NORTHWESTERN UNIVERSITY

Utilization and Computational Generation of Enzymatic Reaction Rules to Predict and Analyze Biochemical Pathways

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemical and Biological Engineering

By

Andrew Stine

EVANSTON, IL

December 2018

Abstract

Utilization and Computational Generation of Enzymatic Reaction Rules to Predict and Analyze Biochemical Pathways

Andrew Stine

The work in this thesis focuses on computational methods for the identification of novel enzymatic pathways. In particular this work focuses on the utilization of the Biological Network Integrated Computational Explorer (BNICE) software suite to predict *de novo* enzymatic pathways for the production of commercially relevant compounds and on improvements to this program which have the potential to increase both its universality and the ease with which its predictions can be verified.

BNICE uses generalized chemical operators to generate networks of probable biochemical reactions which include not only known enzymatic reactions but also likely reactions not previously found in literature. In the first part of this thesis, BNICE is used to predict enzymatic pathways for the production of propionic acid from pyruvate. 16 such pathways were found which consist of four enzymatic reactions or less. A key reaction in most of these pathways was found to be the reduction of acrylic acid to propionic acid. This reaction was experimentally confirmed by collaborators to be catalyzed by Oye2p from *Saccharomyces cerevisiae*, a previously unknown reaction for this enzyme.

Next, a method is developed for the automatic generation of BNICE chemical operators. Previously, these operators were generated manually, leading to the inability of the operators to describe many enzymatic reactions. This new method allowed for the generation of operator sets capable of describing every atom-balanced reaction in the MetaCyc database. Furthermore, a process is introduced for intuitive adjustment of the specificity of the generated operators by allowing the user to specify the groupings of reactions that should be described by each operator.

Finally, this new technique for automatic operator generation in combination with conserved domain database (CDD) superfamily information is used to create a set of operators such that each operator describes reactions associated with similar genes. BNICE is then utilized to apply these operators to every compound in *Escherichia coli* generating a list of 688,787 compounds. This list of compounds is then compared to the DrugBank database to identify 205 pharmaceutically relevant products which only require the addition of a single reaction for production from *Escherichia coli*. Furthermore, this method associates each predicted reaction with a CDD superfamily expediting the identification of promising enzyme candidates. These results illustrate the power and flexibility of BNICE and this operator generation program to identify promising enzymatic reactions and to associate these reactions with promising enzymes.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1: Motivation

Enzymes are capable of catalyzing an incredible variety of chemical reactions, often more efficiently, specifically, and safely than traditional chemical processes. However, the diversity of activity that makes enzymes so promising can also make the manual identification of the best enzymatic pathways for a given application quite challenging. Furthermore, once a potential biological pathway has been found it is often difficult to identify enzymes and genes associated with the constituent reactions. This thesis is concerned with methods to address these problems by computationally predicting promising enzymatic pathways and guiding the identification of candidate enzymes which may catalyze the constituent reactions in these pathways.

Many computational techniques for enzymatic pathway prediction exist [1]. Unfortunately, most of these methods rely upon databases of known enzymatic chemistry and are thus inherently restricted to exploring only previously discovered enzymatic reactions [2-5]. Our knowledge of enzymatic reactions is far from complete, and limiting our exploration to known enzymatic reactions precludes the possibility of discovering pathways which involve likely enzymatic reactions which have simply not yet been observed. A computational program that instead explores probable enzymatic reactions, even if they have yet to be observed, can guide the experimental search for new enzyme activity by identifying reactions which would be commercially important if an enzyme can be found to catalyze them. One such program is the Biological Network Computational Explorer (BNICE) [6]. BNICE uses mechanistically generalized reaction rules, known as *chemical operators*, to generate networks of probable biochemical reactions. These chemical operators are compiled from databases of known enzymatic chemistry and describe reactions with generic, flexible criteria for substrate acceptance. Applying these operators to a substrate not only reproduces known reaction routes and product compounds but also predicts novel biochemical reactions and products not previously found in literature. Applying these rules iteratively to the products of the predicted reactions generates a network of compounds that can be produced from the starting compound. Exploring this network allows for the identification of promising enzymatic reaction pathways. In this thesis, the utilization of BNICE is explored as are improvements to the software that greatly increase its flexibility and the ability of its prediction to guide the identification of promising enzymes for predicted reactions.

1.2 Research Outline

Chapter 2 of this thesis focuses on the utilization of BNICE to predict novel pathways for the production of propionic acid from pyruvate. This work has two important purposes. First, it identifies promising novel biosynthetic routes for the production of propionic acid, a commonly used preservative and chemical precursor. Secondly, and perhaps more importantly, it serves as a demonstration of the ability of BNICE to make actionable predictions of novel enzymatic chemistry. Using BNICE and 282 manually generated operators, seven pathways of four reaction steps or less were predicted for the production of propionic acid from pyruvate. Of the 16 reactions present in the pathways, five were known reactions used to create the operators, and two were known reactions which had not been used to create the operators but were instead identified following a more in-depth literature search. This confirms the ability of BNICE to correctly predict enzymatic reaction chemistry not present in its training set. The analysis of the predictions found that for 15 of the 16 pathways the reduction of acrylic acid to propionic acid was a key reactive step. Promising enzymes for catalyzing this reaction were identified through a literature search. Three such enzymes were identified: NADPH dehydrogenase (Oye2p), fumarate reductase, and 2-enoate reductase. Fumarate reductase and 2-enoate reductase were found not to be effective at performing this chemistry; however, Oye2p was found to successfully catalyze this reaction. This activity of Oye2p has not been previously observed in the literature. This represents the first time the BNICE program has been utilized to predict novel enzymatic activity which was subsequently confirmed experimentally.

The results in Chapter 2 successfully demonstrate the power and utility of BNICE. However, the dependency on manually generated operators limits its scope and flexibility. The process for the creation of operators by hand is very slow, requiring researchers to search through examples of enzymatic chemistry and identify patterns of activity, and consequently the generation of new operators can struggle to keep up with the discovery of new enzymes. This in turn can lead to large amounts of enzymatic chemistry not being described by any chemical operator. Chapter 3 of this thesis describes efforts to rectify this problem by developing a method for the automatic generation of BNICE chemical operators. We begin this work by exploring the limits of operator specificity by generating two operators sets based upon the MetaCyc database of enzymatic chemistry: one as specific as possible and one as general as possible [7]. In addition to describing every reaction in the MetaCyc database, these operators were found to be extendable to additional databases despite only utilizing MetaCyc reactions in their creation. The most generalized set of operators were able to describe 79.4% and 92.3% of the reactions in the KEGG database of enzymatic chemistry and the iMM904 metabolic model, respectively [8, 9]. This is an improvement over the manually created operators which were only able to describe 43.3% of the reactions in the iMM904 model and 48.2% of the reactions in KEGG despite being largely based upon reactions in that database [6].

We further improve upon our operator generation program by introducing the idea of generating operators of intermediate specificity between these two limiting cases by a process we call *reaction grouping*. In this method the user specifies groupings of similar reactions. For each grouping, the most specific operator capable of describing all the reactions in the grouping is generated. This method allows for the generation of operators of intermediate specificity in a manner that is intuitive to the user and easily adaptable to different applications. This method was utilized to generate operators based upon the reaction groupings present in the most generalized limiting case operators. The resulting set of operators again described every reaction in MetaCyc; however, using BNICE to apply these operators to the 22 common amino acids resulted in a 50 fold decrease in the number of predicted reactions compared to the most generalized limiting case operators. This process therefore provides the ability to sharply limit the number of potential reactions predicted by BNICE to those which are most similar to observed enzymatic chemistry. Furthermore, an analysis of the results found that for the intermediate operators generated by this process larger groupings of reactions produced more generalized operators while smaller groupings produced more specific ones. Users can therefore control the specificity of the generated operators by changing the size of the reaction groupings.

Chapter 4 of this thesis consists of a demonstration of the power of the new reaction grouping based method of operator generation. In this work chemical operators are generated from reaction groupings based upon genetic similarity of the associated enzymes as determined by Conserved Domain Database Superfamily classification [10]. BNICE was used to apply the resulting 2,210 operators to all the compounds present in the iJO1366 model of *E. coli* resulting in 319,075 predicted reactions and 688,787 predicted products. These compounds were compared to the compounds found in the DrugBank database of pharmaceutical compounds. Of the predicted compounds, 206 were found to be present in this database, and 20 were found to be FDA approved drugs. The tabulated results therefore provide a list of pharmaceutical compounds which are likely reachable in one reaction step from *E. coli*. Furthermore, by providing the CDD superfamily associated with the predicted reactions these predictions can help guide the identification of enzymes capable of performing a chemistry of interest. These results demonstrate the ability of the reaction grouping-based chemical operator generation method to allow BNICE to be quickly adapted to different applications.

In Chapter 5 the results from Chapters 2, 3, and 4 are summarized, and additional applications are suggested for BNICE and reaction-grouped operator generation.

Chapter 2: Exploring *De Novo* Metabolic Pathways from Pyruvate to Propionic Acid

2.1 Introduction

Industrial biotechnology, the practice of using enzymes or whole organisms for synthesis of commercial chemicals, has been widely explored to supplement or replace existing industrial processes [11, 12]. These new, biological processes often have considerable environmental and economic advantages over traditional chemical processes, such as operating at lower temperatures and pressures, forgoing the need for expensive catalysts, allowing for variable feedstock, and performing highly selective chemistry. These advantages make industrial biotechnology an attractive solution to manufacture chemicals in a sustainable economy. For biosynthesis of a target compound, designing the metabolic pathway and selecting enzymes capable of catalyzing each reaction step are central to the development of a viable biological process. Due to the sheer diversity of enzymes, designing a pathway and selecting enzymes can be very challenging.

To help overcome this challenge, several comprehensive biochemical databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Metacyc database, and the Braunschweig Enzyme Database (BRENDA), have been developed to store the information of experimentally elucidated enzymatic reactions extracted from scientific literature [13-15]. Computational methods ranging from path-finding algorithms to metabolic models of entire biological systems have also been developed to expedite enzymatic pathway discovery by searching such databases [16, 17].

These computational methods are limited to using only known enzymatic reactions. Numerous studies have shown that enzymes are often capable of catalyzing the reaction of substrates which are similar but not identical to their native substrate [18]. Such promiscuous activities are of interest in industrial biotechnology because they allow for the utilization of non-biological compounds in a biological process and provide pathway recommendations even when the desired chemical transformation is not directly found in biology. To exploit this opportunity, we have previously developed a program called the Biological Network Integrated Computational Explorer (BNICE) to automatically design biological pathways containing novel promiscuous reactions and novel metabolic intermediates [19-21]. By exploring these pathways, one can identify promising new routes for the production of a compound of interest.

In this work, BNICE is used to investigate pathways from pyruvate to propionic acid. Propionic acid is mainly used as a food preservative and as a precursor to other commodity chemicals and is primarily produced from petrochemicals. Propionic acid is naturally produced by the *Propionibacterium* genus of bacteria and by *Clostridium propionicum*, but current production is limited by significant lactic acid and succinic acid formation and low yields [22, 23]. New metabolic pathways that avoid lactic acid or succinic acid production would be very advantageous. Pyruvate was chosen as a starting substrate because of its high concentration as a glycolytic intermediate, and its carbon number (3) is the same as propionic acid, simplifying the transformations required.

In this study, we tabulated and analyzed all biochemical pathways predicted by BNICE for producing propionic acid from pyruvate in four reaction steps or less. Seven biochemical pathways were found including known and novel reactions. Upon analyzing five-step pathways, many converged on the final step of acrylic acid reduction to propionic acid. Our collaborators show experimentally that *Saccharomyces cerevisiae* Oye2p can carry out this predicted reaction. Although BNICE has previously been tested for its ability to predict known and novel enzymatic pathways [24-26], this is the first example of BNICE predictions being used to guide the experimental identification and realization of novel enzymatic activity.

2.2 Methods

2.2.1 Putative Network Generation and Exploration using BNICE

The BNICE program has been described in detail elsewhere [19]. Its operation will be described here briefly for convenience. BNICE automatically generates putative enzymatic reactions based on observed generalized enzyme functionality, as deduced from databases of known biochemical chemistry. Generalized enzyme functions are based on the Enzyme Commission four-tiered hierarchical classification system: EC i.j.k.l [27]. BNICE defines *operators* which account for the common chemical moieties of all the enzymes in a particular i.j.k class. To reflect this, BNICE operators are named using an i.j.k.a scheme where i.j.k correspond to the EC classification while the final letter identifies a particular operator. For this work, a set of 282 operators created from reactions in the KEGG database, the UMBBD database, and the iAF1260 *E. coli* metabolic model were utilized [13, 28, 29]. Applying these operators iteratively generates a network of compounds and reactions which are potentially reachable from the initial compound (Figure 2.1). Once a compound of interest has been reached, enzymatic routes to that compound are explored by applying a depth-first search pathway searching algorithm to the network. To avoid impractical pathways, we limited the search to four and five step pathways.

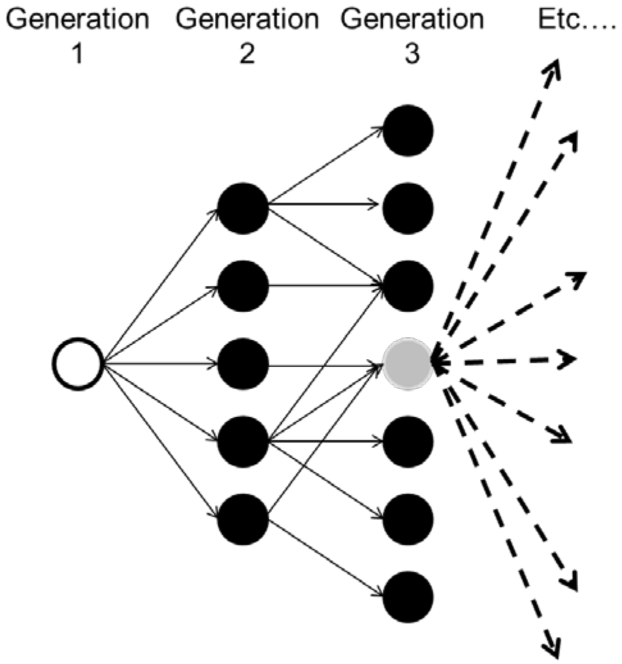


Figure 2.1: Generation of a putative network of compounds using BNICE. Circles represent compounds while arrows represent enzymatic reactions which are predicted through the application of BNICE operators. The white circle represents the initial compound (pyruvate) while the solid grey circle represents a compound of interest (propionic acid). Each generation indicates an additional iterative application of the operators.

2.2.2 Bidirectional Network Generation

To mitigate the combinatorial explosion of a unidirectional search, we utilized a method called bidirectional network generation to reduce the number of compounds that must be searched to build the network (Figure 2.2). This method works as follows. First, the BNICE algorithm is utilized to generate an *n* generational network from the starting compound of interest. The BNICE operators are then reversed and are used to create an *m* generational retro-synthesis network from the target compound. Compounds which are present in both networks are then identified. Next, any compound or reaction not descended from one of the compounds common between the two networks is removed. The merged network is now *n+m* generations and is analyzed by our standard pathfinding methods.

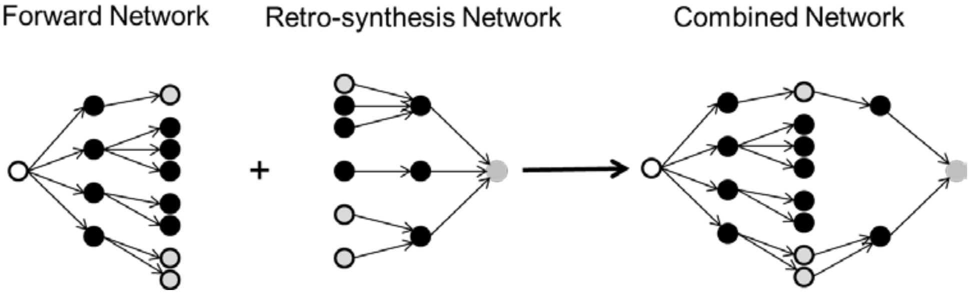


Figure 2.2: Bidirectional network generation.

In these networks the white circle represents the starting compound (pyruvate), the solid grey circle represents the target compound (propionic acid), and the grey circles with a black border represent common compounds between the two networks. The network on the left is an example of a forward network generated “down” from the starting compound. The center network is an example of a retro-synthesis network generated “up” from the target. The network on the right shows how these networks can be combined by finding the common compounds between them. Note that the combined network contains several pathways from the starting compound to the target.

2.2.3 Materials & Chemicals

The experimental work was carried out in the Tyo laboratory by Dr. Miaomin Zhang and is described in the joint publication on which this chapter is based. The full study is detailed here to show the complete sequence of using BNICE to predict novel enzymatic reactions. Chemical compounds, including tris base, acrylic acid, analytical standard grade propionic acid and reduced nicotinamide adenine dinucleotide (NADH), were purchased from Sigma (St. Louis, MO). The protein expression vector pET21a, and the *Escherichia coli* strains DH5α and BL21(DE3) were obtained from Prof. Michael Jewett (Northwestern University).

2.2.4 Enzyme Cloning, Expression and Purification

The *OYE2* gene was cloned from *Saccharomyces cerevisiae* CEN.PK113-5D (from Prof. Jens Nielsen, Chalmers University of Technology) using primers MZ0057 and MZ0058 (Table 2.2). The fumarate reductase gene *frdC* was cloned from *Lactococcus lactis* (ATCC 19435D-5 genomic DNA: American Type Culture Collection, Manassas, VA). The NAD(P)H-flavin reductase gene *fre* was cloned from *E. coli* K-12. The enoate reductase gene *enr* from *Clostridium tyrobutyricum* (GenBankY09960) was codon-optimized using an in-house program by replacing the codons rarely used in *E. coli* and was synthesized by Life Technologies. All genes were fused with N-terminal polyhistidine tags for purification. The vector backbone was amplified from pET-21a using primers MZ0055 and MZ0056 for *OYE2* cloning and equivalent linker-primers for *frdC, fre,* and *enr* (Table 2.2). The amplified gene and pET backbone were assembled to form pET-HIS-OYE2, pET-HIS-frdC, pET-HIS-fre, and pET-HIS-enr by Gibson assembly [30]. The assembly product was transformed into *E. coli* DH5α for construct screening.

E. coli BL21(DE3) was transformed with pET-HIS-OYE2, pET-HIS-frdC, or pET-HIS-fre and was grown in either Luria-Bertani broth (LB) or terrific broth (TB, containing 12 g tryptone, 24 g yeast extract, 4 ml glycerol, 17 mM KH_2PO_4, and 72 mM K_2HPO_4 in 1 L broth) with 100 ug/ml ampicillin at 37^0C to an optical density (OD) of 0.6. Enzyme expression was induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM and incubating the culture at 25 ^{0}C or 30 ^{0}C overnight. pET-enr was oxygen-sensitive, and accordingly was induced in anaerobic Hungate tubes in LB broth flushed with nitrogen to an OD of 0.6, and were induced with 2 mM IPTG at 30^0C overnight to mimic the native anaerobic conditions). Anaerobic procedures were carried out under a nitrogen atmosphere in a UNIlab glove box workstation (M. Braun, Garching, Germany).

For the enzyme activity discovery study, proteins were purified as follows: The culture was spun down at 10,000 x g, 4^0C for 20 min and resuspended in lysis buffer D (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0) at 1:10 (wet weight : volume), frozen at -80^0C, and thawed. Cell lysis was achieved by either sonication or chemical lysis. Sonication of Oye2p: The suspension was sonicated using a Qsonica Q500 sonicator (Newtown, CT) at 50% amplitude for 2 minutes in 10-second pulses, followed by 10-second cooling intervals. Chemical lysis of FrdC, Fre, and Enr: The cell pellet was lysed in BugBuster protein extraction master mix (EMD Millipore, Billerica, MA) at 1/10 of culture volume. For both methods, the cell lysis product was centrifuged at 16,000 x g for 15 minutes at 4 ^{0}C on an Avanti J-E centrifuge (Beckman Coulter, Brea, CA). The supernatant was transferred onto a Qiagen Ni-NTA spin column (Qiagen,Venlo, Netherlands). The His-tagged enzyme was purified by washing the *E. coli* apparatus proteins off the column twice with 600 ul wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH 8.0), and elute twice with 300 ul elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 500 mM imidazole, pH 8.0). Soluble expression was confirmed by SDS-PAGE (Biorad), stained using SimpleBlue (Life Technologies, Grand Island, NY), and imaged on a ChemiDoc system (Biorad). The purified Oye2p was verified to reduce 1-cyclohex-2-enone to cyclohexanone. Oye2p concentration was measured using Bradford assay (BioRad). Molecular mass was confirmed by SDS-PAGE for all enzymes except Enr, which could not be expressed in soluble fraction (Figure 2.5 C). Native enzyme activity was confirmed to verify active protein as follows: Oye2p was found able to reduce 1-cyclohex-2-enone to cyclohexanone, and FrdC with Fre was able to reduce fumaric acid to succinic acid. Enr was carried forth for activity test with crude cell lysate in the event that acrylic acid reduction could be achieved, but the lysate was not found able to reduce its native substrate (crotonic acid).

For large scale purification for the kinetic study, the Oye2p-expressing *E. coli* was grown in TB, induced with 1 mM IPTG at 25^0C, and was spun down at 10,000 x g, 4^0C for 20 min. The pellet was suspended in lysis buffer K (1.5 mM magnesium acetate, 1mM $CaCl_2$, 250 mM NaCl, 100 mM ammonium sulfate, 40 mM Na_2HPO_4, 3.25 mM citric acid, 5% glycerol, 5 mM imidazole, 5 mM β-mercaptoethanol). Cell lysis was performed on an EmulsiFlex-C5 High Pressure Homogenizer (Avestin, Ottawa, Ontario, Canada). After centrifugation at 16,000 x g, 4^0C for 40 min, the supernatant of the cell lysis was collected. The His-tagged Oye2p was purified first by HisTrap FF 5 ml column (GE Healthcare Bio-Sciences, Pittsburgh, PA), and second by size exclusion purification on HiLoad 26/600 Superdex 200 prep grade column (GE), using the AKTAExpress purification system (GE). For affinity purification, the elution buffer contained 10 mM Tris-HCl, 500 mM NaCl, and 500 mM imidazole. The eluent of size exclusion purification contained 10 mM Tris-HCl, 500 mM NaCl, and 5 mM β-mercaptoethanol. Collected protein eluent was further concentrated in Sartorium molecular weight cutoff filter (10 KDa, 20ml) at 4000 x g, 4^0C. Size and purity confirmation of purified Oye2p was carried out as the small scale purifications.

