MetaMolES : A Biochemical Retrosynthesis Tool Using Logistic Regression Model Based on Enzyme Promiscuity

Stephen Blaskowski, Yeon Mi Hwang, Ellie James, Cholpisit(Ice) Kiattiseswee, Phil Leung
Molecular Engineering and Science Institute

Introduction

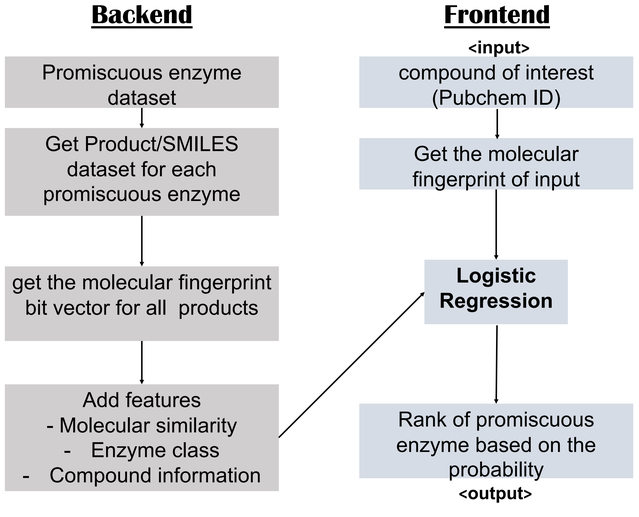
Background

Substrate promiscuous enzymes can biochemically transform several substrates. The use of promiscuous enzymes in metabolic engineering is particularly advantageous because it relieves the burden on the engineered organism. However, most enzyme databases only link one main substrate to each enzyme instead of a repertoire of suitable compounds, which limits the chance for researchers to utilize promiscuous enzymes. Also, another limitation for metabolic engineers comes from the limited access to closed-source biochemical retrosynthesis tools. To address these limitations, herein, we present a user-friendly open-source tool that helps curate the most plausible enzymatic transformation based on substrate promiscuity.

Goal

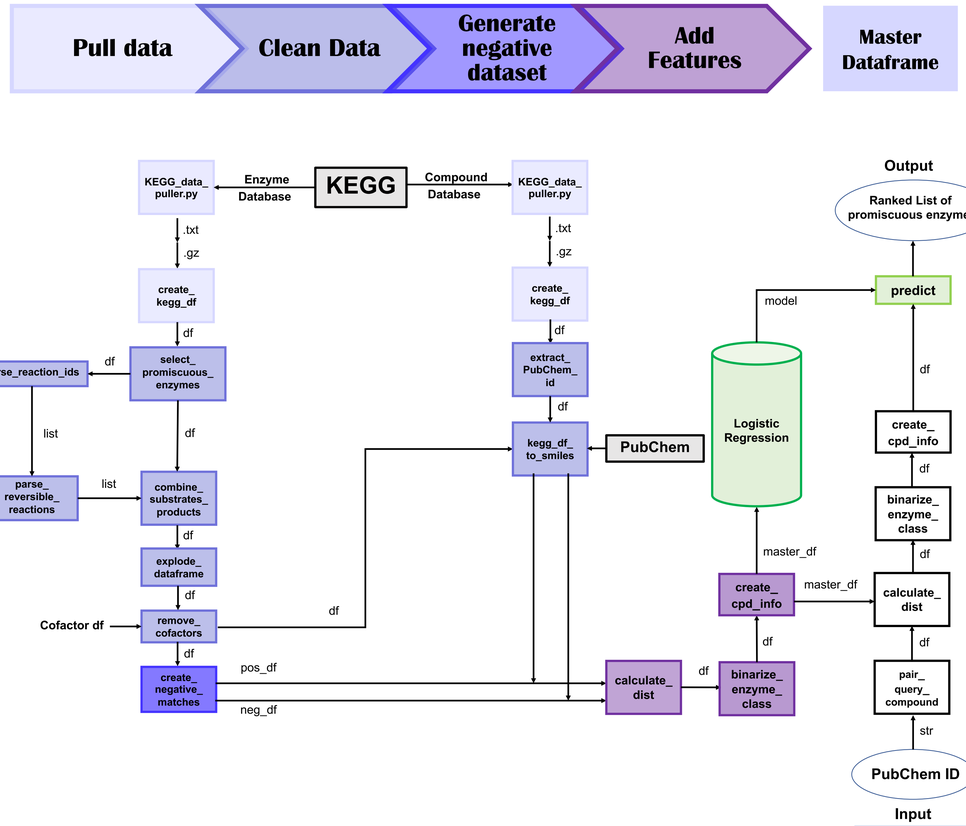
- Aim to utilize data science and software engineering intuition to find, and predict, a plausible metabolic pathway for production of a given molecule with retrosynthetic analysis approach.
- Find a novel promiscuous substrate for enzymatic transformation

