampled parameter point θ, a data set, $\hat{D}$, is then simulated under the chosen statistical model M, specified by the sampled θ. If the generated $\hat{D}$ is too different from the observed data D, then θ is rejected, and a new θ generated. $\hat{D}$ is accepted with a user defined tolerance that meets the requirement $\epsilon \geq 0$ if:

$$p(\hat{D}, D) \leq \epsilon$$

where the distance $p(\hat{D}, D)$ determines the discrepancy between $\hat{D}$ and D, assuming use of the Euclidean distance metric, where the distance between two points a and b is the length of the line connecting them ($\overline{ab}$).

While the algorithm is computed, D is replaced with a set of lower dimension summary statistics $S(D)$, which are selected to describe all relevant information in D. Given this alteration, the acceptance criterion becomes:

$$p(S(\hat{D}), S(D)) \leq \epsilon$$

For example, with an enzyme dataset selected, the observed dataset μ is given a prior distribution θ, from which a series of samples are taken $\theta_1, \theta_2, \theta_3 ... \theta_n$. Given the model M, n simulations are then performed using the sampled prior distribution θ_i, in which a summary statistic μ_i is computed for each simulation such that the following statement is either accepted or rejected:

$$p(u_i, u) \overset{?}{\leq} \epsilon$$

where based on the tolerance ϵ and the distance $p(\cdot, \cdot)$, the summary statistic is checked against the observed data to see how close it is. The posterior distribution of θ is then approximated from the distribution of accepted parameter values θ_i.

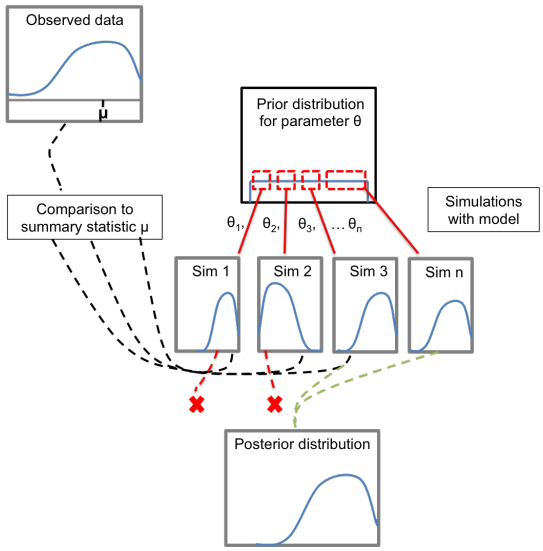

Figure 1: Visualisation of the ABC rejection algorithm.

It is therefore apparent that a focus on developing realistic prior distributions for enzyme kinetic measurements has clear potential benefits. If open-access data is combined with the enzymatic structural information available to the researcher, than producing these realistic distributions becomes a very real possibility. Previously, a log-normal distribution has been shown to fit each branch of the first level of EC kinetic measurements (n.*.*.*), where the log-normal probability density function is given as:

$$f(x, u, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnx-u)^2}{2\sigma^2}}$$

This distribution, although significantly more representative than a uniform distribution over the potential range of kinetic values, only describes and utilises the very upper level of enzyme functional/structural differences to separate them into categories. This begs the question, can we investigate deeper into the EC classification system to produce prior distributions that are more representative of a specific enzyme in question?

Through utilising the information currently available in enzyme kinetic databases such as Sabio-RK, it may be possible to produce realistic priors with high specificity for any enzyme that is needed, resulting in a personalised prior distribution for each specific unknown entity. This undertaking would therefore have the potential to allow quantitatively minded biologists to model the behaviour of their enzymes before any large body of experimental work has been completed, significantly reducing potential associated financial and time limitations.

2 Materials & Methods

Retrieval of values from the Sabio-RK database was achieved with Python (version 3.3) scripts, calling the RESTful interface. Analysis of the data was achieved with Matlab (version R2014b), Python (version 2.7) using the Pandas module, Mathematica (version 10.0.1.0), and R (version 3.1.2). Machine learning scripts were written using the Python Scikit-learn, Numpy, and Scipy modules.

Databases used for information retrieval include the Sabio-RK database (open access, available from http://sabio.villa-bosch.de/) for enzyme kinetic parameter measurements, the UniProt database (open access, available from http://uniprot.org/) for protein sequence data and functional information, and the KEGG database (open access, available from http://www.genome.jp/kegg/pathway.html) for reaction pathway information.