2.2.5 Aerobic and Anaerobic Oye2p Reaction Conditions

Substrate solutions (100-400 mM acrylate) were prepared by dissolving acrylic acid in deionized water and deprotonating with stoichiometric amounts of 6 N sodium hydroxide. For enzyme activity discovery, aerobic reactions with Oye2p were carried out by incubating 1 µM Oye2p in 50 mM Tris-HCl buffer, pH 7.5, with 300 mM NaCl, 10 mM NADH and 10 mM acrylate, at 30^0C, overnight. Controls were performed either without the cofactor NADH or without Oye2p to determine any spontaneous reduction of acrylic acid to propionic acid.

Kinetic assays were performed anaerobically. Solutions of Oye2p, NADH and acrylate were freshly prepared in separate anaerobic Hungate tubes (Chemglass, Vineland, NJ), flushed with nitrogen and mixed under nitrogen atmosphere in an Atmosbag glove bag (Sigma). Reactions with 10 µM Oye2p, 400 µM NADH and acrylate (1, 10, 20, 50, 100 mM) were performed at room temperature (22 ^{0}C) for 80 minutes. Samples were extracted and analyzed on gas chromatography-mass spectrometry (GC-MS) as described below. The acrylate turnover rate was calculated by taking the average rate of propionic acid generation within the 80-minute period. The kinetics data of acrylate reduction by Oye2p in the presence of NADH were fitted to a Michaelis-Menten equation using the nonlinear regression function *nlinfit* in Matlab (Mathworks, Natik, MA).

The extent of anaerobic reaction for Oye2p with 50 and 100 mM acrylate was also measured at progressive times both analytically (on GC-MS) and spectrophotometrically to track its evolution. NADH oxidation by OYE with 50 and 100 mM acrylate was monitored at 340 nm on a Shimadzu UV-2450 spectrophotometer (Kyoto, Japan), in sealed quartz cuvettes with 1 cm path lengths (VWR).

2.2.6 Propionic Acid Measurement

Reactions were prepared for analysis by adjusting to pH 2 with 6 N hydrochloride and extracting with 1:1 (v:v) ethyl acetate with vigorous vortexing. The ethyl acetate phase was collected and concentrated from a starting volume of 2 ml to 250 □l in a Vacufuge vacuum desiccator (Eppendorf, Hauppauge, NY) for propionic acid quantitation. A final concentration of 37 µM 2-ethylbutyric acid was added to each analytical sample as internal standard. Solutions of 10, 40, 100 and 200 µM propionate were extracted by the same procedure and analyzed along with the reaction products and used to create a propionic acid standard curve.

Propionic acid quantitation was performed on an Agilent 5973 GC-MS (Agilent, Santa Clara, CA) installed with a FFAP 30 m x 0.25 mm (0.5µm) column (Agilent). The GC was operated at 2 ml/min He flow, 250 ^{0}C injector temperature, with a 2 µl injection volume, and 1:25 split ratio. The GC oven was programmed to hold 2 minutes at 55 ^{0}C, and then rise 20

^{0}C/min to 240 ^{0}C, where it stayed for 5 minutes. The MS was operating at 230 ^{0}C, -70eV, and a 3.5-minute solvent delay was imposed.

Propionic acid had an average retention time of 8.17 minutes. The m/z 74 (±0.5) ion was extracted for peak integration.

2.2.7 Acrylate Binding Affinity Assay

The flavin redox center in Oye2p changes spectral absorbance upon enzyme-substrate binding [31]. The UV-visible spectra from 300- 600 nm were recorded with variable acrylate concentrations (0-100 mM) and 10 µM Oye2p to investigate their binding affinity. This was performed aerobically and in the absence of NADH. Dissociation constant K_d was calculated from $K_d = ([OYE]_0 – [OYE\text{-}acrylate]) [acrylate] / [OYE\text{-}acrylate]$, where $[OYE]_0$ is the starting Oye2p concentration (10 µM), [acrylate] is the acrylate concentration, and [OYE-acrylate] is the concentration of Oye2p-acrylate complex, which was determined by the absorbance change at long wavelengths (504 nm).

2.3 Results

2.3.1 Pathways from Pyruvate to Propionic Acid Predicted using BNICE

BNICE was used to predict three, four, and five-step enzymatic pathways from pyruvate to propionic acid. In the three-step pathway (Figure 2.3), the first reaction adds a carboxyl group to pyruvate and consumes NADH to reduce one carbonyl group to a hydroxyl group to form malic acid. The reverse of this reaction is known to be catalyzed by several dehydrogenases including the NAD-dependent malate dehydrogenase in *E. coli* and the NAD-linked malic enzyme in *E. coli* [32]. However, no enzyme is known which catalyzes the reaction in this direction at an appreciable rate under most physiological conditions. In the second reaction, malic acid loses a carboxyl group in the form of carbon dioxide and is dehydrated to form acrylic acid. This reaction has not been observed experimentally. The final reaction is the reduction of acrylic acid to propionic acid. There is a report that this activity can be catalyzed by fumarate reductase in *Lactococcus lactis*, though the activity was low [33]. The reduction of acrylic acid to propionic acid will be common across many predicted pathways and will subsequently be discussed in more detail.

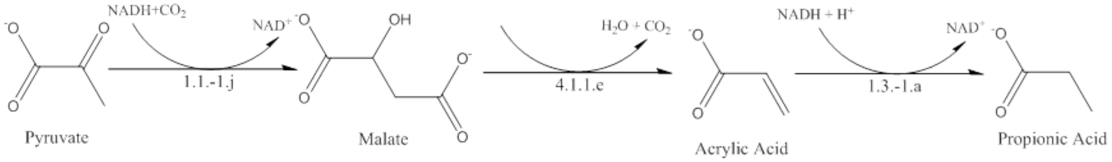


Figure 2.3: Three-step pathway between pyruvate and propionic acid.

The associated BNICE operator is shown for each reaction. Predicted cofactors for each reaction are also labeled.

Bidirectional network generation was used to predict four-step reaction pathways by running BNICE forward three generations and backwards one generation. Seven such pathways were discovered that used a combination of sixteen different reactions (Figure 2.4). Five of these reactions are native enzymatic reactions upon which the BNICE operators were based while the remaining eleven are promiscuous reactions (Figure 2.4).

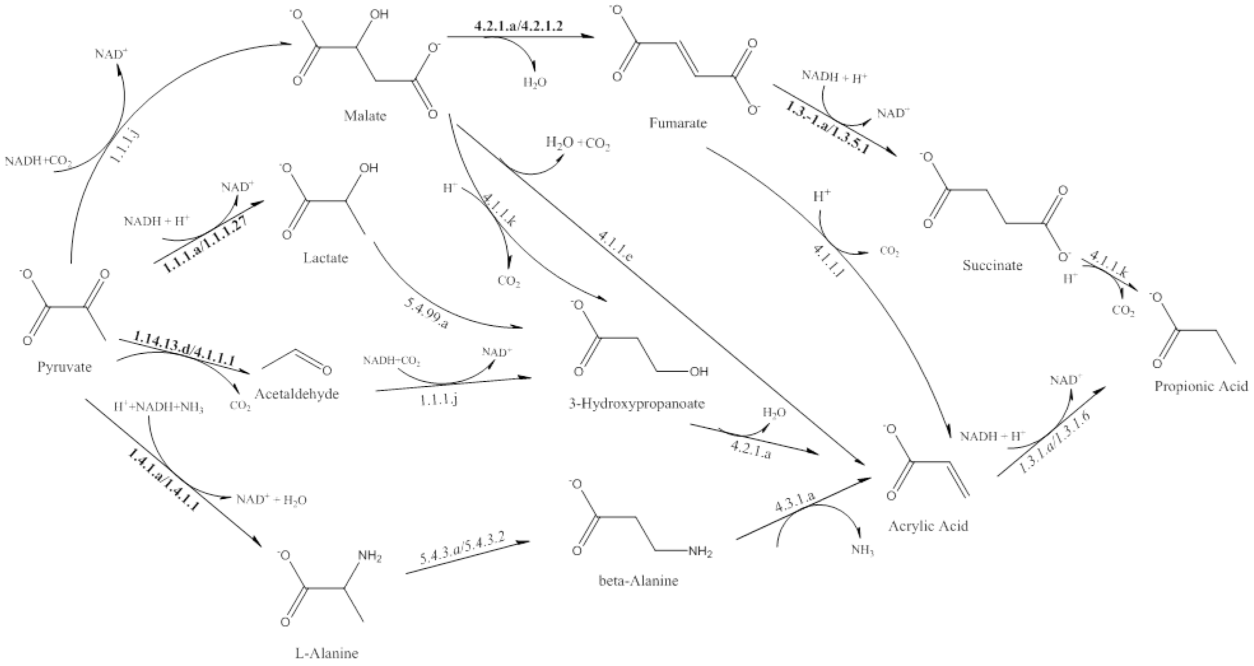


Figure 2.4: Four-step pathways between pyruvate and propionic acid.

The associated BNICE operator is shown for each reaction. For known reactions, both the BNICE operator and the full i.j.k.l designation of an enzyme which performs the reaction are shown and are bolded. Promiscuous reactions which were not in the BNICE training set but were confirmed by a subsequent analysis of the literature are labeled in italics with both the BNICE operator and a full i.j.k.l designation of an enzyme which performs the reaction. Predicted cofactors for each reaction are also labeled.

A literature analysis revealed that two of the promiscuous reactions had previously been observed experimentally (Figure 2.4). These reactions were not in the training set for BNICE operators and therefore illustrate correct predictions by BNICE of enzyme promiscuity. One of these reactions is the transfer of the secondary amine in L-alanine to the terminal carbon to form beta-alanine. This reaction has been experimentally shown to be catalyzed through promiscuous activity of lysine 2,3-aminomutase in *Clostridium subterminale* [34]. The other is the previously mentioned reduction of the olefinic bond in acrylic acid to form propionic acid catalyzed by fumarate reductase.

Six of the seven predicted pathways consisting of four reaction steps or fewer require the reduction of acrylic acid to propionic acid. To investigate whether this pattern continues for longer pathways, five-step reaction networks were generated based on three forward BNICE generations from pyruvate and two retrosynthesis generations from propionic acid. Among the 1,410 pathways of five or fewer reaction steps, 89.3% involves acrylic acid reduction as the last step (Table 2.1).

Table 2.1: Reactants of the final step in predicted five-step pathways.

Reactants of the final step	# of pathways
Acrylic Acid	1259
Propanal	36
Propanyl-CoA	32
Succinic Acid	24
Total Pathways	1410

The results for the five-step reaction network illustrate the importance of the predicted reaction for the reduction of acrylic acid to propionic acid. The prevalence of this reaction in the BNICE network made it a natural and interesting candidate for experimental validation. A list of candidate enzymes was generated by annotating enzymes which catalyze reactions that obey the BNICE operator for this reaction (1.3.1.a). A manual literature survey of these enzymes narrowed the list to three enzymes of particular interest: NADPH dehydrogenase (EC 1.6.99.1), fumarate reductase (EC 1.3.1.6), and 2-enoate reductase (EC 1.3.1.31).

NADPH dehydrogenase (EC 1.6.99.1) was chosen for experimental validation because it is known to catalyze the native reactions of enzymes in the EC 1.3.1.a category and can reduce many α,β-unsaturated aldehydes [35]; its activity against α,β-unsaturated carboxylic acids has not been documented before. Oye2p was specifically selected because it is the *S. cerevisiae* NADPH dehydrogenase responsible for reductive detoxification of acrolein [35, 36] and has been shown to be active against a broad range of substrates [35, 37, 38]. The enoate reductase, Enr from *Clostridium tyrobutyricum* (CAA71086.1) was selected because it has been recombinantly expressed under T_5 promoter in *E. coli* M15(pREP4) [39]. The *L. lactis* fumarate reductase, FrdC, was included to reproduce the acrylate reduction assay conducted by Hillier, et al. [33]. FrdC (BAL51025.1) is the only gene annotated as a fumarate reductase in the *Lactococcus lactis* genome (AP012281.1). Fre is an *E. coli* enzyme used in the FrdC assay to reduce FAD to $FADH_2$, which FrdC uses as a cofactor.

2.3.2 Oye2p Showed Catalytic Activity for Reducing Acrylic Acid to Propionic Acid

Oye2p, Enr, and Fre/FrdC were cloned, expressed, purified, and assayed against acrylic acid. His-tagged Oye2p, Fre, and FrdC were successfully produced and purified, as confirmed by SDS-PAGE (Figure 2.5 A-B). Enr could not be solubly expressed, even during anaerobic expression, and was subsequently discarded (Figure 2.5 C). Catalytic activity was tested using 1 µM enzyme (and 1:1 co-factor Fre for the FrdC assay), 10 mM acrylate, and 10 mM NADH (substrate:cofactor = 1:1), incubated aerobically at 30 ^{0}C overnight. While Fre/FrdC did not produce detectable reduction (Figure 2.6), Oye2p was able to reduce acrylic acid to produce propionic acid (Figure 2.7). Propionic acid, verified by comparison to MS spectra of authentic standards (Figure 2.7 B-D), was produced to 16.96 µM, significantly higher than the background propionic acid in the controls without either NADH or Oye2p (Figure 2.7 A). Background propionic acid depended on acrylate concentration (data not shown), and was likely a result of spontaneous hydrogenation of the acrylate olefinic bond during sample acidification/extraction.

Table 2.2: Oligonucleotides used for constructing expression plasmids.

Primer	Sequence	Gene/ Construct
For gene cloning:		
MZ0057 MZ0058	5'-CCTGTATTTTCAAAGCCCATTTGTTAAGGACTTTAAGCCACAAG-3' 5'-GGCTTTGTTAGCAGTTAATTTTTGTCCCAACCGAGTTTTAGAG-3'	*OYE2*
MZ0061 MZ0062	5'-CCTGTATTTTCAAAGCAAAATTTGGACTAAACTAGGCTTGCTAACG-3' 5'-	*frdC*

	GCTTTGTTAGCAGTTAATTGCTTGTTTTAGCATAGGCCGCAGA-3’	
MZ0065 MZ0066	5’-CCTGTATTTTCAAAGCACAACCTTAAGCTGTAAAGTGACCTCG-3’ 5’-CTTTGTTAGCAGTCAGATAAATGCAAACGCATCGCCAAA-3’	*fre*
MZ0013 MZ0015	5’-GCTTTGTTAGCAGTTAGCAGTTTAAGCCAATCTCG-3’ 5’-GTCTATAATGCATCATCATCATCATCACAAAAACAAAAGTTTGTTCGAGCCT-3’	*enr*
For pET-21a backbone amplification:		
MZ0055 MZ0056	5’-TGGGACAAAAATTAACTGCTAACAAAGCCCGAAAGGAAGC-3’	pET-HIS-OYE2

	5’-TCCTTAACAAATGGGCTTTGAAAATACAGGTTTTCGTGATGATGATG ATGATGCATATGTATATCTCCTTCTTAAAGTTAAAC-3’	
MZ0059 MZ0060	5’-CTAAAACAAGCAATTAACTGCTAACAAAGCCCGAAAGGAAGC-3’ 5’-TAGTCCAAATTTTGCTTTGAAAATACAGGTTTTCGTGATG-3’	pET-HIS-frdC
MZ0063 MZ0064	5’-CGTTTGCATTTATCTGACTGCTAACAAAGCCCGAAAGGAAGCT-3’ 5’-AGCTTAAGGTTGTGCTTTGAAAATACAGGTTTTCGTGATG-3’	pET-HIS-fre

MZ0012	5’-CTTAAACTGCTAACTGCTAACAAAGCCCGAAAG-3’	pET-HIS-enr
MZ0014	5’-TTTGTGATGATGATGATGATGCATTATAGACCTCCTTAGAAAGCGGA ATTGTTATCCGCTCACAATTC-3’	

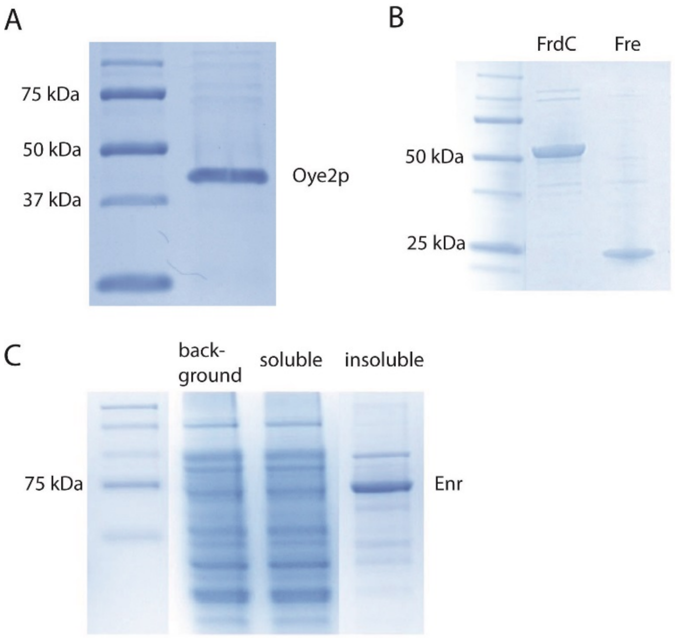


Figure 2.5: SDS-PAGE analysis of Oye2p, FrdC, Fre and Enr expression.

A) Purified His-Oye2p. The expected molecular weight of His-Oye2p is 45 kDa. B) Purified His-FrdC and His-Fre. The expected molecular weights are 55 kDa and 26 kDa, respectively. C) Enr expression. After induced anaerobically in *E. coli*, Enr (74 kDa) was not present in the soluble fraction of cell lysate, but in the insoluble fraction. The background was the soluble lysate of *E. coli* transformed with blank pET21a vectors.

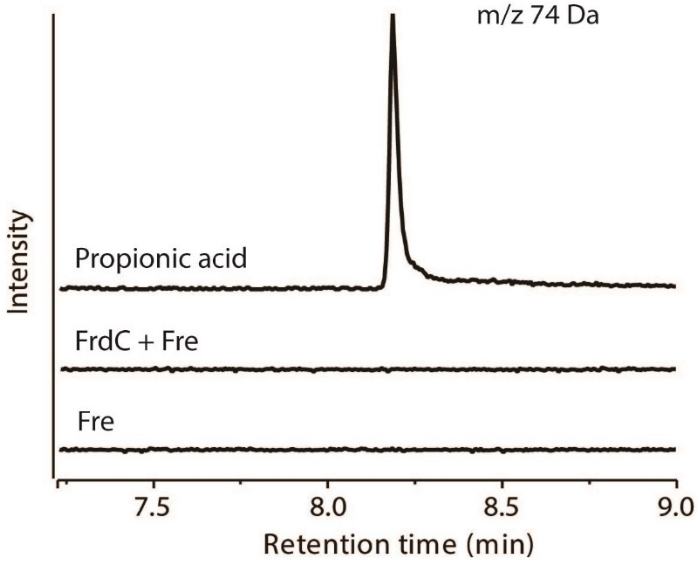


Figure 2.6: Catalytic Activity of FrdC.

FrdC was inactive against acrylic acid. Overnight incubation of FrdC with acrylic acid did not produce the propionic acid peak at 8.17 minutes in the m/z 74 extracted ion chromatogram.

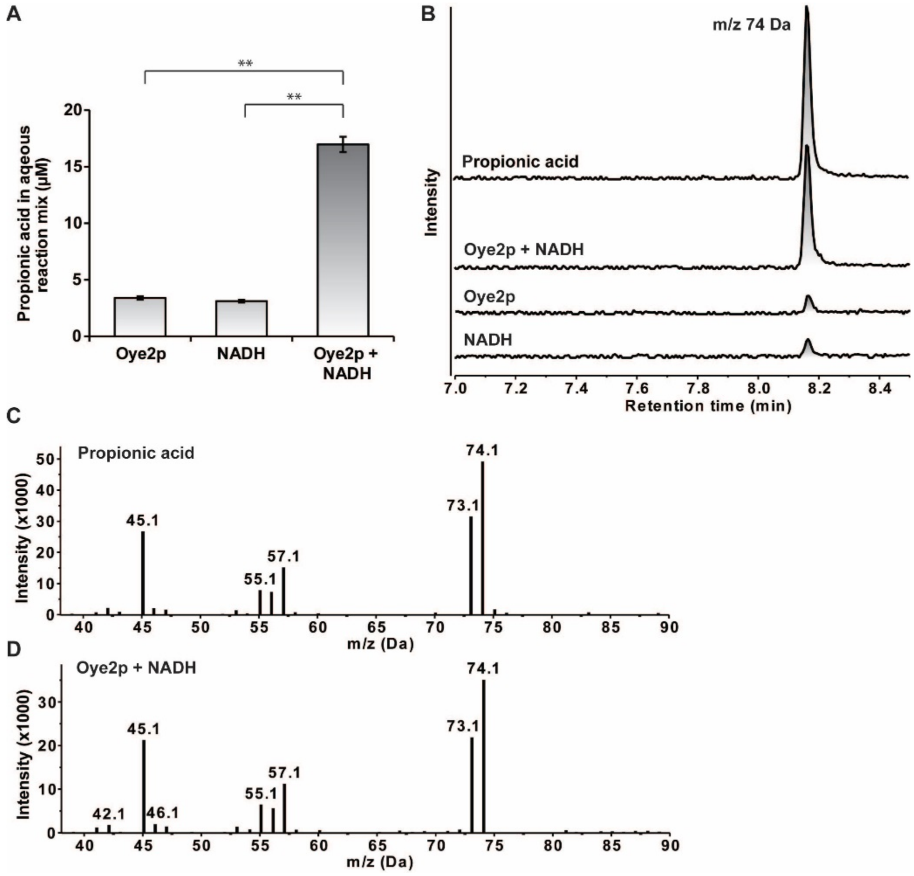


Figure 2.7: Catalytic Activity of Oye2p.

