Biosynthetic Pathway Determination for Intermediary Enriched Fragments between Plant Products and Drugs

By

Sunandini Sharma 121513

Project Report submitted in partial fulfillment of the requirement

For the degree of

Bachelor of Technology

To

Department of Biotechnology and Bioinformatics

Jaypee University of Information and Technology

Waknaghat, Himachal Pradesh

Under supervision of

Dr. Chittaranjan Rout

Jaypee University of Information and Technology

Waknaghat, Himachal Pradesh

and

Dr. Dhundy kiran Bastola

Associate Professor

University of Nebraska Omaha

Certificate

This is to certify that the work titled —" **Biosynthetic Pathway Determination for Intermediary Enriched Fragments between Plant Products and Drugs**", submitted by Sunandini Sharma (121513) in partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics to Jaypee University of Information Technology,Waknaghat, Himachal Pradesh has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor .....

Name of Supervisor Dr. Chittaranjan Rout

Designation Associate Professor Associate Professor

Date .....

Acknowledgement

As I conclude my project with the God's grace, I have many people to thank; for all the help, guidance and support they lent me, throughout the course of my endeavor. First and foremost, I thank Dr. Dhundy Kiran Bastola, my Project Guide, who has always encouraged me to put in my best efforts and deliver a quality and professional output. His methodology of making the system strong from inside has taught me that output is not the END of project. He has always inspired me to look at the positive aspects of research. I really thank him for his time, guidance and patience.

Secondly, I thank Kritika Karri, Graduate Student at University of Nebraska at Omaha for her patient hearing of my ideas and opening up my minds to newer horizons by pointing out my flaws, providing critical comments and suggestions to improve the quality of my work and appreciating my efforts.

I would also like to appreciate and thank the Dr. RS Chauhan for granting me the permission for credit transfer. Without his blessings, I would not have been able to discover myself as a person professionally growing day by day.

Also, I would like to thank Dr. Chitranjan Rout for teaching the course Drug designing and development, this course has been the motivation of my project. I found myself familiar with all the concepts of fragment based drug designing.

Apart from these, countless events, countless people & several incidents have made a contribution to this project that is indescribable.

My special thanks to all the people in the Bioinformatics Lab at University of Nebraska at Omaha.

CONTENTS

Abstract

Fragment based approaches have been proved to be efficacious in the drug discovery process. The fragmentation of drug leads into smaller pieces, or even into discrete functional can be useful to simplify the computational analysis of ligand binding and to map out different pharmacophoric elements required for high-affinity binding [1]. An ideal fragment is chemically diverse, structurally less complex; possess aqueous solubility and high availability [2]. These fragments can then be elaborated, combined with other molecules to provide novel drug leads. Natural products are the most popular source of drug leads. Exploiting the natural sources has now become the greatest interest of pharmaceutical industries for discovery of new leads. Genetic manipulations of micro-organisms or metabolic engineering have now been seen as a striking innovation in the natural product drug discovery process pathways leading to efficient production of drugs and drug precursors [3]. Alteration of natural pathways in organism leads to increased titer and production in the most inexpensive way [4]. In this study, we are investigating the fragments that are significantly enriched between plant products and known drug compounds and identify the biosynthesis pathway with the enriched fragment as an intermediate. A pipeline has been developed to engineer the enriched fragments in living organisms such as E.coli, Arabidopsis thaliana, Pea sativum etc. Once the fragments from the natural sources have been obtained, further scope of this research is to find all possible pathways which yield compounds with therapeutic activities (drugs) using the enriched fragment as the starting material in microorganisms.