The complete dataset of enzyme kinetic parameters linked to Michaelis-Menten reaction systems was first downloaded from the Sabio-RK database and sorted by top-level EC association. This was then stored in a simple '.csv' format to simplify future analysis efforts.

The first approach to analysing this dataset involved determining whether groups of enzymes lower than the first level of the EC hierarchy (n.n.*.* - n.n.n.n) were still best represented by a log-normal distribution. This task was approached by writing a data grouping script with the format visible below (algorithm 1). This script was then used as the basis for many of the following analysis methodologies, with the grouped data now able to be iterated over to produce Q-Q plots, determining how well a log-normal distribution fit the data.

```
for n_i in n.*.*.* do
    if group n_i exists then continue;
     else group data by n_i;
    for n_j in n_i.n.*.* do
        if group n_j exists then continue;
         else group data by n_j;
        for n_k in n_i.n_j.n.* do
            if group n_k exists then continue;
             else group data by n_k;
            for n_l in n_i.n_j.n_k.n do
                if group n_l exists then continue;
                 else group data by n_l;
            end
        end
    end
end
```

Algorithm 1: Data Grouping

Summary statistics, including the mean, standard deviation, and number of entries per EC number, were then computed for each level of the EC hierarchy, and the distributions were analysed.

At this point, a k-nearest neighbours classifier implementation into the dataset was attempted in the hope that it would predict enzyme kinetic parameters based on the enzymes that were closest to it within the EC hierarchy. Each value within this hierarchy was determined as being a separate point in parameter space, with each entry for a specific EC value occupying that same position, and therefore having an equal chance of being chosen. For example, if enough information was known about our enzyme to narrow it down to the third level (n.n.n.*), then if $k = 15$, that amount of neighbours would be randomly selected and averaged from n.n.n.* to predict the parameters for the unknown enzyme. If enough information was known about the enzyme to assign a full EC number to it (n.n.n.n), then k neighbours would be selected from its group and averaged. Many of these groups do not contain enough entries to satisfy the value of k, and therefore until k is satisfied, neighbours are selected based on the minimisation of the Euclidean distance from the EC value of k (i.e. if our target had an EC value of n.n.n.1, then n.n.n.2 would be preferable over n.n.n.16).

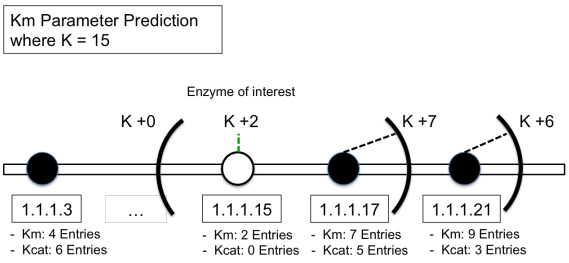

Figure 2: Visualisation of the k-nearest neighbours algorithm in the context of the EC hierarchy.

After some issues were encountered with the k-nearest neighbours methodology, a variety of statistical tests were then conducted to determine how distinct each branch of the EC hierarchy was in-terms of the values of its kinetic parameters. For this, analysis of variance (ANOVA) tests were predominantly used, along with pairwise Student's t-tests of the mean of each individual parameter for each branch of the hierarchy. From these analyses, the distance between these mean values was also investigated to determine whether it grows from a given point of origin.

Lastly, in-order to follow-up on the results from the previous tests, the root-mean-square error was investigated while using the values contained within each level of the EC hierarchy from n.*.*.* to n.n.n.n as a predictor for every given reaction parameter. To do this, a script was written to calculate the mean of a given parameter from each level and permutation of the EC hierarchy. To compare that result with a specific reaction, the error value was stored and averaged across all reactions within a group. For example, for a given reaction (i.e. EC number: 1.4.7.16), the average K_m value for 1.*.*.*, 1.4.*.*, and 1.4.7.*, would be used as a predictor for one of the K_m entries within 1.4.7.16. The error between the predicted value, and the actual value, would then be stored and averaged against all other calculations for all other reactions, determining the most effective level of the EC hierarchy as a predictor for enzymatic kinetic parameters. The equation for this measure (the root mean-square error) is given as:

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

where n gives the number of observations, y_j gives the value of the data-point of interest, and $\hat{y}_j$ gives the observed value to be compared with y_j.