Oye2p reduces acrylic acid to propionic acid. A) Estimated propionic acid concentrations (by integrating m/z 74 Da peaks at 8.17 minutes in the GC-MS chromatograms). Oye2p in the presence of NADH (n = 5) produced more propionic acid than either Oye2p (n = 5; **, $p < 5 \times 10^{-5}$, two-tailed Student t-test) or NADH (n = 5; **, $p < 5 \times 10^{-5}$, two-tailed Student t-test) alone. B) The m/z 74 Da extracted ion chromatogram of Oye2p-acrylate reaction showed the propionic acid peak at 8.17 minutes, matching that of authentic propionic acid standard (100 µM propionic acid in ethyl acetate). C-D) Mass spectra of propionic acid standard (D) and Oye2p-acrylate reaction (E) at 8.17 minutes.

The acrylic acid turnover by Oye2p was found to be slow. The kinetics of Oye2p acrylate reduction were characterized using propionic acid concentration data generated using 1-100 mM acrylate. The Michaelis constant K_m was estimated as 5.92 ± 6.23 mM (95% confidence interval, Figure 2.8 A). The maximum reaction velocity V_{max} was 0.0041 ± 0.0010 μM acrylate/s/μM OYE (95% confidence interval; compared to cyclohexanone reduction by OYE1, oxidative half-reaction: $k_{cat} = 102\ s^{-1}$, $K_d = 32\ \mu M$ [31]). Reaction rates at higher acrylate concentrations were also confirmed by monitoring NADH consumption (Figure 2.8 B), and measuring propionic acid production at progressive times (Figure 2.8 C-D). Both showed a linear regime from 15- 65 minutes. Acrylate binding to Oye2p results in increased absorbance at 504 nm (Figure 2.8 E). Using this property, the acrylate binding constant (K_d) was estimated as 8.93 ± 1.58 mM (95% confidence interval), close to the K_m estimate.

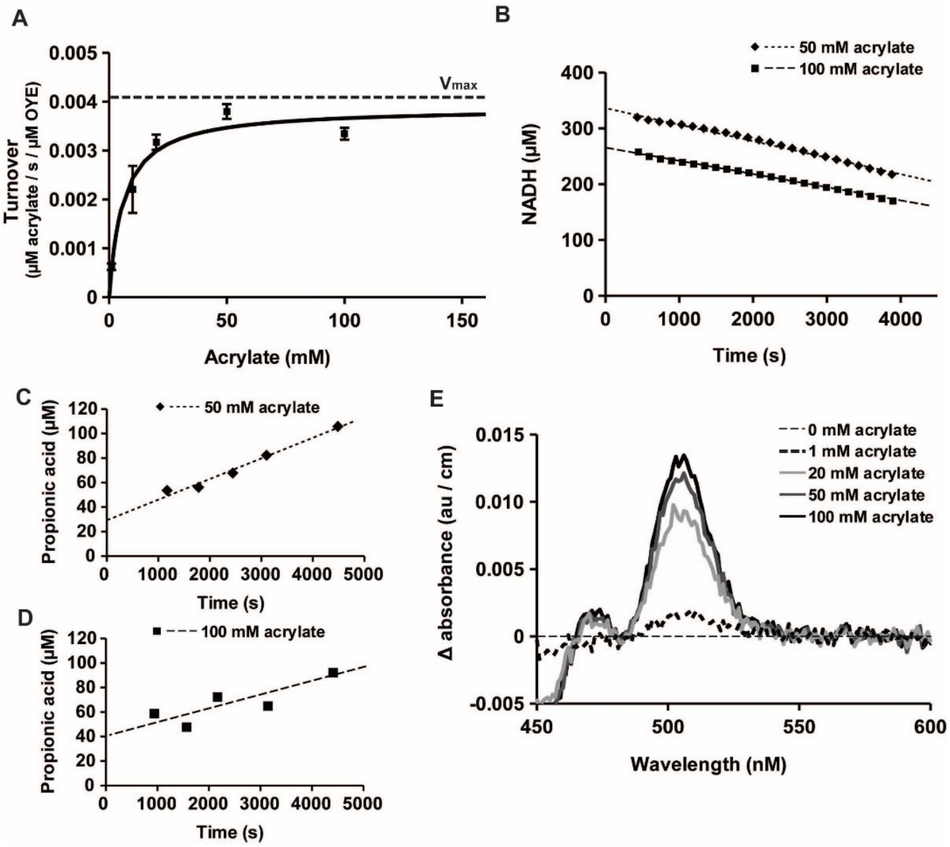


Figure 2.8: Kinetic characterization of Oye2p-acrylate reaction.

A) The apparent propionic acid production rate within the first 80 minutes of reaction as a function of acrylate concentration. Kinetic assays were conducted anaerobically with 10 µM Oye2p, 400 µM NADH and 1-100 mM acrylate. Error bars are standard errors (n=3). The data were fitted to a Michaelis-Menten equation (solid line). B) NADH consumption showed linear trend from 8- 67 minutes (diamonds and dotted line: 50 mM acrylate, R^2 = 0.99; squares and dashed line: 100 mM acrylate, R^2 = 0.99). C-D) Time courses of propionic acid production were also linear from 16-75 minutes with 50 mM acrylate (C, R^2 = 0.98) and 100 mM acrylate (D, R^2 = 0.70). E) Oye2p absorbance spectrum shifts in the presence of 1-100 mM acrylate. The binding between Oye2p and acrylate caused an absorbance increase at 502-506 nm.

2.3 Discussion

2.3.1 Pathway Prediction for Propionic Acid Production

Using BNICE, we predicted seven novel four-step pathways for the production of propionic acid from pyruvate. Five of these seven pathways avoided the production of the undesired side products lactic acid and succinic acid. The reduction of the olefinic bond in acrylic acid to yield propionic acid is the common final step for all six four-step pathways which avoided producing succinic acid. The prediction of five-step pathways was performed to evaluate the potential importance of acrylic acid reduction in propionic acid synthesis. Over 89% of the 1410 five-step pathways contain this last step. It is possible that a broader array of options would emerge given a larger operator set.

This convergence of pathways indicates that converting acrylic acid to propionic acid is a key reaction to implement for propionic acid production. Beyond our proposed pathways, acrylic acid biosynthesis is already being commercialized [40]. It would be straightforward to extend the acrylic acid enzymatic process to produce propionic acid as well, if an enzyme that is able to convert acrylic acid to propionic acid could be identified. Therefore we focused on this reaction for experimental validation of the BNICE prediction.

In the only four-step pathway that does not involve acrylic acid reduction, the single step decarboxylation of succinic acid to propionic acid is a novel reaction proposed by BNICE. While no enzyme has yet been experimentally proven to catalyze this reaction, decarboxylation is known to be catalyzed by a few enzymes on substrates with slight differences. These enzymes include oxaloacetate decarboxylase (EC 4.1.1.3, oxaloacetate to pyruvate), oxalosuccinate decarboxylase (EC1.1.1.42, oxalosuccinate to 2-oxoglutarate), and aspartate 1-decarboxylase (EC 4.1.1.11, L-aspartate to beta-alanine)[41-43]. The transformation of succinic acid to propionic acid has also been found to occur in multiple reaction steps, as in the native production of propionic acid by bacteria of the genus *Propionibacteria* [44]. Again, a single-step conversion of succinic acid to propionic acid may be preferable to multi-step reactions, because having fewer reaction steps reduces side-product formation and can increase product titer.

Comparing BNICE predicted pathways with known native pathways can be informative in helping us choose the best pathways to pursue on an industrial scale and bringing new insights into our understanding of biochemical synthetic pathway construction. The native *Clostridium propionicum* pathway for propionic acid production is the following five-step pathway: pyruvate→lactate→lactyl-CoA→acryloyl-CoA→propionyl-CoA→propionic acid [45]. This pathway was among the five-step pathways predicted by BNICE (data not shown). As already described, BNICE has produced seven pathways with fewer steps than the *Clostridial* pathway (Figure 2.3, 2.4). Most of these novel pathways involve converting acrylic acid to propionic acid without the need of CoA. Here, BNICE's capability of predicting promiscuous enzyme activities has allowed it to identify potential pathways which bypass the addition and removal of the CoA in the native pathway and reduced the number of reaction steps required to convert pyruvate to propionic acid. On the other hand, BNICE predicted that, instead of a single-step conversion of lactyl-CoA to acryloyl-CoA, excluding CoA from the pathway necessitates transforming lactate to acrylic acid in two steps, first from lactate to 3-hydroxypropenoate, then from this intermediate to acrylic acid through dehydration (Figure 2.4). This suggests that CoA is required to stabilize the single-step conversion of lactyl to the acryloyl structure. It also highlights the ability of BNICE to discern between those spectator atoms which are required to stabilize a reaction and those that may be unnecessary for the reaction.

2.3.2 The Discovery and Characterization of Oye2p Acrylic Acid Reducing Activity

We show that in the presence of NADH, Oye2p can reduce acrylic acid to propionic acid. Under aerobic conditions, Oye2p propionic acid yield was 0.17% overnight. This reactivity of an old yellow enzyme (OYE) against a mono-carboxylic acid was previously reported to be unlikely [31, 38, 46, 47]. However, in previous studies, the substrate concentrations used in assays were rarely above 5 mM, making product detection (likely less than 3 µM propionic acid) difficult. We were able to identify this activity by making a concerted effort to find the product based on BNICE prediction. The Oye2p reduction of α,β-unsaturated carboxylic acids or aldehydes occurs in a Ping-Pong mechanism that involves two half reactions: 1) the reductive half reaction, where NADH binds to OYE and transfers two electrons to the OYE-bound flavin mononucleotide (FMN), and 2) the oxidative half reaction, where the substrate binds to OYE and receives two electrons from the reduced $FMNH_2$. Molecular oxygen is a known oxidizer of reduced OYE [46, 48]; therefore, the Oye2p-acrylate reaction kinetics were measured under anaerobic conditions. The monitoring of both NADH consumption and propionic acid production showed that, after an initial phase of nonlinearity, the reaction proceeds at a constant rate. This apparent steady-state kinetics behavior results from the low turnover rate of the oxidative half reaction (10^{-3} µM acrylate/s/µM OYE). The reductive half reaction with NADH (K_d 100 µM, k_{cat} 0.9 s^{-1} [31]) is two orders of magnitudes faster than the oxidative half reaction with acrylate. This implies that

Oye2p spends more time in the reduced state (after NADH reduces FMN) than the oxidized state (after FMN reduces acrylic acid), which was confirmed by the fact that the Oye2p bound FMN stayed colorless when NADH was present. The K_d of the OYE-acrylate complex (8.93 ± 1.58 mM), measured by the Oye2p spectral change at variable acrylate concentration, is not smaller than the K_m of the OYE-acrylate reaction. This indicates that the conversion of acrylate to propionate is much slower than the dissociation of OYE-acrylate complexes. Hence, the slowness of the OYE-acrylate reaction is likely due to weak enzyme-ligand binding rather than a slow catalytic step.

Oye2p turnover of acrylic acid can potentially be improved by modifying substrate binding and transition state coordination (catalysis). Oye2p binds α,β-unsaturated aldehydes through H191 and N194, two residues in the hydrophobic ligand binding site that form hydrogen bonds with the carbonyl oxygen. This places the β-carbon in an excellent position to receive a hydride from flavin N-(5) [48]. The α-carbon is then protonated by Y196. Variations of the ligand binding residues in other OYE family enzymes suggest that a more electropositive environment is required to improve carboxylic acid binding. For example, the 12-oxophytodienoate reductase 1 and 3 from tomato (OPR-1/3) and the *Bacillus subtilis* homolog of OYE, YqjM, have been reported to accept some dicarboxylic acids and their esters as substrate [49]. Both ligand binding residues in these enzymes are positively charged (OPR1: H187, H190; OPR3: H185, H188; YqjM: H164, H167) [50]. This implies that site-directed mutagenesis of Oye2p may improve the catalytic rates. On the other hand, better substrate conversion has been achieved with an electron withdrawing group at the α position of the substrate that activates the α,β-carbon double bond, e.g. a halogen [47] or a carboxylic acid (diacids) [49, 51]. Since both the Oye2p-acrylate binding and the reduction of α,β-carbon double bonds are inefficient, rational engineering of OYE for higher propionic acid yield would have to 1) enhance substrate binding by making the ligand binding pocket more electropositive, and 2) introduce amino acid residues in the active site that can coordinate acrylic acid, such that the necessary electron withdrawing groups would stabilize the transition state. The fumarate reductase (1.3.1.6) from *Lactococcus lactis* has been shown to have low reactivity against acrylate [33], but we were unable to reproduce this result (with 1 μM enzyme and 2 mM acrylate). Ambiguity in the particular strain of *L. lactis* may explain this discrepancy.

2.3.3 Enzymatic Pathway Discovery using BNICE

We have demonstrated the use of the BNICE bioinformatics software suite to identify potential enzymatic pathways from pyruvate to propionic acid and showed that the generalized BNICE operators can predict reactions not previously found in literature. Our analysis identified the conversion of acrylic acid to propionic acid to be a key reaction step such that realizing this reaction would make many avenues for propionic acid bio-production possible. We found that Oye2p can reduce acrylic acid in the presence of NADH at a rate of 10^{-3} μM acrylate/s/μM OYE. This activity has not been previously observed. Although the turnover rate is low for practical purposes, further enzyme engineering should be possible. We believe that the strength of BNICE lies in its ability to include promiscuous enzyme activities not previously documented in pathway searches.

Chapter 3: Automatic Generation of Chemical Operators

3.1 Introduction

Researchers have been increasingly investigating the use of enzymatic reactions to find alternative routes to compounds of interest. Our knowledge of enzymatic chemistry has expanded rapidly in recent years. Progressively more researchers are turning to computational tools in order to rapidly investigate potential biosynthetic routes. Databases such as KEGG, BRENDA, and MetaCyc have been developed to speed the search through known enzymatic chemistry [7, 8, 15]. Additionally, pathway tools such as DESHARKY, FMM, CarbonSearch, and MetaHype have been developed to rapidly search through these databases and identify promising enzymatic pathways [2-5]. However, much enzymatic chemistry has yet to be discovered, and there is increasing interest in using computational techniques to predict *de novo* enzymatic reactions which have not yet been observed and use these predictions to guide the search for new enzymatic activity.

One method of exploring *de* novo enzymatic chemistry is to use generalized reaction rules known as *chemical operators* [6]. These operators describe reactions in general terms allowing for the prediction of reactions that are similar, but not identical, to known reactions. The atoms described in reaction operators can be separated into two categories: *reaction site atoms* and *spectator atoms*. Reaction site atoms are those atoms that change bonds across the reaction. These atoms must be included in an operator in order for the predicted *de novo* reactions to be coherent, or otherwise reactions may be predicted that form or break the bonds of atoms not present in the predicted substrates. Spectator atoms are those atoms which do not change bonds across the reaction. Certain spectator atoms often stabilize reaction intermediates and are therefore important for the prediction of feasible *de novo* reactions. The inclusion of spectator atoms is the primary method by which the specificity of a reaction operator is controlled. Requiring the presence of many spectator atoms results in the prediction of fewer reactions, but those predicted are often more similar to previously observed reactions and therefore may be more likely to occur. Requiring few spectator atoms to be present typically allows for the prediction of many *de novo* reactions, but these predicted reactions can be dissimilar to known enzymatic reactions, reducing the likelihood that they can be performed experimentally.

Several computational tools have been developed which utilized chemical operators to predict *de novo* enzymatic chemistry [6, 52-55]. Some initial success has already been achieved utilizing these tools to identify novel enzymatic chemistry [56]. However, the operators used by these programs often must be created manually. This is a slow process which requires researchers to comb through examples of enzymatic chemistry to identify patterns, thereby potentially struggling to keep up with the rate of discovery of new enzymes. This in turn can lead to large segments of enzymatic chemistry not being described by any chemical operator. This problem is exacerbated by the fact that different applications often require different sets of operators. For example, one program known as the Biological Network Integrated Computational Explorer (BNICE) has been utilized to study problems in bioremediation, industrial biotechnology, and biomimetic catalysis [24, 26, 57]. The requirements for the specificity and allowable chemistry in these different applications can vary greatly. Ideally, the chemical operators utilized should be tailored for the particular application of interest. However, making a new set of chemical operators for each new application further reduces the speed at which *de novo* chemistry can be explored, reducing the utility of these tools.

To address these issues, methods for the automatic generation of chemical operators have been developed [58, 59]. Unfortunately, these techniques either do not include spectator atoms [59], causing the specificity of the operator set to be fixed at the most general rules possible, or include spectator atoms based upon parameters whose correct value for a given application is not intuitive [58].

In this work, we introduce a new method for computationally generating chemical operators. We begin by exploring the limits of operator specificity by generating two operator sets based upon the reactions in the MetaCyc database of enzymatic chemistry: one as specific as possible and one as general as possible. Both of these operator sets are capable of describing every relevant reaction in MetaCyc. We demonstrate the utility of these limiting operator sets by utilizing BNICE to apply them to the KEGG database of enzymatic chemistry and the iMM904 *Saccharomycotina cerevisiae* (yeast) model [9].

We next introduce an algorithm for the generation of operators of intermediate specificity. Importantly, the user determines the specificity of these operators by separating reactions into groups. Specific operators can be generated by only grouping together reactions whose substrates are very similar, while general operators can be created by grouping together reactions whose substrates are dissimilar. We finally utilize this algorithm to generate a set of operators based upon groupings from the limiting operator sets. The resulting operators are capable of describing every relevant reaction in MetaCyc. Finally, we compare the specificity of this intermediate operator set with that of the limiting operators by utilizing BNICE to predict *de novo* reactions involving the 22 amino acids.

3.2 Methods

3.2.1 Format of Automatically Generated Operators

In this work all computationally generated operators were represented as BNICE chemical operators [6]. BNICE has the capability of representing operators in three dimensions, which allows chirality to be taken into account; however, all operators were constructed in a two-dimensional format in the present work. A BNICE operator contains three basic parts: a *bond-electron matrix* (BEM) [60] representing the reaction site and spectator atoms of the substrates, a *reaction operator matrix* indicating which bonds change across the reaction, and a *list of acceptable atom types* for each atom in the BEM. For clarity and convenience each of these elements will be briefly described here.

The BEM is a matrix representation of the substrates of the reaction. Each entry in the matrix represents a bond in the compound with the row/column of that entry indicating the atoms that are connected by that bond and the value of the entry indicating the order of the bond.

The reaction operator matrix is a matrix representation of the bond changes during the reaction. Like the BEM, each entry in this matrix corresponds to a bond in the substrates with the row/column of the entry indicating which atoms the bond is between and the value of the entry indicating how the bond order changes across the reaction. This matrix can be obtained by representing both substrates and products as BEMs and then subtracting the substrates' BEM from the products' BEM. In order for the resulting matrix to correctly describe the bond changes across the reaction, the rows/columns in both BEMs must correspond to the same atoms. Identifying which atom in the substrates corresponds to each atom in the products is known as the *atom-mapping problem*. Fortunately, much recent work has been done developing algorithms to solve this problem [61-63]. For this work, we utilized the atom-mappings provided by MetaCyc which are generated using the minimum weighted edge-distance (MWED) metric [61]. Finally, the list of acceptable atom types indicates the allowable chemical elements for each atom (row/column) in the matrices and allows for additional specifications on these atoms as will be described below.

3.2.2 Identifying and Labeling Cofactors

Small molecule cofactors such as adenosine triphosphate (ATP) and water often complicate the search for enzymatic pathways by creating short routes between metabolic compounds which are an artifact of their presence in many reactions rather than a true route for the enzymatic production of a compound [64]. BNICE prevents these artifacts by labeling atoms from common cofactors in the list of the acceptable atom types of the reaction operators. This ensures that the reaction described in the operator can only occur in the presence of the specified cofactor(s) and allows BNICE to prevent the generation of these artifacts when searching for pathways. In order to make operators that are compatible with BNICE, we need a method for automatically identifying and labeling cofactors in the reactions.

To this end, our operator generation program maintains two lists of common cofactors. One list contains compounds which are always labeled as a cofactor whenever they appear in a reaction. This includes mostly very small molecules such as water, oxygen, and carbon dioxide. The other list contains compounds which appear as part of a cofactor pair. These compounds are only labeled as cofactors if the other half of the pair appears on the opposite side of the reaction. Such pairs include ATP and adenosine diphosphate (ADP) and oxidized nicotinamide adenine dinucleotide (NAD^+) and reduced nicotinamide adenine dinucleotide (NADH). In order to increase the applicability of our operators some cofactors are treated as interchangeable such as NADH and nicotinamide adenine dinucleotide phosphate (NADP) or ATP and guanosine triphosphate (GTP). The full list of the cofactors and cofactor pairs can be found in Appendix A.

3.2.3 The Limiting Reaction Rules

As stated above, the generality of a chemical operator is controlled by the presence of spectator atoms. If many spectator atoms are included in the operator then fewer substrates meet the requirements for the operator, causing the prediction of fewer *de novo* reactions. Conversely, if fewer spectator atoms are included then more *de novo* reactions are predicted.

The limiting cases for reaction operator specificity can therefore be explored by investigating the limiting cases for the inclusion of spectator atoms, namely the inclusion of all the spectator atoms in a reaction and the inclusion of no spectator atoms. In this work we will call the former operator set *exact operators* because they specify each reaction rule exactly, allowing no *de novo* reactions to be predicted. We will call the latter operator set re*action site operators* because they only include the reaction site atoms.

3.2.4 Computational Generation of Exact Operators

A program was created in Python 2.7 which can generate an exact operator for any reaction for which an atom mapping is known. This program uses the following procedure.

1. Create the BEM for the substrate(s) of the reaction
2. Identify the cofactors among the substrates using the cofactor lists
3. Create the list of acceptable atom types for the substrate(s)
4. Create the BEM for the product(s) of the reaction
5. Use the atom mapping to transmute the BEM for the products so that the rows/columns refer to the same atoms as the rows/columns in the BEM of the substrates
6. Subtract the BEM for the substrate(s) from the transmuted BEM for the products to obtain the reaction operator matrix

In order to ensure that the generated operators are as specific as possible, additional information is added to the list of acceptable atoms during step 3 indicating whether each atom is a terminal atom. BNICE was modified to disallow operators from being applied to substrates which have nonterminal atoms in place of terminal atoms. This prevents an operator from performing *de novo* reactions due to substrates containing other substrates as subgraphs.

Consequently, any reaction described by an exact operator is identical to the reaction upon which it is based (using two-dimensional representations).

This program was utilized to generate an exact operator for every reaction in MetaCyc 16.0 for which MetaCyc provides an atom mapping. Reaction direction was determined by directional information provided by MetaCyc. If no such directional information was provided, the reaction was assumed to be reversible and separate operators were generated for both reaction directions.