Chapter 1

Introduction

The emergence of natural products has occurred due to the failure of the alternative drug compounds involved in metabolic diseases. As natural products, contain a large number of bioactive structures compared to synthetic compounds, using this as an advantage the pharmaceutical companies are making huge investments in natural compound drug discovery [5]. Over the years, attempts have been made to obtain drugs from natural sources because synthetically processed drugs are nonspecific to their targets. Synthetic drugs are susceptible to cause toxicity and side-effects in the body. It is known from the literature that natural compounds undergo positive selection to possess relevant biological function such as advantageous (ADME) absorption, distribution, metabolism and excretion properties (Atanasov, 2015) and are easily available. As a result, the pharmaceutical industry is now focused towards developing new plant- or bio-based drugs. Although, there has been a lot of struggle to extract compounds entirely from natural sources because of minuscule yield and expensive intermediate processing. Most pharmaceutical companies follow HTP screening of combinatorial libraries, but the screened products undergo severe disadvantage of poor interactions with the macro molecule.

The solution to this problem is “fragment based approach” meaning drugs are made by stepwise addition of building blocks (fragments) rather than total synthesis. Such fragments have sterically stable side groups which ensure favorable ligand-protein interactions [6]. The other advantage is the synthesis of partial analogues which may be therapeutic in nature. Further, cloning techniques may be used to produce the desired amount of the products in microbial or plant systems. The aim of this research is to extract only the biosynthetic intermediate or a bioactive fragment instead of harvesting the entire lead compound. The desired fragment is produced by engineering the plant or microbial cells (Benner and Sismour, 2005). Once the fragments are synthesized in a biological system, the further objective is to determine the pathway for the production of a drug with the fragment as an intermediate reactant. The production of drugs from natural sources has been difficult due to miniscule yield and tedious extraction. With the help of synthetic biology and systems biology, the pathways within the micro-organisms and plants can be easily modified to produce drugs in the most natural way. Later, the specific genes can be tweaked and cloned to produce products in huge quantity.

Interestingly, there are several pieces of evidence of therapeutic drug fragments in plants with anti-fertility properties, which are used as intermediate in the synthesis of contraceptive drugs from the natural source. Iridoid is an example of such fragment, which is found in Clerodendrum bucchanii and it is known to have anti-fertility property [7]. The history of Morphine and Quinine reveals that the constituents of

medicinal plants can be extracted. There are various advantages of using fragment based approach for drug discovery. First, fragments from plant sources reduce down the risk of severe toxicity and side effects in the body. The likelihood of unstable steric groups which produce unfavorable interaction is reduced. Fragment based approach allows the identification of small-molecule fragments that bind to specific regions of a protein target. These fragments can then be elaborated, combined with other molecules, or combined with one another to provide high-affinity drug leads [8]. Natural products with therapeutic properties are an important source of drugs or medicines. Several important drugs such as Taxol, camptothecin, morphine and quinine have been isolated from plant sources [9]. Though a large number of anti- Cancer drugs and chemotherapies are available, there is still a need of drugs that are less toxic, more selective. The aim of the research is to find new drugs from plant sources and alter the nature of the drug so that the independent fragment that has novel therapeutic activity can be biosynthesized. Alteration in the nature of the drug administered will ensure that drug acts only on the target cells and manifest fewer side effects. Clinical, pharmacological, and chemical studies of these traditional medicines, which were derived predominantly from plants, were the basis of most early medicines such as aspirin, digitoxin, morphine, quinine, and pilocarpine [10].

An enrichment of Drugs and natural products is a source of variety of substructures for the design of novel bioactive molecules. With the help of Chemo informatics parameters, common fragments between drugs and natural compounds are obtained. According to Peter et.al (2008) NPs are again the center of attention of the pharmaceutical industry as a promising and reliable source of new bioactive molecules. Another area where the availability of new effective drugs is becoming a pressing need is the treatment of infectious diseases, as antibiotic-resistant bacteria are becoming more common and widespread and are a cause for serious concern.

The natural product derived structure plays a significant role in the discovery of novel pharmaceutical agents and/or bioactive molecules. The anti-diabetic activity in lupins has been attributed to quinozolidine alkaloids and a review of the literature shows many such examples of natural products as sources of new drugs including Paclitaxel, which is one of the most widely prescribed anticancer drugs on the market. Cinchon (a flowering plant) contains 25 closely related alkaloids , of which the most important are quinine , quinidine , cinchonine, and cinchonidine [11]. Most of the natural products are biologically active and have favorable absorption, distribution, metabolism, excretion and toxicology properties. Plants are often the predominant source for the discovery of natural products due to the relative ease of access. However, more recently microbial as well as marine sources have been identified as alternative resources, particularly for antibiotics. Several databases of natural products have been published and reviewed.