Workflow

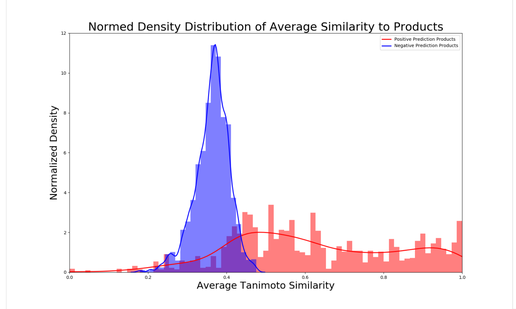


Method

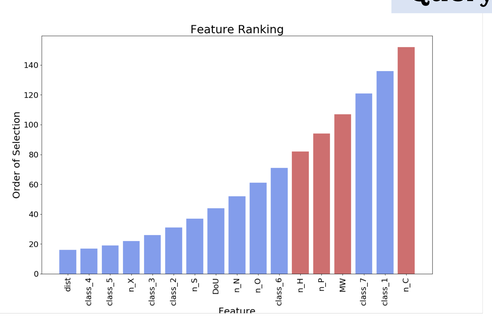
Data Curation



Logistic Regression



normalized counts of calculated average Tanimoto similarity for positive and negative predictions.

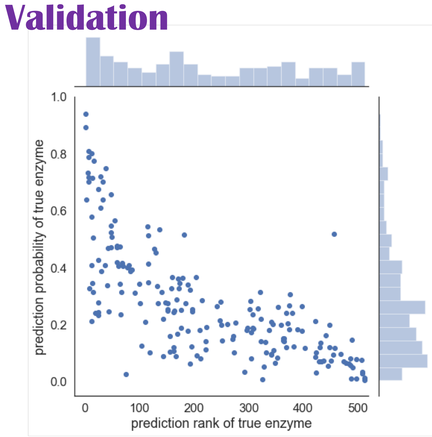


Out of 17 features, 13 features were chosen for logistic regression based on the ranking in their significance.

We selected logistic regression over an SVM to perform enzyme/compound reaction pairs because of our relatively small dataset and our desire to select and rank outputs based on predicted likelihood of reaction. The model was generated with sci-kit learn from 13 features with balanced feature weights and a liblinear solver. The output, after organizing and sorting, is the likelihood of reaction with a specific enzyme. The enzyme can then be looked up in the KEGG database.

Result

Model Validation



We set aside 207 (10%) known promiscuous enzyme-product pairs from our dataset prior to training and testing the logistic regression model. As an evaluation of predictive power, we fed each of the products from this set into the model and recorded the prediction probability for the known true enzyme, as well as the relative rank of the enzyme suggestion out of the master set of 516 promiscuous enzymes. We found a negative correlation between probability and rank (slope=-9.21e-4, p=1.70e-35, R^2=0.54). This suggests a bias away from false positives and towards false negatives. Top and left marginal figures are marginal histograms of rank and probability, respectively.

Conclusion

- The logistic regression model is ~90% predictive when 20% of the data was reserved for testing
- The model correctly predicted reactivity (P>0.5) for 14% of the validation reactions, and among these positive categorizations, the median prediction rank was 28
- Results suggest that using distance of chemical similarity is a reasonable approach
- Validation testing revealed a negative correlation between prediction rank and prediction probability, suggesting a bias away from false positives, and targets for model improvement

Future Work

- Add compounds with canonical SMILES string but no isomeric SMILES string
- Test inclusion of additional features, such as full chemical fingerprints, and enzyme descriptors
- Explore alternative models, such as SVMs, neural networks, decision trees/random forests, and ensemble methods.
- Extend approach to include non-promiscuous enzymes
- Include simple chemical transformation for biocatalysis application

Github repo: https://github.com/theicechol/metamoles

Dependency : Rdkit, bioPython, Pubchempy, scipy, sklearn, pandas, numpy

References

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M.; New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590-D595 (2019).
Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).
Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).
Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019 Jan 8; 47(D1):D1102-1109. doi:10.1093/nar/gky1033.
Pertusi, Dante A., et al. "Predicting novel substrates for enzymes with minimal experimental effort with active learning." Metabolic engineering 44 (2017): 171-181.
Segler, Marwin HS, Mike Preuss, and Mark P. Waller. "Planning chemical syntheses with deep neural networks and symbolic AI." Nature 555.7698 (2018): 604.