3.2.5 Computational Generation of Reaction Site Operators

As explained above, the reaction site operators only specify the reaction site atoms, causing them to be as general as possible. The primary difference between the exact operators and the reaction site operator is that the reaction site operators contain no spectator atoms. These rules can therefore be generated by identifying the spectator atoms and removing these atoms from the operator. Spectator atoms can be readily recognized by examining the reaction operator matrix and isolating atoms for which the rows/columns have all zeroes for entries.

Unlike the exact operators, many reactions can have identical reaction site operators. It is often useful to group these reactions together into a single operator. In order to do so we need a method of determining whether two operators are identical. For our purposes, operators can be considered identical if there is a transmutation that can be applied to the BEM, the reaction operator matrix, and the list of acceptable atom types for one operator which results in these three components being exactly the same.

One potential method for determining whether such a transmutation exists is by trying all possible atom numbering schemes for the atoms. However, this method is inefficient for large operators. Fortunately, topological symmetry perception methods can be used to determine atoms which cannot possibly be the same based on molecular structure and reduce the number of numbering schemes that must be attempted. In this work we use the method of Shelley and Monk for this purpose [65].

Adding these considerations to our previously developed procedure for the generation of exact operators allows for the following procedure to generate reaction site operators describing a set of reactions.

A. For each reaction in the set of reactions
    1. Create the BEM for the substrate(s) of the reaction.
    2. Identify the cofactors among the substrates using the cofactor lists.
    3. Create the list of acceptable atom types for the substrate(s).
    4. Create the BEM for the product(s) of the reaction.
    5. Use the atom mapping to transmute the BEM for the products so that the rows/columns refer to the same atoms as the rows/columns in the BEM of the substrates.
    6. Subtract the BEM for the substrate(s) from the transmuted BEM for the products to obtain the reaction operator matrix.
    7. Isolate all atoms which have no non-zero entries in the reaction operator matrix from the BEM, acceptable atom type list, and reaction operator matrix.

8. Check whether the current operator is identical to a previously created operator. If it is not, create the new operator.

This procedure was used to create a program in Python 2.7 for the generation of reaction site operators. This program was utilized to generate a reaction site operator for every reaction in MetaCyc 16.0 for which MetaCyc provides an atom mapping. As with the exact operators, reaction direction was determined by directional information provided by MetaCyc, and all reactions were assumed to be reversible if no information was provided.

3.2.6 BNICE Mapping

One measure of the scope of a set of operators is the set of reactions which it is able to reproduce. For this work this attribute of the operator sets was investigated using the Mapping Module of the BNICE software suite. This module works as follows. The substrates for each reaction of interest are loaded into BNICE, and the set of operators is applied to them. The products from each resulting predicted reaction are compared to those in the loaded reaction. If they match then the reaction is marked as mapped, and the operator which describes the reaction is recorded. If none of the predicted reactions map the loaded reaction, then the reaction cannot be described by any of the operators, and it is consequently marked as not mapped. In this analysis all cofactors which are grouped together in Appendix A were treated as interchangeable so that our analysis can focus on the non-cofactor substrates and products. The analysis was performed for all automatically generated operator sets and a manually generated BNICE operator set of 247 operators against the atom mapped reactions in MetaCyc, KEGG, and the reactions in the iMM9904 *S. cerevisiae* model.

3.2.7 Identify the Minimum Operator Set to Describe All of MetaCyc

The generation of the reaction site operators is heavily dependent upon the accuracy of the atom mappings. An error in the atom mapping for a reaction can result in more nonzero entries in the change matrix. This in turn will cause additional atoms to be labeled as reaction site atoms by our algorithm and thus will cause a change in the reaction site operator. For this work we investigate the extent to which the atom mappings affect the generated operators by identifying and removing any operators which only map reactions which are all also mapped by another operator and studying the results.

3.2.8 Developing a Measure of Specificity

The limiting cases of operator specificity can be useful for analysis, but for many applications an operator set of intermediate specificity is required. Ideally, the specificity of this operator set will be adjustable for different applications and will be determined using an intuitive criteria which is easily interpreted by the researcher. In this work we introduce the idea of generating such a set of operators based upon *reaction grouping*. In this method the user generates operators by grouping together sets of reactions which he/she wishes to be described by a single operator. The program then identifies *the most specific operator capable of describing all the reactions in the group.* In this work, it is assumed that all the reactions in the group contain the same reaction site. As stated above, we consider this the limit for the most general meaningful operators. This restriction therefore ensures that an operator can be generated which is capable of describing all of the reactions in the group. This method for operator generation allows the specificity of the operator set to be easily adjusted by changing the reaction groupings. Users desiring the prediction of many *de novo* reactions which may act upon substrates very different than those of known enzymatic reactions would create large reaction groupings of very dissimilar reactions. On the other hand, researchers who want to predict a small number of *de novo* reactions very similar to known enzymatic chemistry would create small reaction groupings of very similar reactions.

In order to implement this method we need to define a mathematical definition of *specificity cost* of an operator. One of the goals of this work is to produce an intuitive method for the generation of reaction operators. We therefore defined what we deemed to be an intuitive measure of specificity cost, which was simply the number of generalizations that are included in the operator. In other words, our measure of specificity cost is simply the number of additional atom types and bond types that must be included in the operator in order to describe all of the reactions in the reaction grouping. Notice, however, that this definition does not account for generalities due to atoms and bonds not being present in the operator such as, for example, when one substrate is smaller than another. For the purposes of this calculation we therefore introduce the “None” atom type and “None” bond type to calculate our specificity cost.

3.2.9 Generating Optimal Reaction Rules for a Group of Reactions

Next, an algorithm for the generation of the most specific operator capable of describing two noncyclic substrates was developed. The substrates of the two reactions were represented as networks with *both* atoms and bonds represented as nodes and bidirectional edges used to indicate the endpoints of the bonds. In these networks, each node lists the acceptable atom types and bond types for the atom or bond. Additional nodes of type “None” are added to these networks to allow for the possibility that an atom/bond in one reaction’s substrates may not have a corresponding atom/bond in the other reaction’s substrates. Once we know the optimal mapping of the substrates of one reaction onto the substrates of the other, we generate an operator by, at each point in the mapping, recording both the original node’s atom/bond type and that of the node to which it is mapped.

Unfortunately, identifying such a mapping is non-trivial. A brute-force method which explores all possible mappings would have factorial time complexity. By making two simple assumptions we can greatly increase the tractability of this problem. Our first assumption is that the reaction site for the first reaction must map to the reaction site for the second reaction. As stated above, this is necessary to guarantee that the two reactions can be described by the same operator. Our second assumption is that only atoms that are the same distance from the reaction site may be mapped to each other.

The simplest manner to apply these assumptions is to use a greedy algorithm which starts at the reaction site atoms and steps outward, finding the optimal mapping at each step. Unfortunately, this method does not account for the situation where the cost savings of a small similarity between two nodes close to the reaction site is outweighed by a larger subsequent difference farther from the reaction site.

To accurately calculate the most efficient mapping we instead start at those nodes *farthest* from the reaction site nodes and step inward. A program to accomplish this was created in Python 2.7, and its steps are outlined below.

1. Define the network describing the substrates for the first reaction as *I* and for the second reaction as *J*. Also, define a cost matrix of integers *C* and a mapping matrix of tuples *M*. Finally, let *F(x)* be the mapping of the nodes in *I* onto the nodes in *J*.
2. Identify and label the reaction site in both networks using the same method as was used for the reaction site operators.
3. Use a breadth first algorithm to calculate the distance, *L(n)*, of each node n in both networks from the closest reaction site node. Let *k* be the maximum distance from the reaction site in either substrate.
4. For each substrate network *A* identify all nodes at each distance *a* and collect them in an equidistant set of nodes $\phi(A, a)$.
5. For each equidistance set of nodes, $\phi(\mathrm{I}, \mathrm{a})\ s.t.\ 0 < a \leq k$, in *I*, add "None" type nodes for each node in the equidistance set of nodes, ϕ(J,a), at the same distance in *J*. Similarly, for each equidistance set of nodes, $\phi(J, a)\ s.t. 0 < a \leq k$ in *J*, add "None" type nodes for each node in the equidistance set of nodes at the same distance in I, ϕ(I,a) not excluding the recently added "None" type nodes. This will result in $\phi(J,a)$ and $\phi(I,a)$ containing the same number of nodes for each distance *a* in the networks.
6. For each node, *x*, create a list of descendants *D(x)* where we define that node *y* is a descendent of node *x* if *x* and *y* are connected by an edge and $L(y) > L(x)$.
7. Define the cost of descendants, $C_{desc.}(i, j)$ where i and j are nodes. This represents the minimum cost of mapping all of the descendants of i to the descendants of j.
8. $d = k$.

9. Nodes at $d = k$ have no descendants; therefore, $C_{desc.}(i,j) = 0$ for all nodes $i \in \phi(\mathrm{I}, \mathrm{d})$, $j \in \phi(\mathrm{J}, \mathrm{d})$.
10. For each pair of nodes (i, j) $i \in \phi(\mathrm{I}, \mathrm{d})$, $j \in \phi(\mathrm{J}, \mathrm{d})$ calculate the cost *C(i,j)* as follows.
    a. If the atom/bond type of atom *i* is not the same as the atom/bond type for atom *j* then $\mathrm{C}(\mathrm{i}, \mathrm{j}) = 1 + C_{desc.}(i,j)$
    b. Otherwise, $\mathrm{C}(\mathrm{i}, \mathrm{j}) = C_{desc.}(i,j)$
11. d = d -1
12. For each $i \in \phi(I, d)$ and $j \in \phi(J, d)$ solve the following optimization problem for the optimal assignment variables x_{mn}

$$\min C_{desc.}(i,j)$$

where

$$C_{desc.}(i,j) = \sum_{m \in D(i)} \sum_{n \in D(j)} C(m,n) x_{mn}$$

Subject to the constraints:

$$\sum_{m \in D(i)} x_{mn} = 1 \; for \; \forall \, n \in D(j)$$

$$\sum_{n \in D(j)} x_{mn} = 1 \; for \; \forall \, m \in D(i)$$

$$x_{mn} = \{0,1\} \; for \; \forall \, m \in D(i), n \in D(j)$$

This problem is known as the assignment problem and for this work was solved using the Hungarian algorithm [66]. For each pair of nodes (i,j) a list of tuples (m,n) of each nonzero x_{mn} is stored in M(i,j).

13. For each pair of nodes (i, j), $i \in \phi(\mathrm{I}, \mathrm{d})$, $j \in \phi(\mathrm{J}, \mathrm{d})$ calculate the cost C(i,j) as follows.
    a. If the atom/bond type of atom i is not the same as the atom/bond type for atom j then C(i,j) = 1 + $C_{desc.}(i,j)$
    b. Otherwise, C(i,j) = $C_{desc.}(i,j)$
14. Repeat steps 11-13 until d = 0.
15. Once $d = 0$ we have reached the reaction site. The reaction site nodes are required to be identical between the reactions by our assumptions. However, due to symmetry in the reaction site, multiple mappings may still be possible. We therefore repeat step 12 with the additional requirements that atom/bond types and bonds broken are identical between each node and the node to which it is mapped. Rather than adding optimizing assignment variables x_{mn} to *M* as tuples we instead add these tuples to array Q.
16. Finally, the mapping matrix M is then used to recover the full mapping *F(x)* as follows.
    a. Let $(\alpha, \beta) = Q[0]$.
    b. $F(\alpha) = \beta$.
    c. If either of the atom/bond types for α or β is not "None," add *M(α, β)* to Q
    d. Delete Q[0]
    e. Repeat steps a-d until Q is empty

This program is what is known as a dynamic programming algorithm. It begins at the outermost nodes and works its way inward, and at each step, the algorithm investigates all mappings for each node in network *I* onto each node in network *J* at the current distance. For each of these possible mappings the optimal mapping of the descendants of both nodes is known from the results calculated during the previous step. This is the primary advantage of the algorithm. By utilizing the optimal mapping calculated during the previous step we forgo the need to investigate potential mappings for any nodes other than those at the current distance. This significantly reduces the difficulty of the problem.

Unfortunately, the algorithm as described above cannot guarantee the lowest cost mapping for substrates containing cycles. Some nodes in such compounds are the descendants of multiple nodes, requiring the additional restriction that nodes can only be mapped to each other if all of their ancestors, nodes from which they are descendants, are also mapped to each other. Unfortunately, this restriction can result in operators with very few spectator atoms when comparing linear and cyclic substrates since no node with multiple ancestors can ever map to a node with just one. To overcome this obstacle, we create multiple networks to represent the same reaction substrates. Each of these networks is created using a process of cutting edges. For each node in the network with multiple ancestors, edges are cut to create a network for each possible combination of ancestors. Operators are generated for all combinations of these networks, and the lowest cost operator is accepted.

This algorithm was used to approximate the lowest cost operator for a grouping of reactions by combining the reactions in a reaction grouping in a pairwise fashion. The first two reactions in a group are combined into an operator. This operator is then combined with a third reaction to create a new operator and so forth until an operator is generated which describes all the reactions in the grouping.

3.3 Results

3.3.1 Computational Generation of Limiting Chemical Operators and Mapping against MetaCyc

A set of exact chemical operators and a set of reaction site chemical operators were computationally generated from the MetaCyc database. These sets consisted of 8,682 and 2,413 operators, respectively, and both were capable of mapping every reaction in MetaCyc. For the exact operators, each unique reaction was described by only one operator as should be expected since these operators represent the most specific limit. However, how the reactions in MetaCyc were described by the reaction site operators was much more complex (Figure 3.1 and Table 3.1).

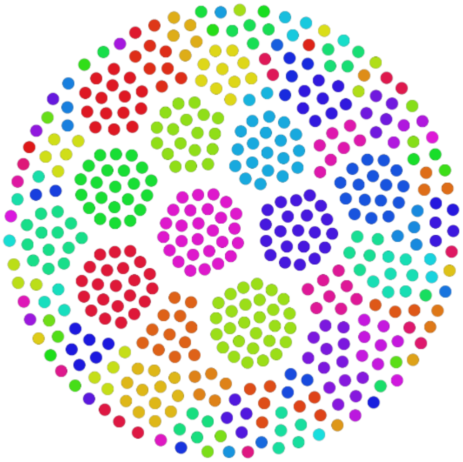


Figure 3.1: MetaCyc reactions as described by the reaction site operators. Visualization of the MetaCyc reactions as described by the reaction site operators. Each dot represents ten reactions in MetaCyc, and each color represents a reaction site rule. Only those rules which describe at least ten reactions are shown.

Table 3.1: Description of MetaCyc reactions by reaction site operators.

Number of Reaction Site Operators	Percentage of Reactions in MetaCyc Described (%)
5	14.1
11	25.3
26	37.9
56	50.0
152	62.5
424	75.0
1237	87.5
2413	100.0

As can be seen in Figure 3.1, the number of reactions described by different operators varies greatly. There are a few operators which describe a large number of reactions, and many operators which describe just a few. This results in the number of MetaCyc reactions described by these operators showing quickly diminishing returns (Table 3.1). Over a quarter of the reactions in the database can be described by just 11 reaction site operators. However, an additional 45 operators are required to cover the next quarter, an additional 368 to cover the next quarter, and an additional 1,989 to cover the final quarter of the reactions.

This operator generation process is highly dependent upon the atom mappings used to describe the reactions. To investigate this dependence we studied the cases where a single reaction is described by multiple reaction site operators. After removing those operators which only described reactions also described by another operator, we were left with an operator set consisting of 1,700 operators. Additionally, 203 reactions were identified for which the substrates and products are identical from the perspective of our program. These mainly consisted of transport reactions and reactions where only the stereochemistry of the substrate was changed. Each of these reactions was described by a single operator. Removing these reactions and operators from our reduced operator set results in a set of 1,497 operators capable of describing every reaction in MetaCyc.

These results represent the best-case scenario for the minimum number of operators required to describe the MetaCyc database and illustrate the ability of our operator generation program to highlight reactions whose atom mappings may warrant further investigation. However, as of the writing of this chapter, the methods used by MetaCyc represent the state of the art in generating atom mappings, and we do not propose an improved method here. We therefore will use the full set of 2,210 reaction site operators (the original 2,413 operators less those that only perform transport or stereochemical reactions) for the remainder of the work in this chapter.

3.3.2 Mapping Automatically Generated Operators against KEGG and iMM904

The 8,682 exact operators and 2,210 reaction operators were mapped against the KEGG database of enzymatic chemistry. Additionally, the full set of 242 manually generated operators previously used by BNICE was also mapped against KEGG for comparison. The results are shown in Figure 3.2.

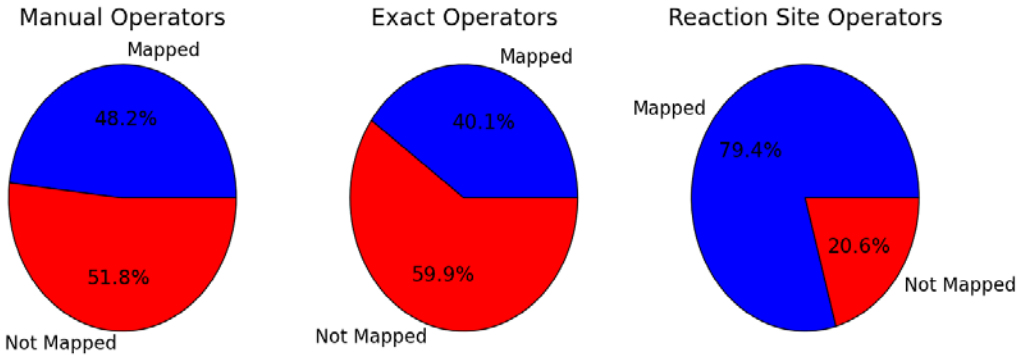


Figure 3.2: Mapping results for KEGG database.

Mapping results are shown for manually generated operators, and the automatically generated exact and reaction site operators generated based upon the MetaCyc database.

Notice that 40.1% of the reactions in KEGG are identical to reactions in MetaCyc. This is a little lower than the 3,895 common reactions out of 8,692 (44.8%) found by Altman et al. [67]. However, in their work, Altman et al. only required that the reactants and products of two reactions be highly similar (cosine similarity of the stoichiometry vector > 0.6) rather than identical (after cofactor generalization) as is the case in this work. Additionally, Altman et al. utilized additional information in their determination of identical reactions such as Uniprot ascension number, enzyme classification, and the name of the catalyzing enzymes. Our analysis solely focuses on the stoichiometry of the reactions.

The automatically generated reaction site operators were able to describe 79.4% of the reactions in KEGG, which is significantly more than the 48.2% described by the manually generated operators. This is despite the fact that the manually generated operators were largely created based on KEGG reactions while the reaction site operators were based on MetaCyc [6].

Next, these three operator sets were mapped against the iMM904 model for yeast. The results are shown in Figure 3.3.

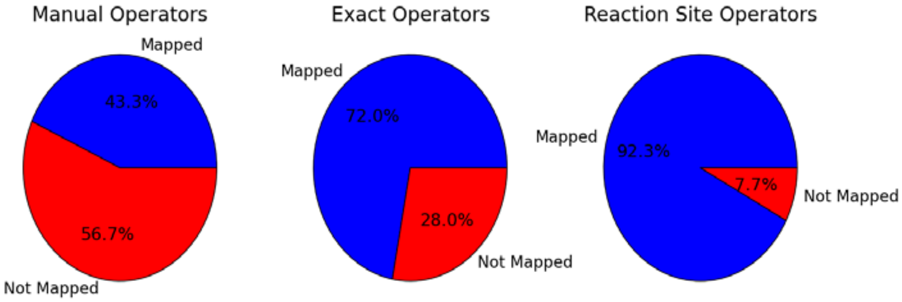


Figure 3.3: Mapping Results for iMM904 metabolic model of yeast.

Mapping results are shown for manually generated operators, and the automatically generated exact and reaction site operators generated based upon the MetaCyc database.

Both automatically generated set of operators showed improved performance when mapped against iMM904 compared to KEGG. This is likely due to KEGG containing more reactions with unusual chemistry than the yeast model. This is highlighted by the fact that 72% of the reactions were mapped by the exact operators indicating a large majority of these reactions have identical stoichiometry to reactions found in MetaCyc. The manual operators, however, performed slightly worse against iMM904 compared to KEGG, likely due to many of the manual operators being based upon KEGG reactions. In fact, the exact operators were able to map significantly more reactions than the manual operators, indicating that the generality of the manual operators was unable to compensate for their lack of breadth. Simply using known reactions from MetaCyc would obtain better coverage of the reactions in this yeast metabolic model than the manual operators. The reaction site operators were able to describe nearly all, 92.3%, of the reactions in the metabolic model.

3.3.3 Intermediate Operator Generation

The groupings provided by the reaction site operators were used to make a set of intermediate operators using the reaction grouping method. Each of these operators is designed to more specifically describe the identical set of MetaCyc reactions as one of the reaction site operators. Therefore, mapping them against MetaCyc results in 100% of the reactions being mapped and the same breakdown of the number of operators required to map a given percentage of MetaCyc as shown in Table 3.1.

The intermediate operators are designed to approximate the most specific description capable of describing all of the reactions in a grouping. These operators can thus be thought of as allowing for the prediction of novel reactions by allowing the operator to be applied to any substrate whose chemical structure is between the substrates of the reactions it describes. We therefore would expect that as the number of reactions in the grouping increases the specificity should decrease since the resulting operator will need to be able to accommodate a wider variety of substrates. This is illustrated by Figure 3.4.