Unless otherwise stated in the results section, findings are in relation to the K_m parameter. This is primarily due to the amount of data available for each parameter, with K_m having the most available entries, k_{cat} having the second-most, and V_{max} having the fewest.

3 Results & Discussion

The EC data series takes on a branched tree hierarchical organisation, with each node on the 4 levels producing a varied number of branches. The moderate complexity of this dataset is best visualised as a tree diagram, pictured below (figure 3).

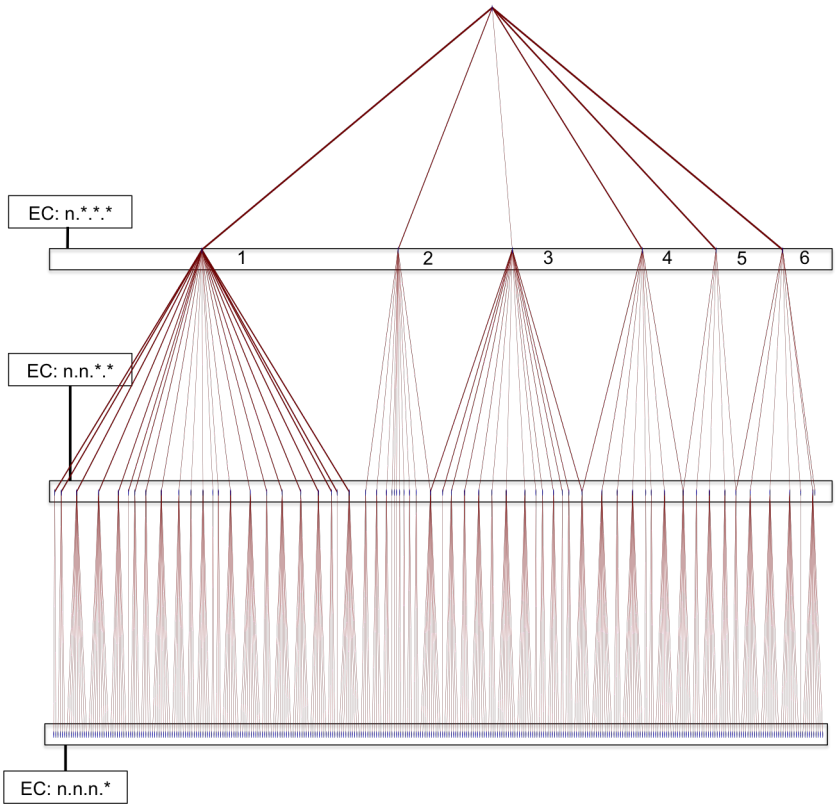

Figure 3: EC hierarchy from n.*.*.* to n.n.n.*.

The dataset itself was surprisingly varied, with little consistency in either the amount of reactions present per category, or the amount of entries per EC number.

EC Class 1: Oxidoreductases	**EC Class 2: Transferases**
1555	1654
EC Class 3: Hydrolases	**EC Class 4: Lyases**
1291	582
EC Class 5: Isomerases	**EC Class 6: Ligases**
253	183

Table 1: Summary table containing the number of enzymes in the top level of the EC hierarchy (n.*.*.*).

The distribution of kinetic parameter entries for K_m and k_{cat} were plotted against their individual EC value and then sorted by the amount of entries they had. The resulting figure (figure 4), demonstrates a steep inverse exponential curve for both of the parameters. This result strongly demonstrates the issue that plagues many modelling efforts. Enzymes that are popular, either because of their function or ideal parameter values, have lots of entires describing their kinetic parameters; yet the vast majority of enzymes have very few entires. This leads to a lack of reliability for the vast majority of entries in databases such as Sabio-RK, and therefore provides support for the use groups of functionally similar enzymes for the production of realistic prior distributions.