Pharmaceutically relevant natural products are of low molecular weight and often restricted to special plant families. While these compounds are not important for the primary metabolism of the plants, they are of great importance for their survival in a given environment. Therefore, medicinally important plants are often collected from the wild or their natural habitat and are more likely to be endangered due to severe over collection. Unfortunately, we still have limited knowledge about plant secondary metabolism, its regulation, molecular mechanisms concerning gene expression and rate-limiting enzymes found within a diverse network of biosynthetic pathways in living organisms.

Another application of Fragment based natural compound drug discovery is the use of low molecular weight probes (Fragments) for screening against array of biological targets. The small size enhances the complementary binding between fragment and targets [12].

Most plants contain metabolites with various biological properties. However, these days most of the chemotherapeutics used in clinical settings are produced through in vitro synthesis. With few exceptions, like vincristine and taxol that are difficult to synthesize in vitro because of complex metabolite structures [13]. Many synthetic drugs cause several side effects that were not acceptable except as treatments of last resort for terminal diseases such as cancer. The presence of metabolites in natural products and medicinal plants avoid side effects over synthetic drugs. It occurs because the metabolites are naturally occurring in plant system and may be used for growth and metabolism. Obtaining a drug completely from a plant source may be difficult as yield from natural source may be small or extraction process may require some chemicals. The solution to these problems is Partial chemical Synthesis or Semi synthesis. In this process, instead of harvesting the lead compound, the aim is to extract only a biosynthetic intermediate or a bioactive fragment [14]. From the intermediate fragment the compound can be prepared by conventional synthesis. This approach has two advantages. First, the intermediate may be more easily extracted in higher yield than the final product itself. Second, it may allow the possibility of synthesizing analogues of the final product. The production of penicillin is an example of this approach. According to M Lohlou (2013) another example is that of Paclitaxel which was manufactured by extracting 10-deacetylbaccatin from the needles of the yew, then carrying out a four-stage synthesis [15].

The growing trend of drug discovery from natural sources along with improved technological inputs promises better returns to the Pharmacy industry. There has been a history of beneficial uses of medicinal plants. Drugs derived from natural sources also served as drug leads suitable for optimization by synthetic means. For certain therapy areas, such as antimicrobials, anticancer antihypertensive and anti-inflammatory drugs, the number was even higher the most of all approved small molecule new chemical entities were derived from nature. Natural products not only complement synthetic molecules, they also exhibit drug

relevant features unsurpassable by any synthetic compound. One key feature of natural products is their enormous structural and chemical diversity. In fact, about 40% of the chemical scaffolds found in natural products are absent in today's medicinal chemistry, and therefore complementary to synthetically produced molecules [16]. Most possibly this is one of the reasons for their historical success in drug discovery, with 45% of today's best selling drugs originating from natural products or their derivatives.

Biosynthesis of natural products involves repeated interaction with modulating enzymes, and the actual biological function of many natural products comprises binding to other proteins. Thus, the ability of natural products to interact with other molecules, an indispensable prerequisite to making an effective drug, might be considered as biologically validated. It is an unsurprising, but often overlooked, fact that many natural products exhibit advanced binding characteristics compared with synthetics. Most probably, the sterically more complex structure of natural products contributes to this. The success of natural products is related to the forces of natural products chemistry, molecular and cellular biology, synthetic and analytical chemistry, biochemistry, and pharmacology to exploit the vast diversity of chemical structures and biological activities of these products. Moreover, the exploration of structural chemical databases comprising a wide variety of chemotypes, in conjunction with databases on target genes and proteins, will facilitate the creation of new chemical entities through computational molecular modeling for pharmacological evaluation. The enrichment process extracts few fragments that can act as potential scaffolds. The computational pipeline gives enriched fragments between plant and drug compounds [17].