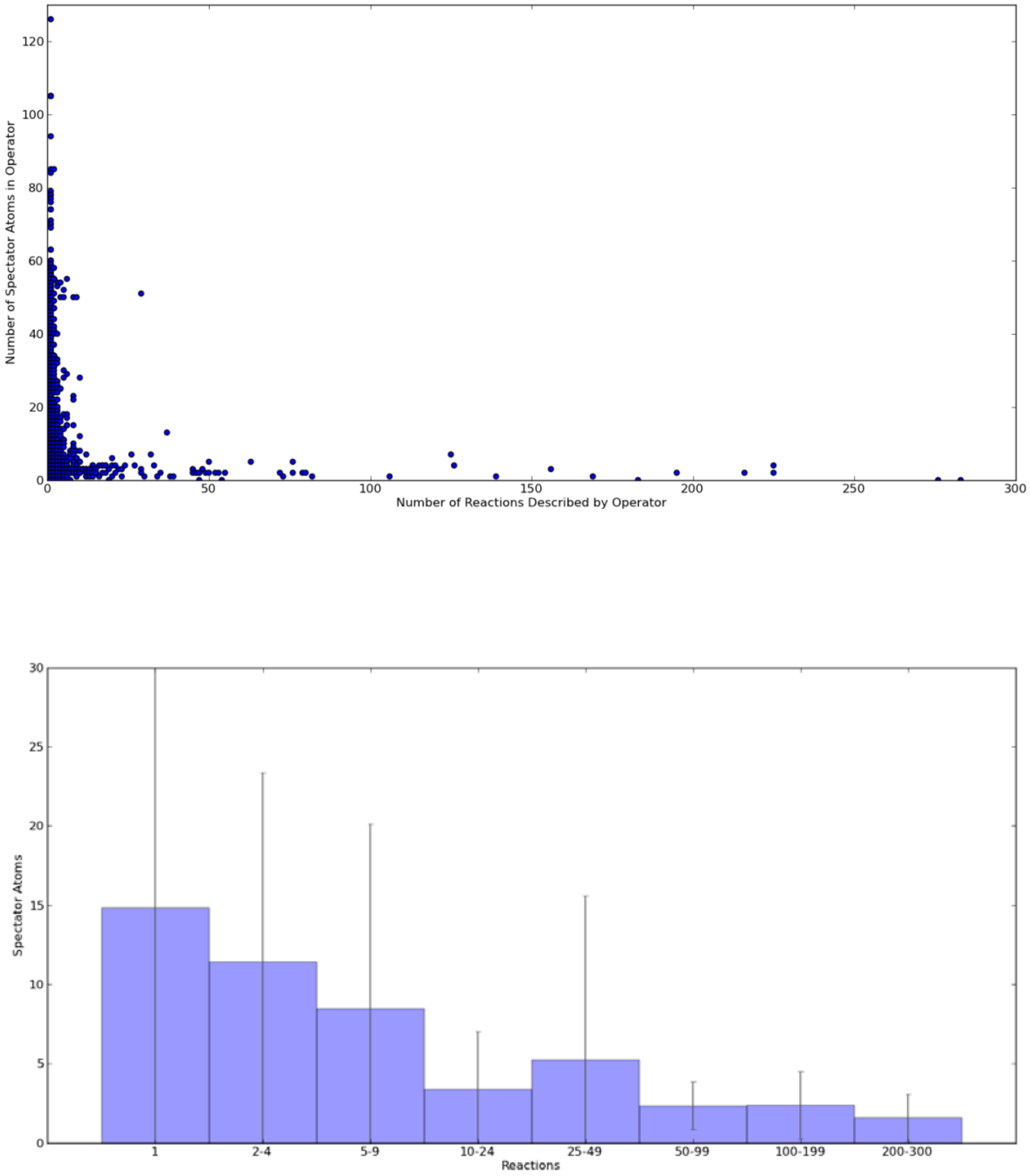


Figure 3.4: Effect of reaction grouping size on operator specificity.

Scatter and bar plot of the number of spectator atoms included in intermediate operators compared to the number of reactions in the grouping used to generate them.

As expected, the number of spectator atoms, and therefore the specificity of the operator, is lower for operators which contain more reactions. The drop-off in spectator atoms occurs relatively sharply with few operators which describe more than ten reactions containing ten or more spectator atoms. There is a large degree of variance in the number of spectator atoms between operators describing the same number of reactions. For operator groupings which contain only a few reactions, this is largely a result of substrate size. However, for operators which describe many reactions this is due to a conservation of spectator atoms across all reactions described by the operator. This occurs when a particular set of spectator atoms must all be present for a reaction to occur. Largely, this is due to distinct functional groups, the most egregious of which is Coenzyme A which contains 48 spectator atoms and is required for the performance of certain types of chemistry. Removing those reactions which contain Coenzyme A from our analysis has a noticeable effect on our results as shown in Figure 3.5.

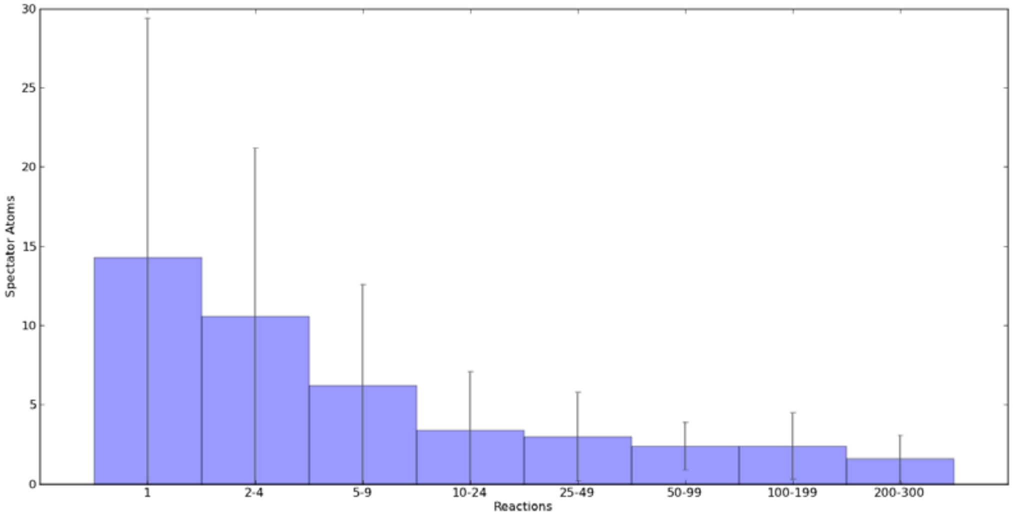


Figure 3.5: Effect of reaction grouping size on operator specificity after removing coenzyme A.

Bar plot of the number of spectator atoms included in intermediate operators compared to the number of reactions in the group used to generate them. In this plot all reactions which contain Coenzyme A were excluded.

BNICE was used to apply these operators and the reaction site operators to the 22 common amino acids. This resulted in 481,764 predicted reactions for the reaction site operators and 17,682 predicted reactions for the intermediate reaction site operators. The addition of spectator atoms therefore resulted in a 50-fold decrease in the number of reactions predicted. This decrease occurs due to two factors: a reduction in the number of operators applied to the substrates and a reduction in the number of reactions predicted by each applied operator. We found that most of the reduction was due to the former.

The number of unique operators applied to the substrates compared to the number of MetaCyc reactions used to create the operators is shown for the reaction site operators and the reaction site intermediate operators in Figures 3.6 and 3.7, respectively.

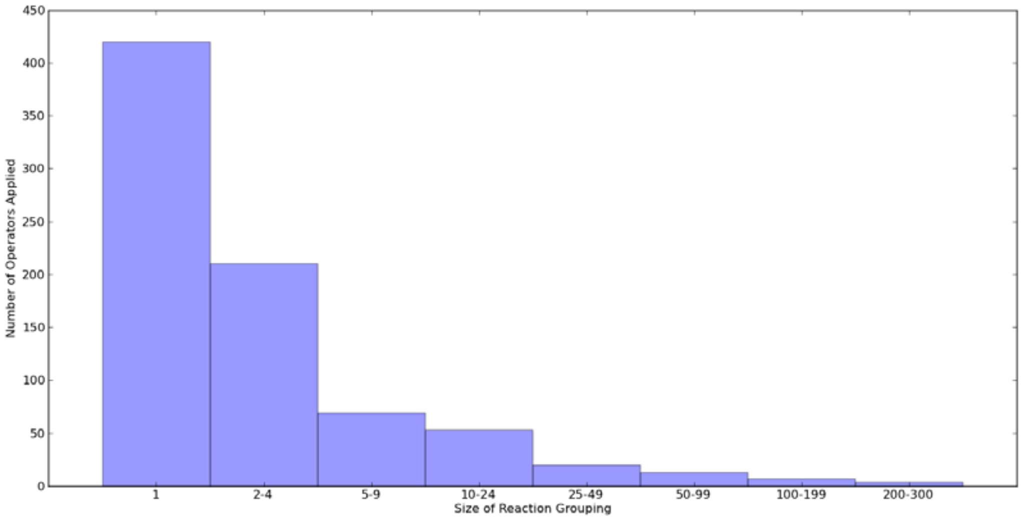


Figure 3.6: Effect of reaction grouping size on utilization of reaction site operators. Number of operators successfully applied to the 22 common amino acids compared to reaction grouping size for the reaction site operators applied to the 22 common amino acids.

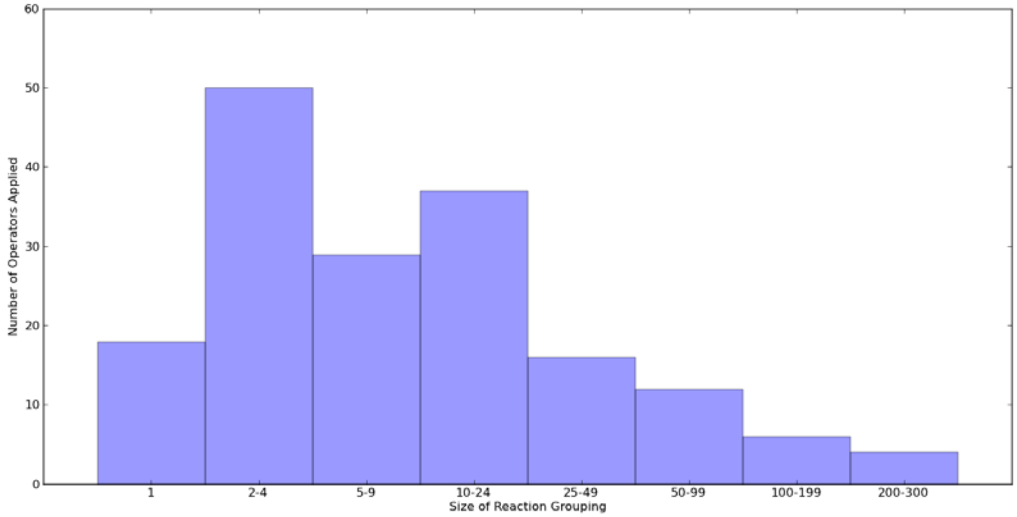


Figure 3.7: Effect of reaction grouping size on utilization of intermediate reaction site operators.

Number of operators successfully applied to the 22 common amino acids compared to reaction grouping size for the intermediate reaction site operators applied to the 22 common amino acids.

As can be seen from Figures 3.6 and 3.7, the addition of spectator atoms reduces the number of applied operators by preventing the application of operators created from small groupings of reactions. In the reaction site operators, the number of applied reactions is inversely related to the size of the reaction grouping used to create the operator. This is largely a result of the number of operators present. As indicated by Table 3.1, there are far more operators describing small groupings of reactions than there are describing large groupings. For the reaction site intermediate operators, the increased specificity of the operators for the small groupings of reactions more than compensates for their increased number and results in a maximum in operator application to occur in the intermediate range of operator size. This seems much more in line with what would be expected. Many of these groupings are small precisely because the constituent reactions are uncommon, requiring particular spectator atoms to be present to stabilize the reaction. The reactions predicted by the intermediate operators are therefore likely higher quality predictions since, unlike the reaction site operators, a large percentage of the predicted reactions are a result of fairly common types of enzymatic chemistry.

3.4 Discussion

3.4.1 Computational Generation of Limiting Chemical Operators

We have described techniques for the computational generation of the most specific (exact) and least specific (reaction site) chemical operators possible for a given set of reactions. Applying these techniques to MetaCyc generated operator sets containing 8,682 and 2,413 operators, respectively, capable of describing every reaction with atom mappings in the database. The exact operator set simply consisted of one operator for each unique reaction in MetaCyc. Meanwhile, the reaction site operator set showed sharply diminishing returns where 11 operators were capable of describing over 25% of the reactions in Metacyc, but 424 operators were required to describe 75% of reactions and 1,497 to describe 100%. This helps explain both the initial success and ultimate difficulty with the development and utilization of manually generated operators. A small number of manually created operators describing the most common types of enzymatic chemistry can describe a relatively large percentage of known biochemical reactions.

However, each further attempt to expand upon the amount of biochemistry described requires an increasing number of operators until the process can become largely untenable manually.

These operator sets generated from MetaCyc were then mapped against the KEGG database and the iMM904 metabolic model. The exact operators were capable of describing 40.1% and 72.0% of the balanced reactions from these sources, respectively, while the reaction site operators were capable of describing 79.4% and 92.3%, outperforming the manually generated operators in both cases. This illustrates an advantage of generalized chemical operators over exact descriptions of reactions. The generalized operators are capable of describing a large majority of the chemistry in these sources despite having been trained on a different dataset. In order to expand our operators to describe all the reactions in these new sources, we need to only focus on the small subsets of reactions (20.6% and 7.7% for KEGG and iMM904, respectively) which are not described by our operators rather than having to analyze every reaction in the entire dataset. This provides the potential for our chemical operators to rapidly grow to describe all the reactions found in many different sources of enzymatic chemistry.

3.4.2 Generation of Intermediate Chemical Operators

We next developed an algorithm for the generation of sets of operators of intermediate specificity. This method relies upon the user creating groupings of similar reactions. The algorithm then approximates the most specific chemical operator capable of describing all of the reactions in the grouping. We initially developed this algorithm using the groupings defined by the reaction site operators. In analyzing the resulting operators we found that as the number of reactions in the reaction grouping increases the number of spectator atoms (and therefore the specificity) of the resulting operator decreases. We have therefore created an operator generation method which allows for the intuitive adjustment of operator specificity through the adjustment of reaction groupings. This gives the user the ability to automatically create operators which are tuned to a particular application.

As an initial demonstration of this technique, we generated intermediate operators for the groupings defined by the reaction site operators. Each of the resulting operators was capable of describing the identical set of MetaCyc reactions as the reaction site operator upon which it was based. However, this operator set saw a 50-fold decrease in the number of predicted reactions when applied to the common amino acids compared to the original reaction site operators. One of the largest problems with *de novo* biochemical reaction prediction is the large number of reactions predicted. This often makes the analysis of the predicted reactions difficult, necessitating the application of various filters to reduce the reaction set. Our technique therefore allows for the generation of a much smaller number of predicted reactions whose substrates are more similar to known enzymatic reactions and therefore may be more likely to occur.

It should further be noted that the reaction site operators represent the most general operators possible, and thus the groupings used to create the intermediate operators in this work represent the most general groupings possible. Further reductions in the number of predicted reactions can be achieved without reducing mapping efficiency by dividing these reaction groupings into smaller subgroups using criteria such as EC classification [27], organism of origin, or similarity in the genes encoding for the associated enzyme. Such a subgroup reduction would have the additional advantage of providing characteristics which could be utilized to identify candidate enzymes capable of catalyzing predicted reactions of interest.

We have demonstrated an automatic and intuitive method for generating chemical operators. Utilization of this method allows for rapid inclusion of new reaction information into operator sets, circumventing the primary shortcoming of manually generating these operators. Furthermore, the flexibility of this method allows the operators to be tailored for specific applications. This method used in conjunction with an operator-based method for enzymatic reaction prediction such as BNICE has the potential to rapidly increase the rate at which promising enzymatic chemistry is discovered and implemented.

Chapter 4: Application of Automatically Generated, Genetically Grouped Chemical Operators to iJO1366

4.1 Introduction

The prediction of *de novo* enzymatic reactions is useful in a wide variety of fields including bioremediation, industrial biotechnology, and biomimetic catalysis [24, 26, 57]. We have previously developed a program known as the Biological Network Integrated Computational Explorer (BNICE) which uses generalized reaction rules known as *chemical operators* to predict *de novo* enzymatic chemistry [6]. This method has been successful at predicting novel chemistry that has subsequently been experimentally confirmed as shown in Chapter 2 of this thesis. However, the BNICE program often identifies a very large number of potential reactions, thereby increasing the difficulty of subsequent analysis of the predictions. Furthermore, as in Chapter 2 of this thesis, the predicted reactions must be associated with candidate enzymes before experimental confirmation can be attempted. Previously, both of these problems were mitigated through the application of chemical fingerprinting methods to filter BNICE predictions and to associate the predicted reactions with promising enzyme candidates [68]. However, ideally we would prefer a flexible framework capable of being rapidly adapted to utilize any of a variety of criteria to both reduce the scope of BNICE predictions to only the most promising reactions and to assist in the association of those reactions with candidate enzymes.

One promising avenue for the development of such a framework is the chemical operators themselves. We have previously developed a method for the automatic generation of chemical operators from reaction groupings (Chapter 3). To use this program, users create groupings of reactions which they wish to all be described by a single chemical operator. Our algorithm then approximates the most specific operator capable of describing all the reactions in the grouping. With this new capability, we can use the operators as the vehicle by which we improve the quality of the predicted reactions by grouping together reactions with similar properties such as the same enzyme classification, similar encoding genes, occurring in closely related organisms, etc. This allows the operators to be more specific by insuring that they are only generalized across related reactions. Additionally, it can highlight probable properties of candidate enzymes to assist with the experimental validation.

In this work, we demonstrate the power of this technique by grouping reactions by the similarity of the genes associated with the catalyzing enzymes. To do this, we generate reaction groupings for reactions in the MetaCyc database of enzymatic chemistry [7] using Conserved Domain Database genetic superfamily classifications [10]. The operators produced by these reaction groupings are then applied to all chemical compounds in the iJO1366 metabolic model of *Escherichia coli* [69]. Finally, the products of all the resulting reactions predicted by BNICE are compared to the compounds in the DrugBank database of pharmaceutical chemicals [70]. Reactions for the production of 206 of these compounds were predicted by BNICE.

4.2 Methods

4.2.1 BNICE

BNICE is explained in full elsewhere [6]. Briefly, this program uses generalized descriptions of enzymatic reactions known as *chemical operators* to predict potential biochemical reactions. The generalization in the chemical operators can allow for the prediction of *de novo* reactions by allowing for reactions to occur which are similar to but not identical to known enzymatic reactions. The degree of generalization affects the number and quality of these predicted reactions. More generalized operators allow for the prediction of a larger number of potential reactions, but many of these reactions are less similar to known enzymatic reactions, often causing them to be difficult to practically implement.

BNICE can also be utilized in a slightly different capacity known as *mapping*. In this method, BNICE is used to predict potential reactions from every compound in a database. These predicted reactions are then compared to all the reactions in the database. Any reaction that has been reproduced is marked as *mapped* while those that are not reproduced are *not mapped*. This technique can be used to measure the amount of chemistry in a database which is capable of being described by a particular set of chemical operators.

4.2.2 Reaction Site Operators

We have previously described methods for the generation of *reaction site operators* (Chapter 3). These operators are generalized representations of biochemical reactions which only contain those atoms whose bonds change across the reaction. This set of operators is the most generalized method of representing a set of reactions. For this work we used the previously generated set of 2,210 operators capable of describing every reaction for which atom mapping exists in the MetaCyc database.

4.2.3 Generation of Intermediate Operators Based upon CDD Groupings

A method for the generation of operators based upon reaction grouping has been developed by our group (Chapter 3). Briefly, the method works by approximating the most specific chemical operator capable of mapping every reaction in a user-defined group. This method affords the user intuitive control over chemical operator generality. By using large groups of dissimilar reactions, the user can create very general chemical operators by forcing the operators to use flexible enough reaction descriptions to be able to describe many disparate reactions. In contrast, specific operators can be created by using small groups of similar reactions.

In this work, we develop a set of intermediate chemical operators by grouping MetaCyc reactions by CDD superfamily. Each of these groupings consists of all reactions which are associated with a gene in the same superfamily *and* which are mapped by the same reaction site operator. This ensures that all the reactions in a group are catalyzed by similar genes and perform similar chemistry. Only reaction groupings containing at least two reactions were utilized.

4.3 Results

4.3.1 Generation of CDD Operators and Mapping against MetaCyc

Utilizing reaction groupings based on CDD superfamilies and reaction site operators, 167 operators were generated from the MetaCyc database. In order to gauge the extent of chemistry described by these operators, three operator sets were mapped against the MetaCyc database: the reaction site operators, the subset of the reaction site operators which can be associated with at least two genes in the same CDD superfamily, and the final set of operators generated from our groupings. The results of these mappings are shown in Figure 4.1.

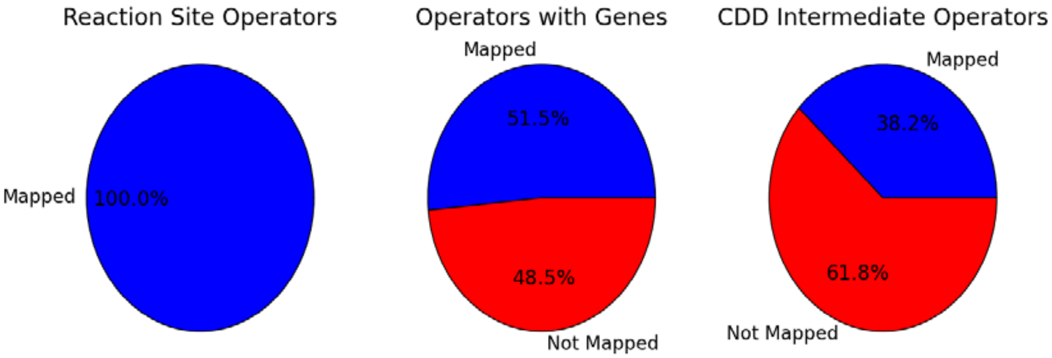


Figure 4.1: Mapping results for MetaCyc database.

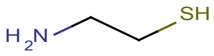
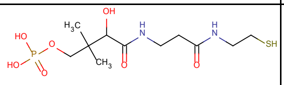
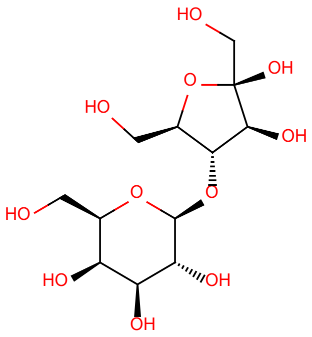
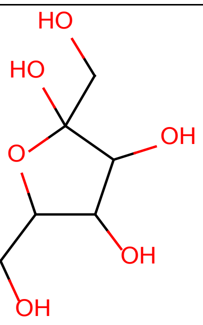
Mapping results are shown for the reaction site operators, those reaction site operators which can be associated with a CDD superfamily and the intermediate operators generated based upon CDD superfamily groupings. All operators were generated from the reactions in the MetaCyc database.