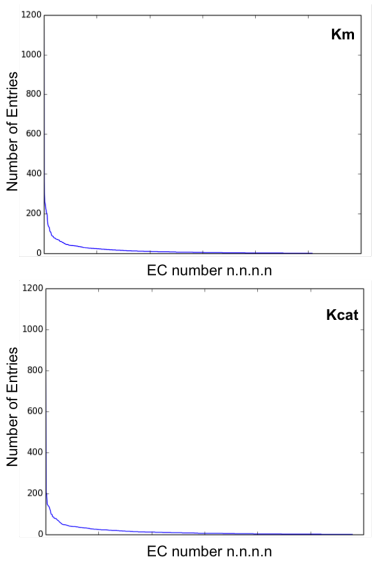

Figure 4: Number of kinetic parameter entries per individual EC number.

As mentioned previously, work completed prior to

this report demonstrated the feasibility of fitting a log-normal distribution to the six classes within the first level of the EC hierarchy (n.*.*.*). For this study, this assumption was tested against EC n.n.*.*, n.n.n.*, and n.n.n.n. At n.n.*.* the assumption holds true, with enough data-points to satisfy the distribution. These data-points appear to be reasonably well aligned with a log-normal distribution as can be seen in the Q-Q plots below (figure 5). This result supports the use of smaller groups that are more closely related to the enzyme of choice when selecting distributions in comparison with the top level of the EC hierarchy. For example, if enough information was known about an enzyme to narrow it down to the second level within the EC hierarchy, then an appropriate prior distribution for this enzyme could be produced using data from the enzymes within its specific sub-subgroup, provided that they give a better prediction than is available at n.*.*.*.

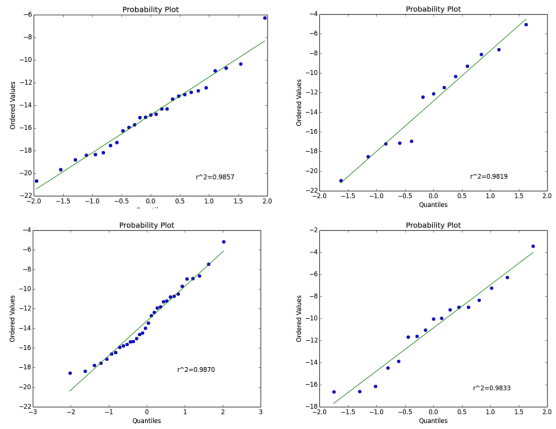

Figure 5: Representative Q-Q plots of n.n.*.* K_m entries against a fitted log-normal distribution.

Below EC n.n.*.* the log-normal distribution doesn't appear to fit the data due to a lack of data points in many of the groups, see below for a representative example (figure 6). Although some groups contain a larger amount of entries, many do not, and so when considering the possibility of dynamically producing a prior distribution for any given enzyme, a safer assumption given this result is to use EC n.n.*.* and above. This issue could potentially be circumvented with a dynamic algorithm that selected a prior distribution from n.n.n.* if there were enough data points to warrant so, and from n.n.*.* if not, but given the scope and time limitations of this project, this was not achievable.

At the fourth level of the hierarchy (n.n.n.n), no examples could be found to satisfy a log-normal distribution. As this level specifies a specific reaction, some of them contain hundreds of entries per kinetic parameter, yet as many of these entries are identical in value, a true log normal distribution was not achievable. Also, as shown previously in figure 4, the number of n.n.n.n entries that contain a sizeable amount of values per parameter are limited to a small subsection of the overall enzymes. At the n.n.n.n level of knowledge about an enzymes structure, there is arguably little need for parameter estimation, as the enzyme values are likely already known and can be looked up, or alternatively, are unknown and therefore are not likely to be predicted from other entries within the same reaction group.

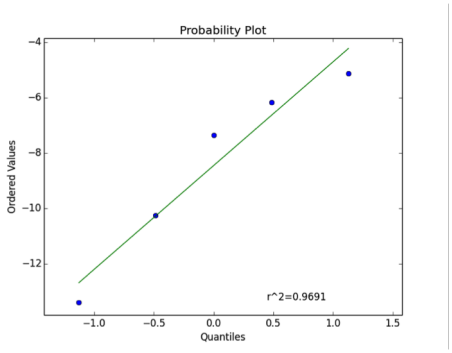

Figure 6: An example Q-Q plot of an n.n.n.* K_m entry against a fitted log-normal distribution.