Limited knowledge is available about the pharmaceutical value of these enriched fragments as an intermediate compound. Therefore, our present work encompasses investigation of these fragments and determines the biosynthetic routes that exist naturally and have been the main target of engineering in the past few years. Currently, a number of computational tools have been published to aid in the pathway reconstruction from source to query. One such algorithm is Retropath [18] which gives the most promising route in terms of metabolite exchange, maximum allowable pathway yield, toxicity and enzyme efficiency. It is linked to KEGG database [19]. Another tool that was used for the analysis was Pathpred [20], a web-based server to predict plausible pathways of multi-step reactions starting from a query compound, which focus on predicting pathways for microbial biodegradation and biosynthesis of plant secondary metabolites.

Chapter 2

Background Pipeline and Tools

MolBlocks: A suite of programs for breaking down sets of small molecules into fragments according to a predefined set of chemical rules, clustering the resulting fragments, and uncovering statistically enriched fragments.

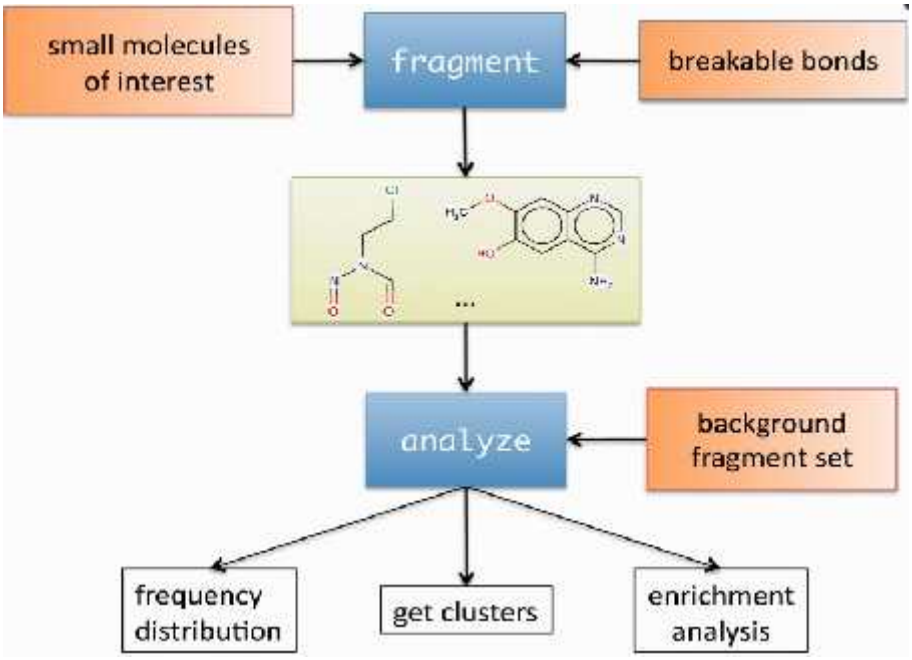

Fig. 1: The fragment program takes as input a set of small molecules and user-defined rules that specify the bonds to break, and then applies these rules to fragment the molecules. As an optional second step, carried out by the analyze program, the user can cluster the fragments and/or determine whether the frequency of any of the fragments is enriched as compared with a background set of fragments [21].

Enrichment Pipeline:

The final step of the pipeline involves the comparison between enriched fragments from the drug dataset against fragments obtained from the natural compounds set. In order to calculate the pairwise similarity between each of the enriched drug fragments and each of the fragments from natural compounds we used the Tanimoto equation. To carry out the calculations we wrote an in-house program that uses the Python API [23] of the OpenBabel library, and retained the drug fragment–natural product fragment pairs that had a Tanimoto similarity > 0.9.