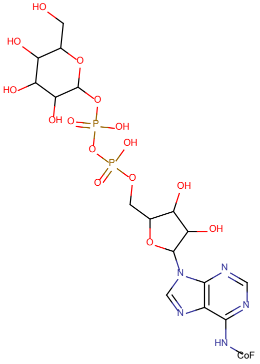
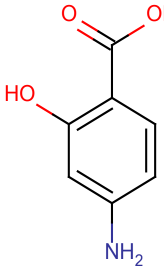
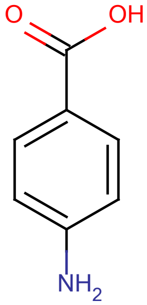
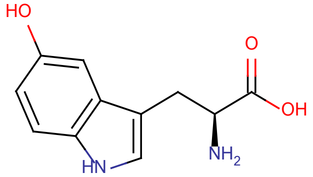
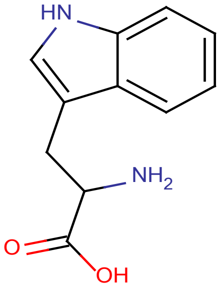
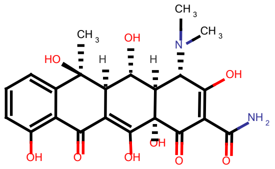
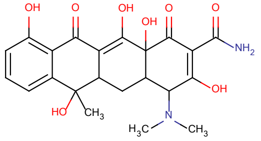
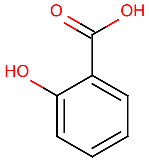
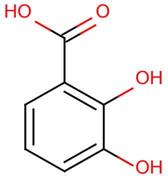
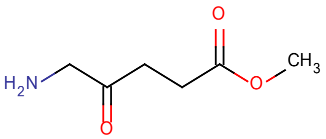
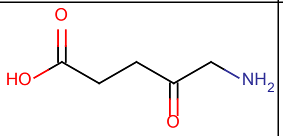
The full set of reaction site operators is capable of describing every reaction in MetaCyc by design. However, many reactions in MetaCyc are not associated with a gene or are associated with a gene which has not been categorized into a CDD superfamily. Thus, when we limit ourselves to those reaction site operators which contain at least two reactions associated with the same CDD superfamily, we remove many of our operators, thereby limiting the percentage of reactions in MetaCyc we can describe to 51.5%. Creating intermediate operators from our reaction groupings makes our operators more specific, further limiting the chemistry we can describe to 38.2% of the reactions in the database. This limits our predictions to only include chemistry for which we have genetic evidence to support the generalization of an operator.

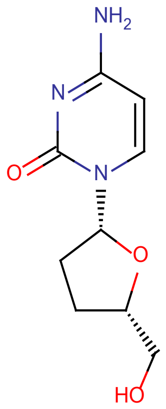
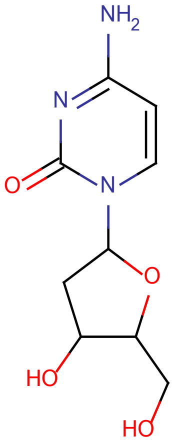
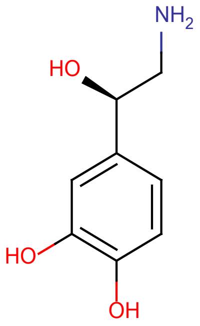
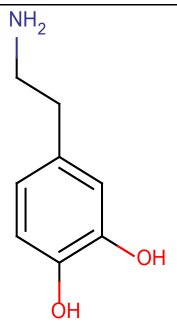
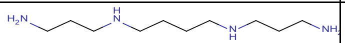
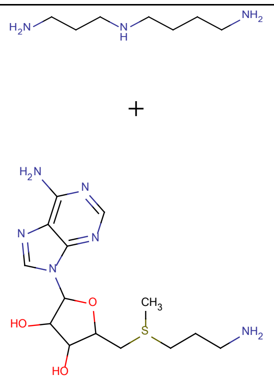
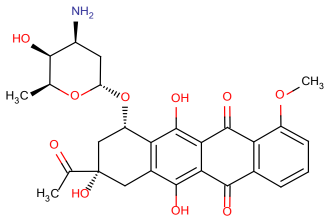
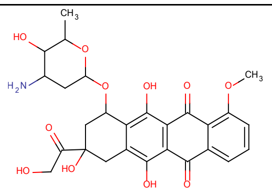
4.3.2 Identifying Potential Reactions for the Production of Pharmaceuticals

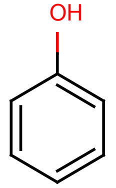
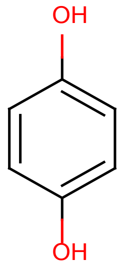
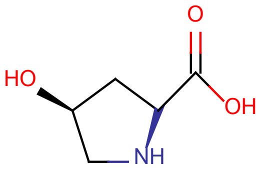
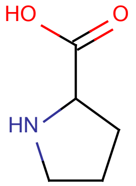
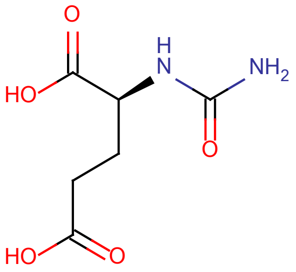
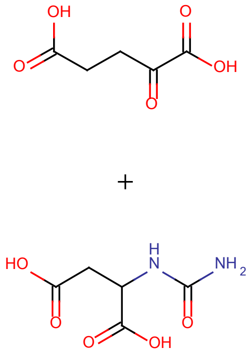
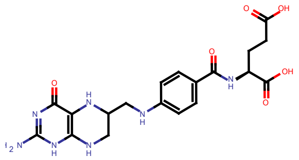
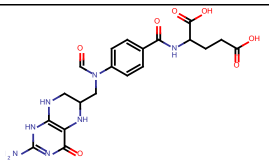
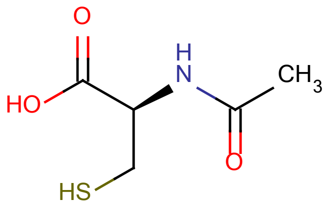
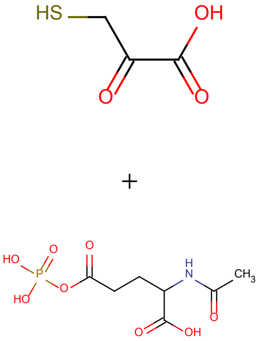
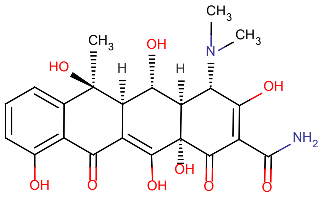
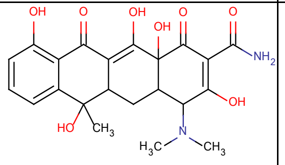
BNICE was used with the operators from our CDD superfamily grouping to predict reactions from every compound found in the iJO1366 model of *E. coli* [69]. This resulted in the prediction of 1,488,070 reactions for the production of 688,787 compounds. These compounds were then compared against those in the DrugBank database of pharmaceutical compounds [71]. We found that 206 of the generated compounds were contained in the database, of which 20 were FDA approved drugs. Table 4.1 shows the production information for these 20 compounds.

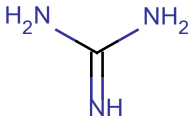
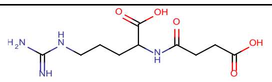
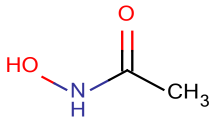
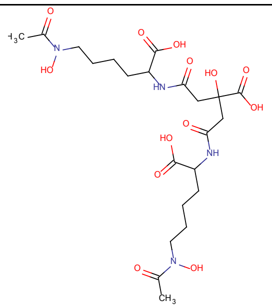
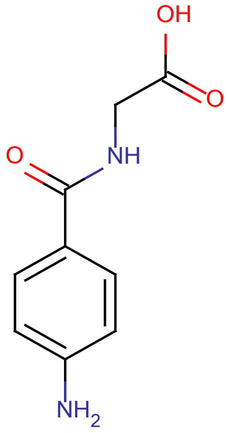
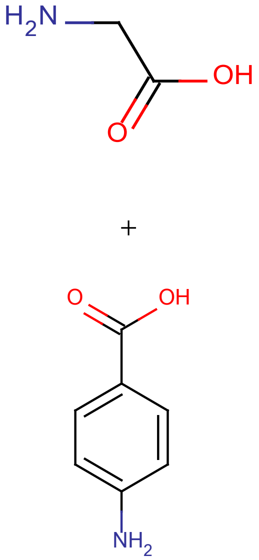
Table 4.1: Generated compounds that are FDA approved pharmaceuticals

Drug Bank ID	Compound Structure	Common Name	Produced From	Super-Family
DB00847		Cysteamine		cl00220
DB00581		Lactulose	 +	cl10013 cl19139

				
DB00 233		Aminosalicylic Acid		cl12078 cl18408
DB02 959		Oxitriptan		cl12078 cl18408 cl18680
DB00 595		Oxytetracycline		cl12078
DB00 936		Salicylic acid		cl12078
DB00 992		Methyl aminolevulinate		cl17173

DB00943		Zalcitabine		cl12078
DB00368		Norepinephrine		cl12078 cl18408
DB00127		Spermine		cl17173
DB00694		Daunorubicin		cl12078

DB03255		Phenol		cl12078
DB08847		Hydroxyproline		cl12078 cl18408
DB06775		Carglumic Acid		cl18945
DB00116		Tetrahydrofolic acid		cl19134
DB06151		Acetylcysteine		cl18945
DB00595		Oxytetracycline		cl12078 cl19134

DB00536		Guanidine		cl19134
DB00551		Acetohydroxamic Acid		cl19134
DB00345		Aminohippurate		cl17255

The reactions which produce these 20 compounds are all predicted to be catalyzed by enzymes belonging to one of ten CDD superfamilies. The most prevalent types of reaction for the production of these compounds are hydroxylation and dehydroxylation with the most common CDD superfamily being cl12078. There are 19 reactions included in both the cl12078 grouping responsible for hydroxylation and in the grouping responsible for dehydroxylation. As shown in our previous work, the specificity of operators generated by our method is inversely proportional to the number of reactions in the reaction grouping (Chapter 3). The result of these large reaction grouping is very general operators; these operators only require that a three carbon chain is present at the site of hydroxylation and that a chain of three carbons is attached to a hydroxyl at the site of dehydroxylation. While this is the largest superfamily grouping to catalyze hydroxylation reactions, it is not the only one; cl19134, cl18408 and cl18680 also catalyze hydroxylation. However, these reaction groupings have four, eleven and nine reactions in them, respectively, and are more specific than the cl12078 grouping.

The formation of guanidine and oxytetracycline serve to demonstrate the situation where the same CDD superfamily, cl19134, performs multiple chemistries. The superfamily classification is a designation of genetic similarity, not of reaction chemistry, and consequently, enzymes with the same CDD superfamily may perform different types of chemistry. In this case the cl19134 superfamily actually performs five types of chemistry: reduction of double bonds, hydroxylation, ether formation, and breaking/forming carbon-nitrogen bonds. Only hydroxylation and breaking carbon-nitrogen bonds produced compounds found amongst the FDA-approved pharmaceutical compounds in the DrugBank database.

4.4 Discussion

4.4.1 Operator Generation from CDD Superfamily Groupings

Using the automatic operator generation program, we generated 167 operators from the MetaCyc database based upon CDD superfamily genetic groupings of enzymes. The reliance upon only those reactions associated with genes contained in the CDD database limits the range of chemistry to 51.5% of the reactions in MetaCyc. Using the automatic operator generation program to add spectator atoms further limits this to 38.2% of the reactions in MetaCyc. However, these operators were created using 610 reactions (7.0% of the MetaCyc database). This indicates that 31.2 % of the reactions in MetaCyc are described by a CDD superfamily operator even though they are not related to a CDD superfamily. In some cases, genes may be associated with these reactions which are merely not listed in the database. However, in the cases where no gene is known to encode the enzyme catalyzing a reaction, these results may be useful in identifying such genes by suggesting probable CDD superfamilies to which they may belong. Although this is not a primary focus of this work, this illustrates an additional advantage of grouping reactions by enzyme/gene properties: the potential for a positive feedback loop where the resulting operators can guide the identification of reactions with similar properties, further improving the quality of the generated operators.

4.4.2 Application of CDD Superfamily Operators to iJO1366

The operators created from CDD superfamily groupings were then applied to all compounds in iJO1366 resulting in 1,488,070 predicted reactions, producing 688,787 unique products. Previous work applying the reaction site operators generated 481,764 reactions when just applied to the 22 common amino acids (Chapter 3). The filtering of our BNICE predictions by grouping operators by CDD superfamily therefore has allowed us to explore a manageable number of predicted reactions and compounds which would likely not be the case with the previous set of more general operators.

The predicted products were then compared to the DrugBank database of 6,837 pharmaceutical compounds with 206 compounds found to match. A listing of the details for these matched compounds which are FDA approved drugs is shown in Table 4.1, while a full listing of all matched compounds is shown in Appendix B. These tables not only list the compounds of interest but help to associate them with candidate enzymes through the listed CDD superfamily and thus can serve as a useful guide for experimentalists looking to implement pathways to any of the listed compound in *E. coli* by providing a starting point for the identification of candidate enzymes.

4.4.3 Utility of Automatic Operator Generation

While this chapter has focused on using similarity between encoding genes to group reactions, the process does not require the use of this particular attribute. Automatic operator generation based upon reaction groupings allows users to filter BNICE results and guide candidate enzyme identification using any criteria that can be used to group together enzymatic reactions. The current work is merely one example of the flexibility and utility of this process.

Chapter 5: Conclusions and Recommendations for Future Work

5.1 Conclusions

Overall, this thesis has demonstrated the utility of using a method based on generalized chemical operators such as BNICE to predict *de novo* enzymatic reactions (Chapter 2), the development of a method for the automatic generation of chemical operators from user-defined reaction groupings (Chapter 3), and the utilization of this procedure for automatic operator generation to limit BNICE predictions to the most promising reactions and to guide the association of those reaction with candidate enzymes (Chapter 4).

In Chapter 2, BNICE was used to predict novel pathways for the production of propionic acid. Beyond the usefulness of new biochemical routes for the production of this commonly used preservative and chemical precursor, this serves as an excellent example of the ability of BNICE to discover new and promising methods for the production of a compound. Moreover, it illustrates how BNICE results can be used to guide experimental research. By examining all four-step and five-step pathways from pyruvate to propionic acid, we were able to determine that the reduction of acrylic acid to propionic acid was a key reaction required for the majority of the potential pathways. This allowed us in turn to focus our experimental work on confirming this important reaction. Note that this type of analysis can only be performed due to the large amount of predicted pathways that result from examining many *de novo* reactions. If we had identified potential pathways manually or only explored the known pathways between these compounds, this form of analysis would likely have been difficult due to the smaller number of reactions in the network of pathways. Thus, even those pathways predicted by BNICE which are non-optimal can be useful for guiding experimental research. Finally, we were able to experimentally confirm the reduction of acrylic acid to propionic acid by NADPH dehydrogenase (Oye2p), a reaction this enzyme was not previously known to perform. This is the first time that BNICE has guided the experimental verification of a predicted enzymatic reaction.

In Chapter 3, we improved upon BNICE by exploring the automatic generation of BNICE operators. We began by examining the limiting cases of specificity for chemical operators by generating two sets of operators: one the most specific possible (exact operators) and the other the most general (reaction site operators) for all the reactions in MetaCyc with atom mappings. We then developed a method for the generation of operators of intermediate specificity based upon reaction groupings. In this method, the user separates reactions of interest into groups of similar reactions. For each group, our algorithm generates the most specific operator capable of describing all the reactions in that group. As a demonstration, we used this method to generate intermediate operators using the reaction site operators as our reaction groupings. Each of the resulting operators was able to describe the identical set of MetaCyc reactions as its corresponding reaction site operator. However, when applying these intermediate operators to the 22 common amino acids, we saw a 50-fold decrease in the number of reactions predicted compared to the reaction site operators, indicating a sizable increase in the quality of the predicted reactions. We also found that, as expected, the specificity of the generated operators varied inversely with the size of the reaction groupings. Small groupings resulted in more specific operators, while large groupings tended to result in more general operators. The promise in this technique lies in its flexibility. Different operator sets may be generated for different applications by varying the way the reactions are grouped. For example, a researcher in industrial biotechnology may wish to use reaction groupings based upon Enzyme Commission

(EC) classification [27] or genetic similarity in order to ensure smaller groups of more similar reactions, resulting in more specific operators whose predictions are close to known enzymatic chemistry. However, a researcher in biomimetic catalysis may wish to use more general operators based upon larger reaction groupings such as those formed by the reaction site operators. We believe this flexibility and the intuitive manner in which the operator specificity can be adjusted makes this algorithm a powerful tool for enzymatic research.

In Chapter 4, we used this automatic operator generation program to produce operators based upon MetaCyc reactions which were grouped by the similarity of the associated genes as defined by Conserved Domain Database (CDD) superfamilies. These operators were then applied to the iJO1366 model of *E. coli* resulting in the prediction of 1,488,070 reactions which produce 688,787 unique compounds. These compounds were then compared against the DrugBank database of pharmaceutical compounds, with 205 of the predicted products found to be present in the database. These results provide a useful guide for the production of pharmaceutical compounds from *E. coli* since they provide not only a list of 205 promising targets but also a list of the predicted CDD superfamily of the enzyme(s) capable of producing each target. This information can provide researchers with a starting point for the identification of candidate enzymes for any reaction of interest. They potentially can do this not only by finding annotated genes which catalyze similar chemistry, but also by identifying unannotated ORFs which belong to these CDD superfamilies and may therefore be able to catalyze the reactions of interest. These results, however, are merely an example of the power of this technique. Different sets of operators can be rapidly generated using other metrics for reaction grouping such as Enzyme Commission (EC) classification or organism of origin for the associated enzymes. Additionally, these operators can be applied to different metabolic models and compared against different databases such as the Zinc database of commercially available compounds [72] for different applications. Once again, one of the most powerful aspects of this new tool is its flexibility.

5.2 Recommendations for Future Work

We present here recommendations for potential improvements and new applications for the BNICE framework and the automatic operator generation procedure.

5.2.1 Improved Atom Mapping for Improved Reaction Assignment to Reaction Site Operators

Our study of the minimum reaction site operator set required to describe all the reactions in MetaCyc found that only 1,497 reaction site operators were required to describe the MetaCyc database even though 2,210 reaction site operators were generated. This illustrates the degree of overlap between the reaction site operators. In all the cases where a reaction can be described by multiple reaction site operators, the reaction site operator with which the reaction is grouped is decided by the atom mappings provided by MetaCyc. Since the identification of the reaction site is a key step in the generation of intermediate operators, an error in the atom mapping can result in incorrect intermediate operators. In their generation of the atom mappings in MetaCyc, Latendresse et al. calculated an error rate of 0.9% and generated multiple optimal atom mappings for 2.1% of the reactions [61]. Fortunately, by identifying reactions which are described by multiple operators, we can rapidly find reactions whose atom mappings may require additional study and, if necessary, manual curation.

5.2.2 Improved Cost Function for Intermediate Operator Calculation

The calculation of the intermediate operators in this work utilized a cost function based upon a simple count of the number of generalized atom/bond types present in the operator. However, this simple cost function may fail to correctly weight attributes of the substrates in order to correctly predict the most likely reactions. One attribute that warrants particular study is the distance a node is from the reaction site. Nodes that are closer to the reaction site often have a more prominent role in stabilizing a reaction. Additionally, the cost function may be modified to account for the atom types of the nodes being compared. The difference between some atom types (for example halogens with other halogens) may be significantly less than between other (for example, a halogen with a carbon) in regards to enzymatic activity against a substrate. Both of these changes could be implemented without significant changes to the execution or speed of the dynamic programming algorithm used to generate intermediate operators.

5.2.2 Extending Automatic Operator Generation to Additional Databases

In this work, the operators utilized were based solely upon the MetaCyc database of reactions. As shown in Chapter 3, we found only a 40.1% overlap between MetaCyc and KEGG and a 72.0% overlap between MetaCyc and the iMM904 metabolic model. For more robust reaction prediction, the generated operators should be extended to include these and other databases and metabolic models such as UMBBD [28] and BRENDA [15]. Many databases specialize in particular applications; for example, UMBBD focuses on biodegradation, while BRENDA deals with enzyme promiscuity. Including reactions from these databases into our operators would allow for higher quality results when utilizing the automatically generated operators in these fields.

5.2.3 Gap Filling with BNICE

Most utilizations of BNICE to date have focused on the exploration of novel enzymatic pathways for the production of commercially valuable compounds. However, one underutilized potential application of BNICE is to fill gaps in metabolic models. Metabolic models have a wide variety of applications, including optimizing the production of a compound [73], determining whether a knockout will grow [74], and calculating minimum growth medium [75]. However, the accuracy of these models can be compromised by our incomplete knowledge of biochemistry. Frequently, this lack of knowledge takes the form of a gap in the metabolic model. A gap occurs when a compound is known to exist in an organism, but no reaction is known for its production or utilization. BNICE has the potential to fill many of these gaps through the prediction of *de novo* reactions and pathways to/from other known compounds in the organism. Operators based upon genetic similarity like the CDD superfamily operators utilized in Chapter 4 would be ideal for this application. The genome of the organism being studied can be searched for ORFs similar to each CDD superfamily. If any superfamily is not represented in the genome, the associated operators can be removed from the analysis. This would allow the BNICE predictions to be filtered to ensure better results. Furthermore, if a promising pathway to/from a compound is identified with the CDD superfamily operators, then a list of candidate enzymes can rapidly be determined by comparing the ORFs in the genome to the CDD superfamilies associated with the pathway. This in turn could guide knock-out studies to confirm the predicted enzymatic pathway.

References

[1] M. Moura, L. Broadbelt, and K. E. J. Tyo, "Computational tools for guided discovery and engineering of metabolic pathways," in *Systems metabolic engineering : methods and protocols*, H. S. Alper, Ed., ed New York: Humana Press, 2013, pp. 123-148.

[2] G. Rodrigo, J. Carrera, K. J. Prather, and A. Jaramillo, "DESHARKY: automatic design of metabolic pathways for optimal cell growth," *Bioinformatics,* vol. 24, pp. 2554-2556, Nov 1 2008.