Given the encouraging finding of a log-normal distribution fitting n.n.*.* groups, a k-nearest neighbours machine learning approach was attempted on the dataset. This approach assumed a linear relationship between n.n.*.* groups, where 1.1.1.15 would be closely related to 1.1.1.16 and 1.1.1.14, and groups that are numerically further from the EC number of interest are less likely to be picked as neighbours.

n.n.n.n and n.n.n.* values were determined to be the most appropriate for this machine learning classifier due to the 'neighbour-like' relationship they share with the groups around them. When using n.n.*.* values as input, there were generally far more entries to satisfy the value of$k$then were required, leading to a technique akin to taking small random samples from large data-pools, adding no benefit to the predictive power of the technique. When using n.n.n.n and n.n.n.* values to find neighbours, the predictive power of this technique proved to be very poor. With n.n.n.* EC values given to the predictor, and the number of neighbours (k) set between 15 and 50, the average precision value (calculated as the ratio of true-positive to false positive predictions) was as low as 2%, and as high as 12% depending on the size of the training set given. Higher precision val-

ues could be achieved by overfitting the dataset, however this is artificial and would likely not be useful within a research setting. Accuracy values were similarly disappointing.

This poor output was primarily due to the nature of the algorithm in combination with the assumptions we had placed on the dataset. In assuming that there was a linear relationship in kinetic parameter difference, increasing the further that you were from the starting EC value, we would therefore sample from around the given EC number. The issue with this was that if the value of$k$was too small, then the sample size wasn't representative of the surrounding enzymes, and therefore was ineffective at predicting the true value, yet if a larger$k$value was given then you risk sampling from structures that are dissimilar from the starting enzyme. Picking an appropriate static value for$k$given the dynamic nature of the task was difficult, and assuming that the need for realistic prior distributions will most likely originate from researchers with little knowledge on their enzyme of choice, it was decided that a change in approach would be necessary.

From this point in the project, we began to investigate the relationship between different levels of the EC hierarchy, examining the statistical significance between their respective values and querying the possibility that the relationship between these values wasn't as linear as was first assumed. We began by conducting a pairwise Student's t-test for each element on each level of the EC hierarchy. The results for levels EC n.n.*.* - n.n.n.n were mixed, primarily due to the amount of variables in each of the many groups. Unsurprisingly, some of these groups demonstrated a significant difference ($P \leq 0.05$) when paired against others, however there was little correlation in each groups perceived distance from its pair and the likelihood that it would demonstrate a significant difference. When the test reached EC n.n.n.n, the test was severely limited by the lack of parameter values and variation. For the 6 groups within EC n.*.*.*, the pairwise Student's t-test demonstrated a distinct lack of significant parameter value variation between groups, with no tests giving significant differences, and only 2 giving borderline significant results (1.*.*.* in comparison with 2.*.*.*, and 2.*.*.* in comparison with 3.*.*.*). For the full comparison, see below (table 2).

This lack of significant results provides a clear explanation for why our k-nearest neighbours approach failed, as our assumptions regarding intra-group differences were incorrect.

A follow-up ANOVA test of the n.*.*.* parameter means produced a P-value of 0.09 for K_m parameter values, and 0.18 for k_{cat}. Although arguably borderline signifiant for the K_m parameter, it could also be argued that an ANOVA test is less representative of the variation between groups as it investigates the central tendencies of all groups. Although the groups as a whole have a borderline significant difference in their mean values, this doesn't necessarily support our assumption of a distance-based difference in parameter values. What this result does support however, in combination with the distribution findings, is the use of data within EC hierarchy groups to predict kinetic parameter values and produce realistic prior distributions.

n.*.*.* Set 1	**n.*.*.* Set 2**	**P-value**
1.*.*.*	2.*.*.*	**0.057**
1.*.*.*	3.*.*.*	0.274
1.*.*.*	4.*.*.*	0.872
1.*.*.*	5.*.*.*	0.746
1.*.*.*	6.*.*.*	0.563
2.*.*.*	3.*.*.*	**0.080**
2.*.*.*	4.*.*.*	0.428
2.*.*.*	5.*.*.*	0.226
2.*.*.*	6.*.*.*	0.365
3.*.*.*	4.*.*.*	0.664
3.*.*.*	5.*.*.*	0.522
3.*.*.*	6.*.*.*	0.523
4.*.*.*	5.*.*.*	0.936
4.*.*.*	6.*.*.*	0.409
5.*.*.*	6.*.*.*	0.325

Table 2: Results of the Pairwise Student's t-test for all n.*.*.* pairs of the K_m parameter values.