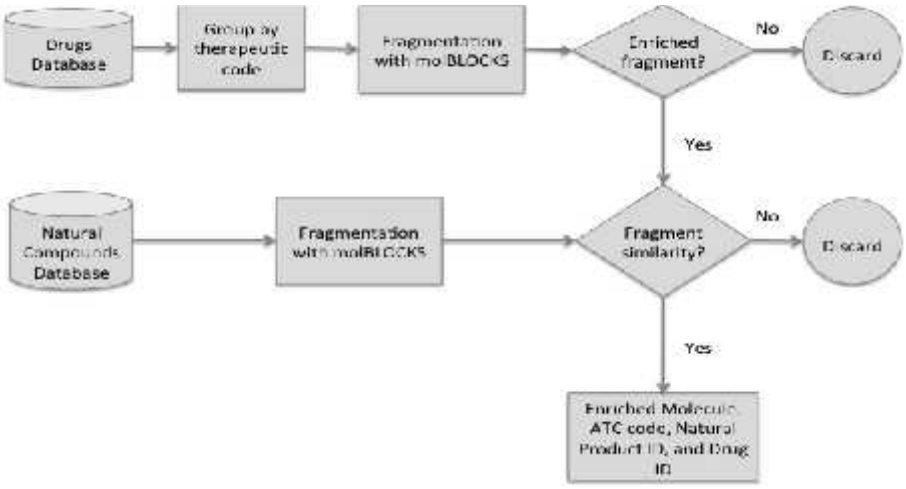

Fig. 2 : Simplified overview of the pipeline. Each approved drug (obtained from Drugbank) is assigned a therapeutic class using the ATC nomenclature. The drugs are then broken down into fragments using the molBLOCKS software, and enrichment analysis is performed on each therapeutic class to identify statistically overrepresented fragments (FDR < 0.05). Each overrepresented fragment is then compared against similarly obtained fragments from a database of natural compounds (SuperNatural II) [22].

Retropath:

A web-based pathway analysis platform available at http://xtms.issb.genopole.fr, which provides full access to the set of pathways that can be imported into a chassis organism such as *Escherichia coli* through the application of an Extended Metabolic Space modeling framework. The XTMS approach consists on determining the set of biochemical transformations that can potentially be processed *in vivo* as modeled by molecular signatures, a specific coding system for derivation of reaction rules for metabolic reactions and enumeration of all the corresponding substrates and products. Most promising routes are described in terms of metabolite exchange, maximum allowable pathway yield, toxicity and enzyme efficiency. By answering such critical design points, XTMS not only paves the road toward the rationalization of metabolic engineering, but also opens new processing possibilities for non-natural metabolites and novel enzymatic transformations. The following figure demonstrates the full working of the tool. [23]

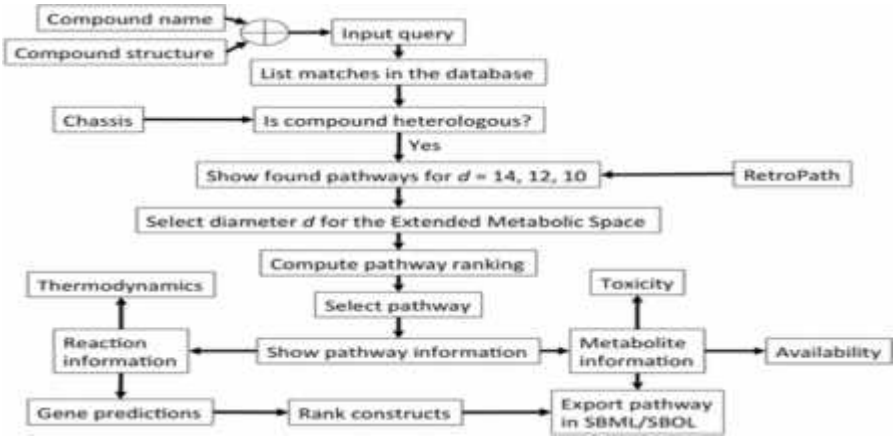

Fig. 3. Schematic representation of the query process in the XTMS server. After the user inputs a query compound, pathways for heterologous compounds that match the query are retrieved from the Extended Metabolic Space by RetroPath. After selecting a desired diameter, pathways and constructs are ranked and information about reactions and metabolites is provided, with the possibility of downloading the desired construct in SBML/SBOL format.