[3] C. H. Chou, W. C. Chang, C. M. Chiu, C. C. Huang, and H. D. Huang, "FMM: a web server for metabolic pathway reconstruction and comparative analysis," *Nucleic Acids Research,* vol. 37, pp. W129-W134, Jul 1 2009.

[4] P. Carbonell, D. Fichera, S. B. Pandit, and J. L. Faulon, "Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms," *Bmc Systems Biology,* vol. 6, Feb 6 2012.

[5] A. P. Heath, G. N. Bennett, and L. E. Kavraki, "Finding metabolic pathways using atom tracking," *Bioinformatics,* vol. 26, pp. 1548-1555, Jun 15 2010.

[6] V. Hatzimanikatis, C. H. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, "Exploring the diversity of complex metabolic networks," *Bioinformatics,* vol. 21, pp. 1603-1609, Apr 15 2005.

[7] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler*, et al.*, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research,* vol. 40, pp. D742-D753, Jan 2012.

[8] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research,* vol. 28, pp. 27-30, Jan 1 2000.

[9] M. L. Mo, B. O. Palsson, and M. J. Herrgard, "Connecting extracellular metabolomic measurements to intracellular flux states in yeast," *Bmc Systems Biology,* vol. 3, Mar 25 2009.

[10] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. N. Lu, F. Chitsaz, L. Y. Geer, *et al.*, "CDD: NCBI's conserved domain database," *Nucleic Acids Research,* vol. 43, pp. D222-D226, Jan 28 2015.

[11] V. G. Yadav, M. De Mey, C. G. Lim, P. K. Ajikumar, and G. Stephanopoulos, "The future of metabolic engineering and synthetic biology: Towards a systematic practice," *Metabolic Engineering,* vol. 14, pp. 233-241, May 2012.

[12] B. A. Boghigian, G. Seth, R. Kiss, and B. A. Pfeifer, "Metabolic flux analysis and pharmaceutical production," *Metabolic Engineering,* vol. 12, pp. 81-95, Mar 2010.

[13] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research,* vol. 42, pp. D199-D205, Jan 2014.

[14] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, *et al.*, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Research,* vol. 42, pp. D459-D471, Jan 2014.

[15] I. Schomburg, A. Chang, S. Placzek, C. Sohngen, M. Rother, M. Lang, *et al.*, "BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA," *Nucleic Acids Research,* vol. 41, pp. D764-D772, Jan 2013.

[16] A. Mithani, G. M. Preston, and J. Hein, "Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison," *Bioinformatics,* vol. 25, pp. 1831-1832, Jul 15 2009.

[17] I. Rocha, P. Maia, P. Evangelista, P. Vilaca, S. Soares, J. P. Pinto, *et al.*, "OptFlux: an open-source software platform for in silico metabolic engineering," *Bmc Systems Biology,* vol. 4, Apr 19 2010.

[18] H. Nam, N. E. Lewis, J. A. Lerman, D. H. Lee, R. L. Chang, D. Kim, *et al.*, "Network Context and Selection in the Evolution to Enzyme Specificity," *Science,* vol. 337, pp. 1101-1104, Aug 31 2012.

[19] J. Gonzalez-Lergier, L. J. Broadbelt, and V. Hatzimanikatis, "Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways," *Journal of the American Chemical Society,* vol. 127, pp. 9930-9938, Jul 13 2005.

[20] C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, V. Hatzimanikatis, and L. J. Broadbelt, "Computational discovery of biochemical routes to specialty chemicals," *Chemical Engineering Science,* vol. 59, pp. 5051-5060, 2004.

[21] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, "Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate," *Biotechnol Bioeng,* vol. 106, pp. 462-73, Jun 15 2010.

[22] J. Coral, S. G. Karp, L. P. D. Vandenberghe, J. L. Parada, A. Pandey, and C. R. Soccol, "Batch Fermentation Model of Propionic Acid Production by Propionibacterium acidipropionici in Different Carbon Sources," *Applied Biochemistry and Biotechnology,* vol. 151, pp. 333-341, Dec 2008.

[23] B. A. Rodriguez, C. C. Stowers, V. Pham, and B. M. Cox, "The production of propionic acid, propanol and propylene via sugar fermentation: an industrial perspective on the progress, technical challenges and future outlook," *Green Chemistry,* vol. 16, pp. 1066-1076, 2014.

[24] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, "Discovery and Analysis of Novel Metabolic Pathways for the Biosynthesis of Industrial Chemicals: 3-Hydroxypropanoate," *Biotechnology and Bioengineering,* vol. 106, pp. 462-473, Jun 15 2010.

[25] D. Wu, Q. Wang, R. S. Assary, L. J. Broadbelt, and G. Krilov, "A Computational Approach To Design and Evaluate Enzymatic Reaction Pathways: Application to 1-Butanol Production from Pyruvate," *Journal of Chemical Information and Modeling,* vol. 51, pp. 1634-1647, Jul 2011.

[26] S. D. Finley, L. J. Broadbelt, and V. Hatzimanikatis, "In silico feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene," *Bmc Systems Biology,* vol. 4, Feb 2 2010.

[27] K. Tipton and S. Boyce, "History of the enzyme nomenclature system," *Bioinformatics,* vol. 16, pp. 34-40, Jan 2000.

[28] L. B. M. Ellis, D. Roe, and L. P. Wackett, "The University of Minnesota Biocatalysis/Biodegradation Database: the first decade," *Nucleic Acids Research,* vol. 34, pp. D517-D521, Jan 1 2006.

[29] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp*, et al.*, "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information," *Molecular Systems Biology,* vol. 3, Jun 2007.

[30] D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison, 3rd, and H. O. Smith, "Enzymatic assembly of DNA molecules up to several hundred kilobases," *Nat Methods,* vol. 6, pp. 343-5, May 2009.

[31] B. J. Brown, Z. Deng, P. A. Karplus, and V. Massey, "On the active site of old yellow enzyme - Role of histidine 191 and asparagine 194," *Journal of Biological Chemistry,* vol. 273, pp. 32753-32762, Dec 4 1998.

[32] M. Yamaguchi, "Studies on Regulatory Functions of Malic Enzymes .4. Effects of Sulfhydryl-Group Modification on the Catalytic Function of Nad-Linked Malic Enzyme from Escherichia-Coli," *Journal of Biochemistry,* vol. 86, pp. 325-333, 1979.

[33] A. J. Hillier, R. E. Jericho, S. M. Green, and G. R. Jago, "Properties and function of fumarate reductase (NADH) in Streptococcus lactis," *Aust J Biol Sci,* vol. 32, pp. 625-35, Dec 1979.

[34] F. J. Ruzicka and P. A. Frey, "Kinetic and Spectroscopic Evidence of Negative Cooperativity in the Action of Lysine 2,3-Aminomutase," *Journal of Physical Chemistry B,* vol. 114, pp. 16118-16124, Dec 16 2010.

[35] H. S. Toogood, J. M. Gardiner, and N. S. Scrutton, " Biocatalytic Reductions and Chemical Versatility of the Old Yellow Enzyme Family of Flavoprotein Oxidoreductases," *ChemCatChem,* vol. 2, pp. 892–914, 2010.

[36] E. W. Trotter, E. J. Collinson, I. W. Dawes, and C. M. Grant, "Old yellow enzymes protect against acrolein toxicity in the yeast Saccharomyces cerevisiae," *Appl Environ Microbiol,* vol. 72, pp. 4885-92, Jul 2006.

[37] A. Muller, B. Hauer, and B. Rosche, "Asymmetric alkene reduction by yeast old yellow enzymes and by a novel Zymomonas mobilis reductase," *Biotechnol Bioeng,* vol. 98, pp. 22-9, Sep 1 2007.

[38] M. Hall, C. Stueckler, B. Hauer, R. Stuermer, T. Friedrich, M. Breuer, *et al.*, "Asymmetric bioreduction of activated C = C bonds using Zymomonas mobilis NCR enoate reductase and old yellow enzymes OYE 1-3 from yeasts," *European Journal of Organic Chemistry,* pp. 1511-1516, Mar 2008.

[39] F. Rohdich, A. Wiese, R. Feicht, H. Simon, and A. Bacher, "Enoate reductases of Clostridia - Cloning, sequencing, and expression," *Journal of Biological Chemistry,* vol. 276, pp. 5779-5787, Feb 23 2001.

[40] G. Ondrey, "A step closer to bio-based acrylic acid," *Chemical Engineering,* vol. 121, pp. 18-18, Oct 2014.

[41] J. Moyle and M. Dixon, "Purification of the Isocitric Enzyme (Triphosphopyridine Nucleotide-Linked Isocitric Dehydrogenase-Oxalosuccinic Carboxylase)," *Biochemical Journal,* vol. 63, pp. 548-552, 1956.

[42] P. Dimroth, "Characterization of a Membrane-Bound Biotin-Containing Enzyme - Oxaloacetate Decarboxylase from Klebsiella-Aerogenes," *European Journal of Biochemistry,* vol. 115, pp. 353-358, 1981.

[43] J. M. Williamson and G. M. Brown, "Purification and Properties of L-Aspartate-Alpha-Decarboxylase, an Enzyme That Catalyzes the Formation of Beta-Alanine in Escherichia-Coli," *Journal of Biological Chemistry,* vol. 254, pp. 8074-8082, 1979.

[44] L. Liu, Y. Zhu, J. Li, M. Wang, P. Lee, G. Du, *et al.*, "Microbial production of propionic acid from propionibacteria: current state, challenges and perspectives," *Critical reviews in biotechnology,* vol. 32, pp. 374-81, Dec 2012.

[45] M. Hetzel, M. Brock, T. Selmer, A. J. Pierik, B. T. Golding, and W. Buckel, "Acryloyl-CoA reductase from Clostridium propionicumi - An enzyme complex of propionyl-CoA dehydrogenase and electron-transferring flavoprotein," *European Journal of Biochemistry,* vol. 270, pp. 902-910, Mar 2003.

[46] R. M. Kohli and V. Massey, "The oxidative half-reaction of Old Yellow Enzyme. The role of tyrosine 196," *Journal of Biological Chemistry,* vol. 273, pp. 32763-70, Dec 4 1998.

[47] E. Brenna, F. G. Gatti, A. Manfredi, D. Monti, and F. Parmeggiani, "Biocatalyzed Enantioselective Reduction of Activated C=C Bonds: Synthesis of Enantiomerically Enriched alpha-Halo-beta-arylpropionic Acids," *European Journal of Organic Chemistry,* pp. 4015-4022, Jul 2011.

[48] V. Massey, "The chemical and biological versatility of riboflavin," *Biochem Soc Trans,* vol. 28, pp. 283-96, 2000.

[49] C. Stueckler, M. Hall, H. Ehammer, E. Pointner, W. Kroutil, P. Macheroux*, et al.*, "Stereocomplementary bioreduction of alpha,beta-unsaturated dicarboxylic acids and dimethyl esters using enoate reductases: enzyme- and substrate-based stereocontrol," *Org Lett,* vol. 9, pp. 5409-11, Dec 20 2007.

[50] G. Oberdorfer, K. Gruber, K. Faber, and M. Hall, "Stereocontrol Strategies in the Asymmetric Bioreduction of Alkenes," *Synlett,* pp. 1857-1864, Aug 2012.

[51] C. K. Winkler, G. Tasnadi, D. Clay, M. Hall, and K. Faber, "Asymmetric bioreduction of activated alkenes to industrially relevant optically active compounds," *Journal of Biotechnology,* vol. 162, pp. 381-389, Dec 31 2012.

[52] A. Cho, H. Yun, J. H. Park, S. Y. Lee, and S. Park, "Prediction of novel synthetic pathways for the production of desired chemicals," *Bmc Systems Biology,* vol. 4, Mar 28 2010.

[53] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto*, et al.*, "PathPred: an enzyme-catalyzed metabolic pathway prediction server," *Nucleic Acids Research,* vol. 38, pp. W138-W143, Jul 2010.

[54] K. C. Soh and V. Hatzimanikatis, "DREAMS of metabolism," *Trends in Biotechnology,* vol. 28, pp. 501-508, Oct 2010.

[55] J. F. Gao, L. B. M. Ellis, and L. P. Wackett, "The University of Minnesota Pathway Prediction System: multi-level prediction and visualization," *Nucleic Acids Research,* vol. 39, pp. W406-W411, Jul 2011.

[56] H. Yim, R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, *et al.*, "Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol," *Nature Chemical Biology,* vol. 7, pp. 445-452, Jul 2011.

[57] D. Wu, D. J. Yue, F. Q. You, and L. J. Broadbelt, "Computational evaluation of factors governing catalytic 2-keto acid decarboxylation," *Journal of Molecular Modeling,* vol. 20, Jun 2014.

[58] P. Carbonell, A. G. Planson, D. Fichera, and J. L. Faulon, "A retrosynthetic biology approach to metabolic pathway design for therapeutic production," *Bmc Systems Biology,* vol. 5, Aug 5 2011.

[59] M. Leber, V. Egelhofer, I. Schomburg, and D. Schomburg, "Automatic assignment of reaction operators to enzymatic reactions," *Bioinformatics,* vol. 25, pp. 3135-3142, Dec 1 2009.

[60] I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, *et al.*, "New Applications of Computers in Chemistry," *Angewandte Chemie-International Edition in English,* vol. 18, pp. 111-123, 1979.

[61] M. Latendresse, J. P. Malerich, M. Travers, and P. D. Karp, "Accurate Atom-Mapping Computation for Biochemical Reactions," *Journal of Chemical Information and Modeling,* vol. 52, pp. 2970-2982, Nov 2012.

[62] E. L. First, C. E. Gounaris, and C. A. Floudas, "Stereochemically Consistent Reaction Mapping and Identification of Multiple Reaction Mechanisms through Integer Linear Optimization," *Journal of Chemical Information and Modeling,* vol. 52, pp. 84-92, Jan 2012.

[63] D. Fooshee, A. Andronico, and P. Baldi, "Reaction Map: An Efficient Atom-Mapping Algorithm for Chemical Reactions," *Journal of Chemical Information and Modeling,* vol. 53, pp. 2812-2819, Nov 2013.

[64] M. Arita, "The metabolic world of Escherichia coli is not small," *Proc Natl Acad Sci U S A,* vol. 101, pp. 1543-7, Feb 10 2004.

[65] C. A. Shelley and M. E. Munk, "Approach to the Assignment of Canonical Connection Tables and Topological Symmetry Perception," *Journal of Chemical Information and Computer Sciences,* vol. 19, pp. 247-250, 1979.

[66] J. Munkres, "Algorithms for the Assignment and Transportation Problems," *Journal of the Society for Industrial and Applied Mathematics,* vol. 5, pp. 32-38, 1957.

[67] T. Altman, M. Travers, A. Kothari, R. Caspi, and P. D. Karp, "A systematic comparison of the MetaCyc and KEGG pathway databases," *Bmc Bioinformatics,* vol. 14, Mar 27 2013.

[68] D. A. Pertusi, A. E. Stine, L. J. Broadbelt, and K. E. J. Tyo, "Efficient searching and annotation of metabolic networks using chemical similarity," *Bioinformatics,* vol. 31, pp. 1016-1024, Apr 1 2015.

[69] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist*, et al.*, "A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011," *Molecular Systems Biology,* vol. 7, Oct 2011.

[70] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. F. Liu*, et al.*, "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research,* vol. 42, pp. D1091-D1097, Jan 2014.

[71] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard*, et al.*, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research,* vol. 34, pp. D668-D672, Jan 1 2006.

[72] J. Irwin and B. Shoichet, "The ZINC database as a new research tool for ligand discovery," *Abstracts of Papers of the American Chemical Society,* vol. 230, pp. U1009-U1009, Aug 28 2005.

[73] H. Alper, K. Miyaoku, and G. Stephanopoulos, "Construction of lycopene-overproducing E-coli strains by combining systematic and combinatorial gene knockout targets," *Nature Biotechnology,* vol. 23, pp. 612-616, May 2005.

[74] A. R. Zomorrodi and C. D. Maranas, "Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data," *Bmc Systems Biology,* vol. 4, Dec 29 2010.

[75] A. K. Chavali, J. D. Whittemore, J. A. Eddy, K. T. Williams, and J. A. Papin, "Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major," *Molecular Systems Biology*, vol. 4, Mar 2008.

Appendix

Appendix A: Cofactors Used for Automatic Operator Generation

In Tables A.1 and A.2 below are listed the cofactors used for automatic operator generation in this thesis. Table A.1 lists compounds that are always treated as cofactors whenever they appear in a reaction. Table A.2 lists pairs of compounds which are treated as cofactors by this program when they appear on opposite sides of a reaction. Both tables lists the common name for the cofactors and the notation used by BNICE to represent them. This illustrates how cofactors are grouped together by the operator generation program. Labels for cofactor pairs all contain the phrase “CoF” (for cofactor) in order to distinguish them from their non-cofactor version.

Table A.1: List of BNICE cofactors.

Common Name	BNICE Nomenclature
Carbon dioxide	CO2
Hydrogen bromide	HBr
Hydrogen chloride	HCl
Hydrogen fluoride	HF
Hydrogen iodide	HI
Ammonia	NH3
Oxygen	O2
Phosphate	Pi
Diphosphate	PPi
Sulfite	Sulfite
Hydrogen peroxide	H2O2
Coenzyme A	CoA
Acetyl coenzyme A	ACETYL-COA
Carbonate	CO3
Water	WATER
Sulfate	SULFATE
Carbon monoxide	CARBON-MONOXIDE
Cyanide	HCN
Hydrogen sulfide	HS
Sulfite	Sulfite

Table A.1: List of BNICE cofactor pairs.

Cofactor 1 common name	Cofactor 2 common name	Cofactor 1 BNICE Nomenclature	Cofactor 2 BNICE Nomenclature
5-formyl-tetrahydromethanopterin	Tetrahydromethanopterin	5-Formyl-H4MPT_CoF	H4MPT_CoF
Adenosine-triphosphate	Adenosine-diphosphate	ATP_CoF	ADP_CoF
Adenosine-triphosphate	Adenosine-monophosphate	ATP_CoF	AMP_CoF
Adenosine-diphosphate	Adenosine-monophosphate	ADP_CoF	AMP_CoF
Cytidine-triphosphate	Cytidine-diphosphate	ATP_CoF	ADP_CoF
Cytidine-triphosphate	Cytidine-monophosphate	ATP_CoF	AMP_CoF
Cytidine-diphosphate	Cytidine-monophosphate	ADP_CoF	AMP_CoF
Guanosine-triphosphate	Guanosine-diphosphate	ATP_CoF	ADP_CoF
Guanosine-triphosphate	Guanosine-monophosphate	ATP_CoF	AMP_CoF
Guanosine-diphosphate	Guanosine-monophosphate	ADP_CoF	AMP_CoF
Uridine-triphosphate	Uridine-diphosphate	ATP_CoF	ADP_CoF
Uridine-triphosphate	Uridine-monophosphate	ATP_CoF	AMP_CoF
Uridine-diphosphate	Uridine-monophosphate	ADP_CoF	AMP_CoF
Generalized Nucleoside-Triphosphate	Generalized Nucleoside-diphosphate	ATP_CoF	ADP_CoF
Generalized Nucleoside-Triphosphate	Generalized Nucleoside-monophosphate	ATP_CoF	AMP_CoF
Generalized Nucleoside-diphosphate	Generalized Nucleoside-monophosphate	ADP_CoF	AMP_CoF
Diphosphopyridine nucleotide oxidized	Diphosphopyridine nucleotide reduced	NAD_CoF	NADH_CoF

Nicotinamide adenine dinucleotide phosphate	Reduced nicotinamide adenine dinucleotide phosphate	NAD_CoF	NADH_CoF
Flavin adenine dinucleotide reduced	Flavin adenine dinucleotide oxidized	FADH2_CoF	FAD_CoF
S-adenosyl-methionine	S-adenosyl-homocysteine	S-Adenosylmethionine_CoF	S-Adenosylhomocysteine_CoF
Phospho-Histidine	Histidine	Phospho-Histidine_CoF	Histidine_CoF
2-oxoglutarate	L-Glutamate	2-oxoglutarate_CoF	L-Glutamate_CoF
2-oxoglutarate	Succinate	2-oxoglutarate_CoF	succinate_CoF
L-Glutamine	L-Glutamate	L-Glutamine_CoF	L-Glutamate_CoF
Oxidized Coenzyme F420	Reduced Coenzyme F420	Oxidized-Factor-F420_CoF	Reduced-Factor-F420_CoF
Adenosine 3',5'-bisphosphate	3'-phosphoadenylyl-sulfate	3-5-ADP_CoF	PAPS_CoF
Dimethylallyl diphosphate	Diphosphate	DMAPP_CoF	PPi_CoF
Ubiquinol	Ubiquinones	Ubiquinol_CoF	Ubiquinones_CoF
Sulfurated-Sulfur-Acceptors	Unsulfurated-Sulfur-Acceptors	Sulfurated-Sulfur-Acceptors_CoF	Unsulfurated-Sulfur-Acceptors_CoF
Demethylated-methyl-acceptors	Methylated-methyl-acceptors	Demethylated-methyl-acceptors_CoF	Methylated-methyl-acceptors_CoF
Deaminated-Amine-Donors	Aminated-Amine-Donors	Deaminated-Amine-Donors_CoF	Aminated-Amine-Donors_CoF

Appendix B: DrugBank Compounds Produced by the CDD Superfamily Intermediate Operators

Table B.1 lists all compounds in the DrugBank database which were produce by the application of the CDD superfamily intermediate operators to the iJO1366 metabolic model of *E. coli.* This table includes the DrugBank ID for each compound, its common name, and every CDD superfamily predicted by BNICE to produce it.

Table B.1: List of DrugBank compounds produced by the CDD superfamily intermediate operators.