In support of our pairwise t-test findings, the root mean-square-error (RMSE) between each value of an example K_m EC dataset was plotted. The EC values used for this plot were 1.1.*.* to 1.99.*.* Based on our previous assumption of a linear increase in difference between parameter values the further the samples were from the EC value of origin, we would expect a linear increase in RMSE from group 1.1.*.* to 1.99.*.* (16 data-points when using appropriately large K_m data from the groups). Our expectations and our findings can be seen below (figure 7).

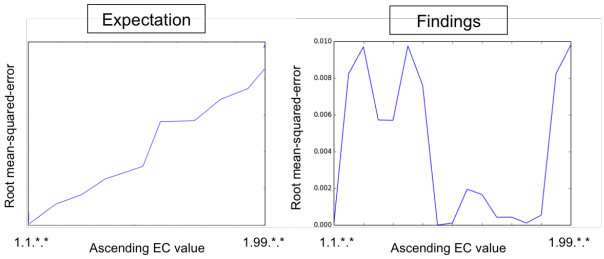

Figure 7: Root mean-square-error of ascending EC group K_m values against the group 1.1.*.*

Based on the results outlined previously, it was decided that using the mean and standard deviation of the closest EC group to the enzyme of choice would be most accurate for fitting a log-normal distribution to for each prior distribution. In order to determine which level of the EC hierarchy (*.*.*.* - n.n.n.*) would be the best overall predictor for the production of realistic prior distributions, we calculated the RMSE of each relevant groups' predictions against the true value of the parameter. For the K_m parameter value, EC *.*.*.* gave a RMSE value of 0.1245, n.*.*.* gave a RMSE value of 0.1241, n.n.*.* gave a RMSE value of 0.1227, and n.n.n.* gave a RMSE value of 0.1242.

This finding suggests that the second level of the EC hierarchy (n.n.*.*) provides the strongest predictive ability for the most abundant parameter entry in the dataset, balancing the amount of data within each group against the structural/functional relationships that the associated enzymes share. At this level of the hierarchy, the data in each group is filtered by the specific action of the enzyme, with the second level dictating what molecule, or group of molecules, the enzyme acts on. When considering this biological information, this result is unsurprising given that any further subclassifications of the enzyme's function is less likely to alter the central tendency measure that dictates the group's predictive ability. Further subclassifications will also significantly diminish the availability of data within each group, potentially skewing the measure of central tendency with outlying values measured in non-ideal circumstances.

This project has produced some interesting results, and beckons for future work to build on the information that has been derived from it. The first step in any future work to be conducted, would be to measure whether prior distributions produced from the second layer of the EC hierarchy influence the outcome achieved with the predictive ABC methodologies outlined previously. The log-normal prior distributions calculated from the mean and standard deviation of the selected EC group could be easily fed into an ABC program and sampled from to predict the posterior distributions. If these realistic priors proved to produce better outputs than the alternative uniform distributions, then the findings can be said to be useful. They may also help by reducing the computation time when using the ABC rejection algorithm by creating a smaller parameter space to sample from. If this is shown to work, then it could potentially be implemented into an easily accessible graphical user interface (GUI), where a user can submit the information they know about the enzyme and receive a realistic prior based on the EC classifications of the information that they provide the program with. The program itself could take on the form of a lookup table, where if the data provided fits up to the second level of the EC hierarchy, then the program retrieves the mean and standard deviation of the group and returns this as a log-normal distribution to the user. The program could regularly update its database of parameter values from the large amount of easily accessible databases, including Sabio-RK, which we already have data retrieval scripts for.

The codebase and theory demonstrated in this project, and in previous work, also has the potential to be applied to similar systems, where realistic prior distributions can be used in the prediction of parameter values. For example, if a transcriptomics dataset containing various transcription rates for specific genes was obtained, then it could be used to predict the transcription rates of structurally similar genes through the use of ABC in combination with realistic prior distributions, using the techniques outlined above to classify and investigate the dataset. Unfortunately, no open-access databases could be found that contained this information, but nonetheless this technique is not strictly limited by the biological context and could conceptually be applied to quantitative modelling efforts in fields such as evolutionary biology, cancer modelling, and interactome analysis.

