Metabolic Engineering 44 (2017) 171–181

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/meteng

Predicting novel substrates for enzymes with minimal experimental effort with active learning

Dante A. Pertusi[a], Matthew E. Moura[a], James G. Jeffryes[a,b], Siddhant Prabhu[a], Bradley Walters Biggs[a], Keith E.J. Tyo[a,*]

[a] *Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, United States*

[b] *Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, United States*

ARTICLE INFO

Keywords:
Machine learning
Active learning
Enzyme promiscuity

ABSTRACT

Enzymatic substrate promiscuity is more ubiquitous than previously thought, with significant consequences for understanding metabolism and its application to biocatalysis. This realization has given rise to the need for efficient characterization of enzyme promiscuity. Enzyme promiscuity is currently characterized with a limited number of human-selected compounds that may not be representative of the enzyme's versatility. While testing large numbers of compounds may be impractical, computational approaches can exploit existing data to determine the most informative substrates to test next, thereby more thoroughly exploring an enzyme's versatility. To demonstrate this, we used existing studies and tested compounds for four different enzymes, developed support vector machine (SVM) models using these datasets, and selected additional compounds for experiments using an active learning approach. SVMs trained on a chemically diverse set of compounds were discovered to achieve maximum accuracies of ~80% using ~33% fewer compounds than datasets based on all compounds tested in existing studies. Active learning-selected compounds for testing resolved apparent conflicts in the existing training data, while adding diversity to the dataset. The application of these algorithms to wide arrays of metabolic enzymes would result in a library of SVMs that can predict high-probability promiscuous enzymatic reactions and could prove a valuable resource for the design of novel metabolic pathways.

1. Introduction

Substrate-level enzyme promiscuity (Humble and Berglund, 2011; Khersonsky and Tawfik, 2010; Sévin et al., 2016) has recently been recognized as a far more ubiquitous phenomenon than previously assumed, having important implications in several research areas. Substrate-level enzyme promiscuity is a phenomenon in which an enzyme can catalyze a reaction on more than one substrate, and the challenges it presents in experimental methods have made development of *in silico* methods for its prediction a subject of intense interest. Understanding this phenomenon is critical to explaining many biological processes. Enzyme promiscuity plays a key role in metabolite damage (Linster et al., 2013), a phenomenon where essential metabolites are converted to non-useful forms by reactions catalyzed by both homologous and heterologous promiscuous enzymes in wild-type and engineered organisms. Metabolite damage can drastically reduce fitness, to the extent that specific repair mechanisms have evolved to convert damaged metabolites back into a form that can be used by the cell (Linster et al., 2013; Van Schaftingen et al., 2013). Enzyme promiscuity is also critical in specific mechanisms of antibiotic resistance. Promiscuous enzymes can compensate for inhibited essential enzymes, allowing bacteria to circumvent the block. For example, methotrexate is an effective antibiotic against many microbes by inhibiting dihydrofolate reductase, an essential enzyme. However, in *Lieshmania major*, a second enzyme, PRT1, a broad spectrum pteridine reductase, is able to catalyze the DHFR reaction and is not inhibited by methotrexate (Gourley et al., 2001; Nare et al., 1997). Finally, enzyme promiscuity is important in metabolic evolution: It has been hypothesized that promiscuous enzymes improve fitness by serving as a starting point in biochemical evolution (DePristo, 2007; Khersonsky and Tawfik, 2010). This would explain instances of related proteins having a wide range of activities (Verdel-Aranda et al., 2015) and the presence of pathways that rescue what would otherwise be lethal knockout strains (Kim and Copley, 2012).

Enzyme promiscuity also has substantial positive and negative effects on industrial biotechnology. Substrate-level promiscuity can enable a tantalizing array of novel biosynthetic routes to drugs and biochemical. Promiscuous enzymes may also catalyze non-canonical

[*] Corresponding author.
E-mail address: k-tyo@northwestern.edu (K.E.J. Tyo).

http://dx.doi.org/10.1016/j.ymben.2017.09.016
Received 1 March 2017; Received in revised form 20 August 2017; Accepted 18 September 2017
Available online 10 October 2017

reactions, allowing for the potential construction of metabolic routes to compounds not known to occur in nature. On the other hand, enzyme promiscuity can also be problematic to industrial biotechnology applications. Heterologous enzymes can promiscuously act on unanticipated native metabolites, diverting carbon from a desired end product (Mafu et al., 2016); in engineered pathways, concentrations of heterologous enzymes and metabolic intermediates are pushed to high concentrations such that the probability of substrate-level promiscuity is high, leading to unintended and deleterious effects on biochemical production (Biggs et al., 2016). To address this, there has a been a wave of computational approaches to design novel metabolic pathways and predict byproduct-producing and/or damage reactions (Campodonico et al., 2014; Carbonell et al., 2014; Cho et al., 2010; Lee et al., 2012). A rising challenge in applying these methods is low accuracy, largely due to a paucity of high-quality data on which to best train *in silico* prediction models.

Enzyme promiscuity is commonly investigated by assaying the activity of an enzyme on several different compounds selected ad hoc. However, selecting substrates that best expand the knowledge about an enzyme's promiscuity is a non-trivial task, as there are a large number of possible substrates. Furthermore, several compounds may be prohibitively expensive or difficult to procure, making an exhaustive experimental study infeasible. To avoid the time and expense of carrying out activity assays, *in silico* methods can be substituted that either make use of existing promiscuity data or reduce experimental effort by collecting informative substrates. Promiscuity characterization is a task that can be addressed by cheminformatics and machine learning methods. While molecular modeling suites and docking approaches are capable of providing more nuanced predictions about the interactions between enzymes and potential substrates than 2D cheminformatics approaches, crystal structure information they require is often not available.