Pathpred: A web-based server to predict plausible pathways of muti-step reactions starting from a query compound, based on the local RDM pattern match and the global chemical structure alignment against the reactant pair library. In this server, we focus on predicting pathways for microbial biodegradation of environmental compounds and biosynthesis of plant secondary metabolites, which correspond to characteristic RDM patterns in 947 and 1397 reactant pairs, respectively. The server provides transformed compounds and reference transformation patterns in each predicted reaction, and displays all predicted multi-step reaction pathways in a tree-shaped graph.

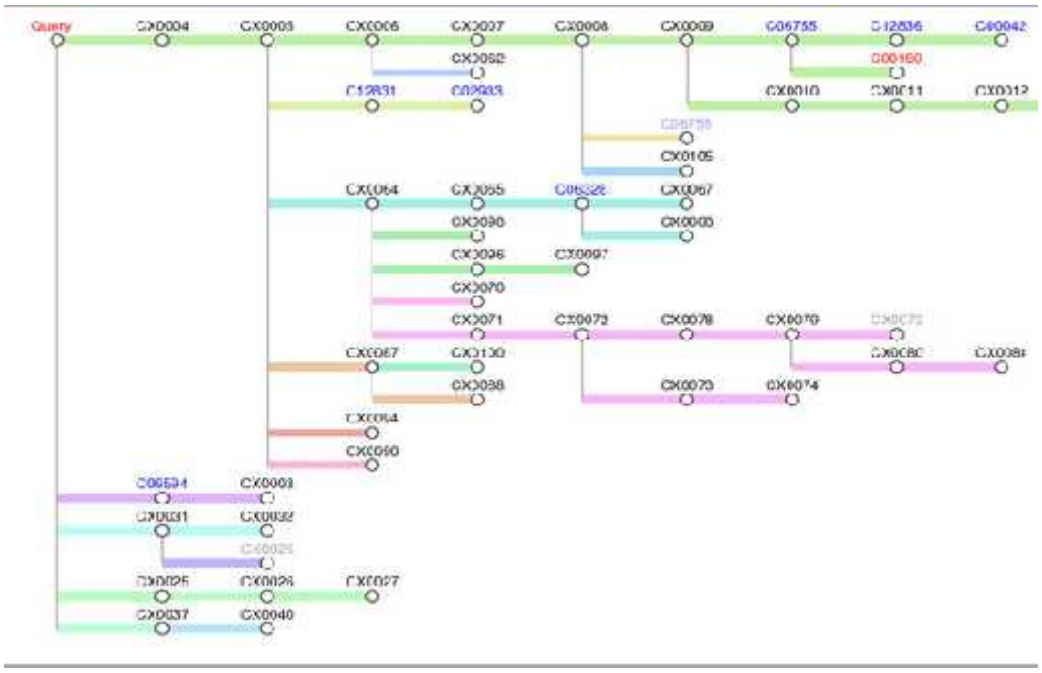

Fig. 4 : The predicted pathway tree consists compounds (node) and reactions (edge). The blue-colored C numbers are compounds in the KEGG pathway and the black-colored CX numbers are compounds not stored in KEGG. A greyed-out node means that there is the same compound elsewhere in the tree. When the pathway reaches to the final compound in bi-directional prediction, the compound is highlighted in red. The consecutive reactions in same color show that the reference reactions are consecutive in KEGG pathway. The thickness of the line shows the plausibility of the predicted reaction based on the reaction score.