DataBank ID	Common Name	CDD Superfamily for Production
DB00341	Cetirizine	cl12078; cl19134

DB00444	Teniposide	cl12078;
		cl18408;
		cl18680

DB00448	Lansoprazole	cl12078
DB00525	Tolnaftate	cl18945
DB00616	Candoxatril	cl18945
DB00665	Nilutamide	cl18408; cl12078

DB00800	Fenoldopam	cl19241;
		cl17182;
		cl00841

DB00926	Etretinate	cl19134
DB01037	Selegiline	cl17173

DB01091	Butenafine	cl18408;
		cl12078;
		cl02872;
		cl09931;
		cl00470

DB01092	Ouabain	cl18949;
		cl19137;
		cl13995;
		cl00289;
		cl00447;
		cl00474

DB01201	Rifapentine	cl17255
DB01298	Sulfacytine	cl17255
DB01320	Fosphenytoin	cl19134; cl12078
DB01324	Polythiazide	cl17068
DB01400	Neostigmine	cl18408; cl12078

DB01413	Cefepime	cl17186;
		cl02872;
		cl09931;

		cl00470
DB01436	Alfacalcidol	cl16912; cl00841
DB01478	desmethylprodine	cl16912; cl02872; cl09931; cl00470
DB01530	3Alpha,17beta-dihydroxy-5alpha-androstane	cl18408; cl12078
DB01531	Desomorphine	cl19134; cl17068; cl00470
DB01539	1-Piperidinocyclohexanecarbonitrile	cl12078
DB01541	Boldenone	cl17240; cl12078; cl02872; cl09931; cl00470
DB01547	Drotebanol	cl16912; cl02872; cl09931; cl00470
DB01563	Chloral hydrate	cl12078
DB01642	O1-Methyl-Glucose	cl18949; cl19058; cl19137; cl00474
DB01649	7-Methyl-Gpppa	cl17068
DB01651	Methyl 4,6-O-[(1r)-1-Carboxyethylidene]-Beta-D-Galactopyranoside	cl18945
DB01679	Propyl Trihydrogen Diphosphate	cl16912; cl02872; cl09931; cl00470
DB01703	N-(2-Ferrocenylethyl)Maleimide	cl18945; cl17068; cl00470
DB01712	(3r)-4-(P-Toluenesulfonyl)-1,4-Thiazane-3-Carboxylicacid-L-Phenylalanine Ethyl Ester	cl19134
DB01749	1,2-Dimethoxyethane	cl12078
DB01785	Dimethylallyl Diphosphate	cl18216; cl03532
DB01795	Phenyl Boronic Acid	cl18945

DB01806	10-{4-Dimethylamino-5-[4-Hydroxy-6-Methyl-5-(6-Methyl-5-Oxo-Tetrahydro-Pyran-2-Yloxy)-Tetrahydro-Pyrane-2-Yloxy]-6-Methyl-Tetrahydro-Pyran-2-Yloxy}-8-Ethyl-1,8,11-Trihydroxy-7,8,9,10-Tetrahydro-Naphthacene-5,12-Dione	cl19134; cl12078
DB01824	(3s)-Tetrahydrofuran-3-Yl (1r,2s)-3-[4-((1r)-2-{[(S)-Amino(Hydroxy)Methyl]Oxy}-2,3-Dihydro-1h-Inden-1-Yl)-2-Benzyl-3-Oxopyrrolidin-2-Yl]-1-Benzyl-2-Hydroxypropylcarbamate	cl12078
DB01854	5-Bromonicotinamide	cl12078
DB01865	3-(6-Aminopyridin-3-Yl)-N-Methyl-N-[(1-Methyl-1h-Indol-2-Yl)Methyl]Acrylamide	cl18945
DB01872	Acetylgalactosamine-4-Sulfate	cl18408; cl12078
DB01897	2-(2f-Benzothiazolyl)-5-Styryl-3-(4f-Phthalhydrazidyl)Tetrazolium Chloride	cl19134

DB01899	Nd1-Phosphonohistidine	cl18949;
		cl19137;
		cl13995;
		cl11399;
		cl00289;
		cl00447;
		cl00474; ; ;

DB01930	2,4-Dihydroxy-3,3-Dimethyl-Butyrate	cl18408;
		cl18680;
		cl12078;
		cl04742

DB01941	6-[1-(3,5,5,8,8-Pentamethyl-5,6,7,8-Tetrahydronaphthalen-2-Yl)Cyclopropyl]Pyridine-3-Carboxylic Acid	cl12078

DB01960	7n-Methyl-8-Hydroguanosine-5'-Diphosphate	cl16913;
		cl17173;
		cl06920

DB01979	Methyl alpha-D-mannoside	cl17240;
		cl12078;
		cl02872

DB02006	Br-Coeleneterazine	cl17190;
		cl15968;
		cl17037;
		cl08484;
		cl00192;
		cl00841

DB02007	Alpha-D-Glucose-6-Phosphate	cl17190;
		cl15968;
		cl17037;

		cl08484;
		cl00192;
		cl00841

DB02010	Staurosporine	cl19134; cl12078

DB02041	4-Aminophthalhydrazide	cl17190;
		cl15968;
		cl08484;
		cl00192;
		cl00841

DB02059	Adenosine-5-Diphosphoribose	cl12078
DB02068	Delta-Amino Valeric Acid	cl18408; cl12078
DB02079	(Aminooxy)Acetic Acid	cl17255

DB02091	4-(2,4-Dimethyl-Thiazol-5-Yl)-Pyrimidin-2-Ylamine	cl19134;
		cl17068;
		cl00470

DB02093	5-Phospho-D-Arabinohydroxamic Acid	cl18945;
		cl16912;
		cl00470

DB02126	4-Carboxycinnamic Acid	cl17190;
		cl12283;
		cl15968;
		cl17037;
		cl08484;
		cl00289

DB02139	(2e)-N-Allyl-4-{[3-(4-Bromophenyl)-5-Fluoro-1-Methyl-1h-Indazol-6-Yl]Oxy}-N-Methyl-2-Buten-1-Amine	cl18408;
		cl18680;
		cl12078

DB02165	Zinc Trihydroxide	cl12078

DB02203	Acetone Cyanohydrin	cl16912;
		cl02872;
		cl09931

DB02232	1,2-Dihydroxybenzene	cl18408; cl12078
DB02251	O-Succinylbenzoate	cl12078

DB02253	(1r)-4-[(1e,3e,5e,7z,9e,11z,13e,15e)-17-Hydroxy-3,7,12,16-Tetramethylheptadeca-1,3,5,7,9,11,13,15-Octaen-1-Yl]-3,5,5-Trimethylcyclohex-3-En-1-Ol	cl18945;
		cl17068;
		cl00470

DB02303	(5s)-5-Iododihydro-2,4(1h,3h)-Pyrimidinedione	cl18949;
		cl19137;
		cl15968;
		cl00289;
		cl00447;

		cl00474

DB02308	4-(1,3,2-Dioxaborolan-2-Yloxy)Butan-1-Aminium	cl16912;
		cl02872;
		cl09931;
		cl00470

DB02355	Adenosine-5'-Rp-Alpha-Thio-Triphosphate	cl12078

DB02363	2'-Monophosphoadenosine-5'-Diphosphate	cl17190;
		cl15968;
		cl17037;
		cl08484;
		cl00289

DB02381	Nor-N-Omega-Hydroxy-L-Arginine	cl16912;
		cl02872;
		cl09931;
		cl00470

DB02402	5-(4-Methoxyphenoxy)-2,4-Quinazolinediamine	cl18945;
		cl16912;
		cl00470

DB02425	Hexadecyl Octanoate	cl17195;
		cl17240;
		cl16912;
		cl02872;
		cl09931;
		cl00470

DB02483	Etheno-Nad	cl19134;
		cl17068;
		cl00470

DB02494	Alpha-Hydroxy-Beta-Phenyl-Propionic Acid	cl17240;
		cl12078;
		cl02872;
		cl09931;
		cl00470

DB02495	9-(4-hydroxybutyl)-N2-phenylguanine	cl12078
DB02496	1-Deoxy-D-xylulose 5-phosphate	cl19134
DB02503	4-(Carboxyvin-2-Yl)Phenylboronic Acid	cl12078

DB02526	CRA_10655	cl17195;
		cl17240;
		cl16912;
		cl02872;
		cl09931;
		cl00470

DB02572	BV4	cl18949;
		cl19137;
		cl11995;

		cl17169; cl17171; cl00289; cl00447; cl00474; cl00490; cl02570
DB02586	4,7-Dimethyl-[1,10]Phenanthroline	cl18408; cl12078
DB02589	Se-Ethyl-Isoselenourea	cl17255
DB02594	2'-Deoxycytidine	cl17255
DB02609	4-Hydroxy-L-Threonine-5-Monophosphate	cl12078
DB02674	4-(2-Oxo-Hexahydro-Thieno[3,4-D]Imidazol-4-Yl)-Butyricacid	cl18949; cl19137; cl00289; cl00447; cl00474
DB02675	(4-Hydroxymaltosephenyl)Glycine	cl17255
DB02683	Inhibitor Bea428	cl17190; cl12078; cl00841
DB02793	Isochorismic Acid	cl16912; cl02872; cl09931; cl00470
DB02800	5-Hydroxymethylene-6-Hydrofolic Acid	cl18408; cl18680; cl12078
DB02808	Trifluorofurnesyl Diphosphate	cl12078
DB02810	N-(2-Acetamido)Iminodiacetic Acid	cl16912
DB02829	4-(Acetylamino)-3-[(Aminoacetyl)Amino]Benzoic Acid	cl17240; cl12078; cl02872; cl00470
DB02859	Soraphen A	cl12078
DB02862	Gluco-Phenylimidazole	cl18945; cl16912; cl00470
DB02869	3-amino-5-phenylpentane	cl19139; cl10013
DB02875	CRA_1802	cl00220
DB02889	4-O-(4,6-Dideoxy-4-{[4,5,6-Trihydroxy-3-(Hydroxymethyl)Cyclohex-2-En-1-Yl]Amino}-Beta-D-	cl19134

	Lyxo-Hexopyranosyl)-Alpha-D-Erythro-Hexopyranose	
DB02893	D-Methionine	cl12078
DB02898	5-{[(2-Amino-9h-Purin-6-Yl)Oxy]Methyl}-2-Pyrrolidinone	cl18408; cl12078
DB02916	[(2r,3s,4r,5r)-5-(6-Amino-9h-Purin-9-Yl)-3,4-Dihydroxytetrahydro-2-Furanyl]Methyl Sulfamate	cl18945
DB02938	Heptanoic Acid	cl12078
DB02957	Orotidine-5'-Monophosphate	cl18945; cl17182; cl00470
DB02963	(5-Chloropyrazolo[1,5-a]Pyrimidin-7-Yl)-(4-Methanesulfonylphenyl)Amine	cl18408; cl18680; cl12078
DB03008	5-Fluoro-Beta-L-Gulosyl Fluoride	cl17186; cl02872; cl09931; cl00470
DB03009	2-[(2-Oxo-2-Piperidin-1-Ylethyl)Thio]-6-(Trifluoromethyl)Pyrimidin-4(1h)-One	cl18945
DB03035	1,8-Di-Hydroxy-4-Nitro-Anthraquinone	cl18949; cl19137; cl15968; cl17037; cl00289; cl00447; cl00474; ; ;
DB03065	7-Nitroindazole-2-Carboxamidine	cl12078
DB03087	2-(Sec-Butyl)Thiazole	cl16912; cl02872; cl09931; cl00470
DB03092	5-Hydroxymethyl-Chonduritol	cl18945
DB03096	N-Aminoethylmorpholine	cl12078
DB03103	Thymidine-5'- Diphosphate	cl19134
DB03126	Mant-Adp	cl19134
DB03186	U-Pi-a-Pi	cl18408; cl18680; cl12078
DB03229	2-Oxo-4-Methylpentanoic Acid	cl12078
DB03232	2-[(2e,6e,10e,14e,18e,22e,26e)-3,7,11,15,19,23,27,31-Octamethyldotriaconta-2,6,10,14,18,22,26,30-Octaenyl]Phenol	cl16912; cl02872; cl09931
DB03239	3',5'-Dinitro-N-Acetyl-L-Thyronine	cl18949;

		cl19137;
		cl00289;
		cl00447;
		cl00474

DB03240	(S)-2-Amino-3-(1,3,5,7-Pentahydro-2,4-Dioxo-Cyclopenta[E]Pyrimidin-1-Yl) Proionic Acid	cl19134; cl12078

DB03241	1-Amino-1-Carbonyl Pentane	cl18949;
		cl19137;
		cl17186;
		cl00289;
		cl00447;
		cl00474

DB03262	Al-6619, [2h-Thieno[3,2-E]-1,2-Thiazine-6-Sulfonamide,2-(3-Hydroxyphenyl)-3-(4-Morpholinyl)-, 1,1-Dioxide]	cl12078
DB03276	4-[(10s,14s,18s)-18-(2-Amino-2-Oxoethyl)-14-(1-Naphthylmethyl)-8,17,20-Trioxo-7,16,19-Triazaspiro[5.14]Icos-11-En-10-Yl]Benzylphosphonic Acid	cl12078
DB03305	N5-Iminoethyl-L-Ornithine	cl19134
DB03325	Tyrosyladenylate	cl11394
DB03331	N-Naphthalen-1-Ylmethyl-2'-[3,5-Dimethoxybenzamido]-2'-Deoxy-Adenosine	cl12078

DB03351	Sri-9439	cl18945;
		cl16912;
		cl00470

DB03352	S-Arsonocysteine	cl17173
DB03355	5'-O-(N-(L-Threonyl)-Sulfamoyl)Adenosine	cl18945
DB03368	5-Methyl-5-(4-Phenoxy-Phenyl)-Pyrimidine-2,4,6-Trione	cl18945

DB03380	L-Tyrosinamide	cl18408;
		cl12078;
		cl00013

DB03411	2-Hydroxymethyl-Pyrrolidine-3,4-Diol	cl17173
DB03444	(3e)-6'-Bromo-2,3'-Biindole-2',3(1h,1'h)-Dione 3-Oxime	cl19134; cl12078

DB03445	Tazobactam Trans-Enamine Intermediate	cl18949;
		cl19137;
		cl15968;
		cl17037;
		cl08484;
		cl00841

DB03494	CRA_10950	cl17255
DB03526	AL5927	cl17346

DB03539	2-(Acetylamino)-2-Deoxy-6-O-Methyl-Alpha-D-Allopyranose	cl17190;
		cl12283;

		cl15968; cl00841
DB03540	Norcamphor	cl18216; cl03532

DB03586	5(R)-5-Fluoro-Beta-D-Xylopyranosyl-Enzyme Intermediate	cl12078;
		cl09931;
		cl00309

DB03623	9-(4-Hydroxyphenyl)-2,7-Phenanthroline	cl17240;
		cl12078;
		cl02872;
		cl09931;
		cl00470

DB03627	Adamantane	cl16912;
		cl02872;
		cl09931

DB03647	3-[Isopropyl(4-Methylbenzoyl)Amino]-5-Phenylthiophene-2-Carboxylic Acid	cl19139; cl10013
DB03651	2,4,6-Trinitrophenol	cl16913; cl17173
DB03699	Succinyl-Coenzyme A	cl12078
DB03715	Pentadecane	cl12078
DB03736	2-Cyclopropylmethylenepropanal	cl12078
DB03754	Tris(Hydroxymethyl)Aminomethane	cl12078
DB03779	Glucosaminyl-(Alpha-6)-D-Myo-Inositol	cl17173
DB03812	3-{2,6,8-Trioxo-9-[(2s,3r,4r)-2,3,4,5-Tetrahydroxypentyl]-1,2,3,6,8,9-Hexahydro-7h-Purin-7-Yl}Propyl Dihydrogen Phosphate	cl18945; cl12078
DB03818	N-[Tosyl-D-Prolinyl]Amino-Ethanethiol	cl19134; cl12078
DB03823	Epigallocatechin	cl17255
DB03826	5,6-Diaminouracil	cl18216; cl03532
DB03834	Tazobactam Intermediate	cl19134
DB03847	Gamma-Carboxy-Glutamic Acid	cl18945; cl17173

DB03859	1-Thio-Beta-D-Glucopyranose	cl17190;
		cl17270;
		cl15968;
		cl17037;
		cl00192;
		cl00841

DB03872	2,3-Dideoxyfucose	cl17255
DB03877	AL4623	cl16913;

		cl17173
DB03878	N-[4-Methyl-3-[[4-(3-Pyridinyl)-2-Pyrimidinyl]Amino]Phenyl]-3-Pyridinecarboxamide	cl12078
DB03887	Alpha-Adenosine Monophosphate	cl18945; cl16912; cl00470
DB03889	S-(N-Hydroxy-N-Bromophenylcarbamoyl)Glutathione	cl18949; cl19137; cl15968; cl11399; cl00289; cl00447; cl00474; ; ;
DB03890	N-[2-(1-Formyl-2-Methyl-Propyl)-1-(4-Piperidin-1-Yl-but-2-Enoyl)-Pyrrolidin-3-Yl]-Methanesulfonamide	cl12078
DB03909	Adenosine-5'-[Beta, Gamma-Methylene]Triphosphate	cl12078
DB03924	5,8-Di-Amino-1,4-Dihydroxy-Anthraquinone	cl16912
DB03926	5-Alpha-Androstane-3-Beta,17-Alpha-Diol	cl18949; cl19137; cl15968; cl02872; cl00470
DB03930	4-Methyl-Pyrroline-5-Carboxylic Acid	cl18949; cl19137; cl13995; cl00289; cl00447; cl00474
DB04027	D-Arginine	cl17187
DB04042	2-[4-(Hydroxy-Methoxy-Methyl)-Benzyl]-7-(4-Hydroxymethyl-Benzyl)-1,1-Dioxo-3,6-Bis-Phenoxymethyl-1lambda6-[1,2,7]Thiadiazepane-4,5-Diol	cl17173
DB04092	Apstatin	cl17190; cl15968; cl17037; cl08484; cl00192; cl00841
DB04123	(P-Iodophenylacetylamino)Methylphosphinic Acid	cl18216; cl03532
DB04195	Heptulose-2-Phosphate	cl17173
DB04840	Debrisoquin	cl19134; cl00013

DB04888	Bifeprunox	cl18945
DB04890	Bepotastine	cl12078

DB04894	Vapreotide	cl18949;
		cl19137;
		cl15968;
		cl17037;
		cl08484;
		cl00841

DB04930	Permethrin	cl18945

DB04961	Troxacitabine	cl18949;
		cl19137;
		cl15968;
		cl17037;
		cl00289;
		cl00447;
		cl00474; ; ;

DB04983	Denufosol	cl18945
DB05003	Imexon	cl18216; cl03532
DB05016	PTC124	cl12078
DB05271	Rotigotine	cl12078
DB05891	HD-O	cl11394; cl00015
DB06090	TC-5619	cl19137; cl11394
DB06402	Telavancin	cl12078
DB06594	Agomelatine	cl17186

DB06691	Mepyramine	cl17240;
		cl12078;
		cl02872;
		cl00470

DB06701	Dexmethylphenidate	cl16912;
		cl02872;
		cl09931;
		cl00470

DB06716	Fospropofol	cl18949;
		cl19137;
		cl15968;
		cl17037;
		cl00289;
		cl00447;
		cl00474; ; ;

DB06751	Drotaverine	cl17255

DB06999	N-{3-[(5-chloro-1H-pyrrolo[2,3-b]pyridin-3-yl)carbonyl]-2,4-difluorophenyl}propane-1-sulfonamide	cl12078
DB07018	5-ETHYL-3-[(2-METHOXYETHYL)METHYLAMINO]-6-METHYL-4-(3-METHYLBENZYL)PYRIDIN-2(1H)-ONE	cl12078
DB07037	(2S)-1-AMINO-3-[(5-NITROQUINOLIN-8-YL)AMINO]PROPAN-2-OL	cl17255
DB07111	(4S,5E,7Z,10Z,13Z,16Z,19Z)-4-hydroxydocosa-5,7,10,13,16,19-hexaenoic acid	cl17255
DB07368	4-(METHYLSULFONYL)BENZENECARBOXIMIDAMIDE	cl00210; cl00230
DB07444	6-(3-AMINOPROPYL)-4,9-DIMETHYLPYRROLO[3,4-C]CARBAZOLE-1,3(2H,6H)-DIONE	cl18408; cl18680; cl12078
DB07447	5-beta-DIHYDROTESTOSTERONE	cl16912; cl02872; cl09931; cl00470
DB07469	(3aS)-3a-hydroxy-7-methyl-1-phenyl-1,2,3,3a-tetrahydro-4H-pyrrolo[2,3-b]quinolin-4-one	cl18408; cl18680; cl12078; cl04742
DB07597	CIS-(1R,2S)-2-AMINO-1,2,3,4-TETRAHYDRONAPHTHALEN-1-OL	cl00303
DB07847	6-CHLORO-N-{(3S)-1-[(1S)-1-METHYL-2-(4-MORPHOLINYL)-2-OXO ETHYL]-2-OXO-3-PYRROLIDINYL}-2-NAPHTHALENESULFONAMIDE	cl17255
DB07859	4-(4-CHLOROPHENYL)-4-[4-(1H-PYRAZOL-4-YL)PHENYL]PIPERIDINE	cl17255
DB08198	[(4R)-4-(3-HYDROXYPHENYL)-1,6-DIMETHYL-2-THIOXO-1,2,3,4-TETRAHYDROPYRIMIDIN-5-YL](PHENYL)METHANONE	cl18949; cl19137; cl15968; cl00858; cl00447; cl00474
DB08199	N-[(BENZYLOXY)CARBONYL]-L-CYSTEINYLGLYCINE	cl18949; cl19137; cl15968; cl00858; cl00447; cl00474
DB08271	N-ISOBUTYL-N-[4-METHOXYPHENYLSULFONYL]GLYCYL HYDROXAMIC ACID	cl18945; cl00261

DB08431	[(3R,4S)-4-HYDROXY-3-METHYL-2-OXOHEXYL]PHOSPHONIC ACID	cl00220
DB08471	1-(thiophen-2-ylacetyl)-4-(3-thiophen-2-yl-1,2,4-oxadiazol-5-yl)piperidine	cl00220
DB08567	(1S,4S)-4-(3,4-dichlorophenyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine	cl12078
DB08574	(5R)-2-SULFANYL-5-[4-(TRIFLUOROMETHYL)BENZYL]-1,3-THIAZOL-4-ONE	cl12078
DB08576	1-(5-TERT-BUTYL-1,3,4-OXADIAZOL-2-YL)-2-(METHYLAMINO)ETHANONE	cl19134
DB08632	1,3,5-BENZENETRICARBOXYLIC ACID	cl17173
DB08823	Spinosad	cl17255; cl17182
DB08846	Ellagic Acid	cl18408; cl16912; cl02872; cl09931; cl00470
DB08985	Etilefrine	cl12078
DB09008	Cephaloridine	cl12078
DB09042	Tedizolid Phosphate	cl19134; cl17173
DB09101	Elvitegravir	cl18949; cl19137; cl15968; cl00858; cl00447; cl00474

A structure data format (sdf) file which provides more detailed structural information about these compounds has been provided as a digital complement to this table.