In order to predict enzyme promiscuity, previous studies (Campodonico et al., 2014; Cho et al., 2010; Pertusi et al., 2015) have deployed similarity-based approaches to predict an enzyme's ability to catalyze a reaction on a given substrate. These methods rely on catalogues of known substrates as listed in databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2014) or the Braunschweig Enzyme Database (BRENDA) (Schomburg et al., 2013). Similarity methods (Willett, 2006) and supervised machine learning methods are subsequently used to rank and/or classify likely substrates, and has been applied to numerous protein targets and off-targets in drug development efforts (i.e., CYP450 (Jacob and Vert, 2008; Terfloth et al., 2007; Wale and Karypis, 2009)). Support vector machines (Schölkopf and Smola, 2002) (SVMs) are an approach that has performed well in these studies and are trained on data belonging to each of two labeled classes of interest. In the context of predicting substrate-level enzyme promiscuity, the labeled classes are enzyme substrates and non-substrates. The resulting model can then be used to predict if an untested compound is a substrate or non-substrate (cf. Methods). With a previous study estimating that 39% of enzymes in KEGG exhibit substrate promiscuity (Carbonell and Faulon, 2010), the extension of this approach to metabolic enzymes can streamline the process of identifying promiscuous enzymes with a desired side activity as a prelude to structure-based engineering efforts.

Existing databases used to train enzyme promiscuity models have limitations. For most enzymes of biosynthetic interest, the datasets of tested compounds are not very large. As an example, only ~5% of enzymes across all *Escherichia coli* strains in the BRENDA database have 20 or more reported substrates (e.g., Fig. 1A, Table S1). Secondly, these datasets contain disproportionately small numbers of inactive compounds, without which the task of distinguishing between substrates and inactive compounds is highly uncertain (Fig. 1C, D). Finally, existing datasets do not generally explore a diverse set of possible substrates may contain many compounds of relatively low diversity, and therefore low information content (Fig. 1B). While there is no consensus on the number of training compounds required to adequately train an SVM classifier, it is clear that larger, more diverse compound datasets will be required to improve classification power (Matykiewicz and Pestian, 2012), and collecting this data efficiently is very important.

In order to probe chemical space in an efficient manner while still avoiding the drawbacks of exhaustive experimentation, an *active learning* method can be employed to direct the task of data collection. Active learning (Settles, 2012; Warmuth et al., 2003) is a term applied to a number of approaches that use information about the set of unlabeled instances (i.e., untested compounds) in order to make strategic choices about which unlabeled instances to query next so that the result can efficiently train the classifier. In the case of SVMs for predicting substrate-level enzyme promiscuity, active learning serves to prioritize compounds in chemical space according to the additional predictive power knowing the result of that substrate may add to the SVM. Active learning could then offer a roadmap to experimentally efficient enzyme characterization. However, a question remains as to whether this approach can provide sensible compounds as suggestion despite the relatively small datasets available.

In this study, we examine the utility of SVM-based machine learning to predict enzyme substrate promiscuity across a range of enzymatic chemistries and examine the ability of active learning to prioritize new compounds for testing that efficiently expand the domain of applicability of the classifier. Specifically, we compile four enzymes' datasets and develop SVMs for each of them using existing data, demonstrating their efficacy on relatively homogenous chemical sets. We then develop an active learning approach to strategically expand the available pool of compounds—both active and inactive—to use as training data in developing SVM classification models for metabolic enzymes. To our knowledge, we are the first to apply active learning to metabolic enzymes and demonstrate that the approach is effective. We use SVM classification models to validate the efficacy of the active learning method by cross validating existing data. We go on to use an active learning approach to rank untested compounds containing the putative molecular active site from the ZINC Is Not Commercial (ZINC) database (Irwin et al., 2012)—a diverse catalogue of biochemically relevant chemicals—as to their ability to add classifying power to the model for a case study enzyme. Finally, we demonstrate that highly-ranked compounds are enriched for chemical features that add chemical diversity to the dataset.

2. Results

2.1. *Existing datasets have inadequate information content to evaluate substrate-level promiscuity*

Datasets with diverse chemical features are essential to predictive SVMs, as datasets containing only highly similar chemicals do not lead to accurate predictions on chemicals that are very different from the training set. Diverse sets of chemicals, however, make more accurate predictions for a wider set of potential substrates. To visualize the diversity of existing substrate-level promiscuity data, we first generated t-distributed stochastic neighbor embeddings (tSNE, cf. Methods) for each of the four compound datasets for the four enzymes considered in this study: 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic acid synthase (MenD) from *Escherichia coli* (Kurutsch et al., 2009), carboxylic acid reductase (Car) from *Nocardia iowensis* (Akhtar et al., 2013; Moura et al., 2016; Venkitasubramanian et al., 2008, 2007a, 2007b), amino acid ester hydrolase (AAEH) from *Xanthomonas citri* (Kato et al., 1980), and 4-hydroxyacetophenone monooxygenase (HAPMO) from *Pseudomonas putida* (Rehdorf et al., 2009) (Fig. 2, S2). This embedding reduces the many-dimensions in which chemicals vary to a two-dimensional projection while preserving the property that highly similar chemical structures will generally cluster closely together in the resulting visualization. In the case of the Car dataset, we noted that there is one larger cluster that dominates (Fig. 3A), with a small