Chapter 3

Methodology

To systematically compare functionally relevant drug fragments and natural products, a computational pipeline described previously by our group was used. The pipeline compared approved drugs obtained from the DrugBank database against a large collection of natural products, assembled in the SuperNatural II database. The statistically enriched fragments, associated with an anatomic therapeutic group known as ATC code in DrugBank, were compared against the fragments derived from the natural compounds. The fragmentation process for both drug and natural compounds was performed using molBLOCKS suite. The fragment information consisted of the SMILE code, SuperNatural Compound ID and DrugBank ID with ATC code which was used further for downstream analysis.

A. Obtaining and representing drugs and natural products

The DrugBank database (version 4.1) was used to obtain information on drugs that were approved for therapeutic use in at least one country. The initial set of drugs contained 1,554 molecules. Natural products were obtained from the SuperNatural II database , containing 325,508 molecules. Drugs and natural products were represented using the SMILES system, a widely used notation that makes it possible to encode chemicals as ASCII strings. SMILES strings for drugs and natural products were directly obtained from the
DrugBank and SuperNatural II databases, respectively.

B. Fragmenting the molecules

Both drugs and natural products were fragmented with the fragment program, part of the molBLOCKS suite, which breaks molecules along chemically important bonds and returns the corresponding fragments (or putative building blocks). The list of chemical bonds that were used by the program to fragment the molecules, and is based on Lewell et al. The minimum size for a fragment was set to four atoms, and the fragmentation was carried out with the “extensive” flag turned on, which yields all possible fragments that can be generated given the list of chemical bonds of interest . It is noteworthy to mention that the fragmentation rules are encoded as SMARTS (SMiles ARbitrary Target Specification), an extension to the SMILES notation created by Daylight Chemical Information System, Inc. and widely used chemistry. Using SMARTS patterns the particular bonds that are to be cleaved are encoded as regular expressions, making it straightforward to add other cleavable bonds to the fragmentation rules.

C. Clustering fragments

Drug fragments obtained as described above were clustered with the analyze program using standard parameters. In order to compute the fragment similarity for clustering, the program converts the fragment to a fingerprint representation, based on linear segments of up to 7 atoms in length (FP2 fingerprints). The fingerprints are stored as bit arrays, where the presence or absence of a particular linear segment is represented by a 1 or 0, respectively. The FP2 fingerprint representation is obtained via the Open Babel library1. Then, the Tanimoto coefficient Ts between two fragments x and y is computed as:

where X and Y are the bit array representations of the linear segments found in fragment x and y, respectively, and ^ and_ are the bitwise and and or operators. The analyze program computes pairwise similarities between fragments and converts them to a graph representation, where an edge between fragments indicates a pairwise Tanimoto greater than the chosen threshold, which was set to 0.7 in this study. Subsequently, the program extracts the connected components of the graph, and selects the representative element for each cluster as the fragment with the highest average similarity against all the other fragments in the cluster.

D. Extracting enriched fragments for each ATC code

In order to assign functional categories to drugs, we used the Anatomical Therapeutic Chemical (ATC) classification system2, a widely used nomenclature that organizes drugs according to the organ or system which they modulate and their therapeutic properties. The ATC code system is hierarchically organized into five levels of increasing specificity. We considered the second level, which describes the therapeutic main groups. We note that a single drug can be annotated with multiple ATC codes, if it has multiple therapeutic indications. For this study, to get meaningful statistics we selected all the ATC codes that annotated at least 10 distinct drugs. Enrichment analysis was carried out in order to identify the specific fragments (or clusters of fragments) that appear in a set of molecules more frequently than expected by chance, given a background distribution. In this study the background was represented by the union of all approved drugs.

The analyze program uses the hypergeometric distribution to model the probability of obtaining a number of fragments (or clusters of fragments) equal to or greater than the observed by chance alone where N is

the total number of fragments; K is number of fragments of the given type; n is the total number of fragments in the main set; and x is the total number of fragments of thegiven type in the main set.
The program returns both uncorrected p-values and False Discovery Rate (FDR) corrected p-values, obtained with the procedure of Benjamini-Hochberg. In this study we selected fragments that were enriched with an FDR < 0:05.