This project, like many others, is well balanced by its strengths and weaknesses. The strengths of this project lie in the depth of the analysis carried out on the EC hierarchy, and the data contained within. Throughout this study, a multitude of statistical tests and analyses were conducted on various aspects of the data, and we feel that the results, although occasionally unexpected, do represent the data effectively. Our findings are supportive of the predictive capabilities of this dataset, and we feel that this result is positive in-the-sense that it further supports the use of approximate bayesian computation. This groundwork that we have conducted effectively provides an information base to build-upon and explore, and leaves a clear direction

for future work.

When looking at the weaknesses of this project, the key one to highlight is the false assumption that we placed on the dataset and maintained for some time. By assuming a linear distance between EC groups and the difference between parameter values, we spent far too much time trialling predictive methodologies that were ineffective without this assumption being true. It would have been beneficial to the work had we tested this assumption before investigating specific machine learning techniques. Also, due to the size of the datasets available to us, some of the analyses were left unfinished as large gaps in the data for parameters such as k_{cat} and V_{max} were difficult to overcome. This limitation, although unavoidable, will likely affect future efforts to produce realistic prior distributions for specific groups of enzymes, but will also likely reduce in scale as more data on enzyme kinetics become available.

4 Conclusion

To conclude, this study has investigated the potential of using the EC hierarchy to produce realistic prior distributions of enzyme kinetic parameters. We found through the relationship between groups within the hierarchy, that the value of parameters varied less between groups then was expected, and that the second level of the hierarchy provided the greatest amount of predictive power for the dataset in question. Attempts to 'learn' the dataset using machine-learning methodologies were ineffective, suggesting that the production of realistic prior distributions was most effective using the mean values obtained from enzyme groups associated with the enzyme of interest.

From these findings, we have provided a strong base to work from, and a clear direction for future work. Further research will focus on implementing these realistic prior distributions within ABC programs such as ABC SysBio to test their feasibility in improving the predictive power of Bayesian approaches to parameter inference.

References

[1] Sidney M Morris. Regulation of enzymes of the urea cycle and arginine metabolism. *Annual review of nutrition*, 22:87–105, 2002.

[2] Y Moriwaki, T Yamamoto, and K Higashino. Enzymes involved in purine metabolism–a review of histochemical localization and functional implications. *Histology and histopathology*, 14:1321–1340, 1999.

[3] Yuguang Shi and Paul Burn. Lipid metabolic enzymes: emerging drug targets for the treatment of obesity. *Nature reviews. Drug discovery*, 3:695–710, 2004.

[4] J Mark Berg, J L Tymoczko, and L Stryer. Biochemistry. 5th edition. In *Biochemistry textbook*, page 1120. 2006.

[5] Robert N. Goldberg, Yadu B. Tewari, and Talapady N. Bhat. Thermodynamics of enzyme-catalyzed reactions - A database for quantitative biochemistry. *Bioinformatics*, 20:2874–2877, 2004.

[6] Johannes C Hermann, Ricardo Marti-Arbona, Alexander A Fedorov, Elena Fedorov, Steven C Almo, Brian K Shoichet, and Frank M Raushel. Structure-based activity prediction for an enzyme of unknown function. *Nature*, 448:775–779, 2007.

[7] Eleanore Seibert and Timothy S. Tracy. Fundamentals of enzyme kinetics. *Methods in Molecular Biology*, 1113:9–22, 2014.

[8] Arren Bar-Even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S. Tawfik, and Ron Milo. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50:4402–4410, 2011.

[9] K Tipton and S Boyce. History of the enzyme nomenclature system. *Bioinformatics (Oxford, England)*, 16:34–40, 2000.

[10] S. Grisolia. Enzyme Nomenclature, 1961.

[11] Katalin Csilléry, Michael G B Blum, Oscar E. Gaggiotti, and Olivier François. Approximate Bayesian Computation (ABC) in practice, 2010.

[12] Juliane Liepe, Chris Barnes, Erika Cule, Kamil Erguler, Paul Kirk, Tina Toni, and Michael P H Stumpf. ABC-SysBio-approximate bayesian computation in python with GPU support. *Bioinformatics*, 26:1797–1799, 2010.

[13] Mark A. Beaumont. Approximate Bayesian Computation in Evolution and Ecology, 2010.