E. Comparing enriched fragments in the drug dataset against fragments from natural compounds.

The final step of the pipeline involves the comparison between enriched fragments from the drug dataset against fragments obtained from the natural compounds set. In order to calculate the pair-wise similarity between each of the enriched drug fragments and each of the fragments from natural compounds we used the Tanimoto coefficient. To carry out the calculations we wrote an in-house program that uses the Python API [23] of the OpenBabel library , and retained the drug fragment–natural product fragment pairs that had a Tanimoto similarity > 0:9.

F.Computational requirements.

The most time-consuming step of the pipeline is represented by the pairwise fragment comparison, which took approximately 12 hours on a 24-core machine. Fragmentation of the 325,509 molecules found in the SuperNatural II database took approximately eight hours on a 24-core machine, bringing
the entire analysis to roughly 20 hours.

G. Data Processing

We used the online SMILE converter program (https://cactus.nci.nih.gov/ translate/) to convert all the chemical structures from SMILE format to MDL mol structural files, to be used later as an input for the pathway prediction algorithms. Chemspider (http://www.chemspider.com/) , a free chemical structure database, was used to retrieve the IUPAC name and chemical information for the enriched fragments.

H. Biosynthesis Pathway Annotation

Our initial goal in the pathway annotation was to determine the natural source and biosynthesis pathway for the enriched fragments obtained from the enrichment pipeline. We used Retropath webserver (http://www.issb.genopole. fr/~faulon/retropath.php) where each enriched fragment was given as an input query in MDL mol structural format and the resulting output was a feasible biosynthesis pathway from the natural source, including the names of the enzymes catalyzing the reactions (Fig 1.).

The next step of pathway annotation was to determine the synthesis pathway from fragments to drug compound. We required an algorithm which could predict the synthesis pathway given the starting and final product. Pathpred server (http://www.genome.jp/tools/pathpred/) was suitable for performing this task. The enriched fragments, with known biosynthesis pathway from Retropath, was given as an initial substrate. The drug compound associated with the enriched fragment, obtained from the enrichment pipeline were given as the final product in order to obtain the probable synthesis routes between the fragment and drug. This server is linked with KEGG databased and the user can input a query compound in the MDL mol file format, in the SMILES representation, or by the KEGG compound/drug identifier.

Chapter 4

Results and Discussions

The data for 112 enriched fragments were obtained from the fragment based enrichment pipeline. From which we were able to retrieve the pathway information for 8 fragments with biosynthesis route from plant source and 1 fragment with a fungi source, using Retropath server. The fragment information with the enzymes catalysing the reaction.

Enriched fragment	**E.C number**	**IUPAC name of the fragment**	**Plant Source**
CCCc1ccc(cc1)O	1.14.13.11	Trans-cinnamate 4-monooxygenase	Brasica - napus(Rapeseed)
CCCCCCCCCCCC	1.11.1.3,4.1.99.5	hydrogen-peroxide oxidoreductase,aldehyde oxygenase	Pea sativums
CCCCCCCCC(=O)O	1.11.1.3	hydrogen-peroxide oxidoreductase	Pea sativums
O=C[C@H](CC(C)C)N	1.4.3.21	primary-amine oxidase	Arabidopsis thaliana
O=c1[nH]c(=O)c2c(n1C)nc[nH]2	2.1.1.160	caffeine synthase	Camellia irrawadiensis
OC(=O)[C@H](Cc1c[nH]c2c1cccc2)N	1.14.13.125	tryptophan N-monooxygenase	Sinabis alba (white mustard)
Cc1c[nH]c(=O)[nH]c1=O	1.14.11.6	thymine dioxygenase	Rhototorula glutinis
O=C1[C@@H](N)[C@@H]2N1C(=C(CS2)C)C(=O)O	6.3.2.26 -	N-(5-amino-5-carboxypentanoyl)-L-cysteinyl-D-valine synthase	Acremonium chrysogenum
C/C=C/c1ccccc1	4.1.1.-	lysine/ornithine carboxy-lyase	Sophora lavascence

These 9 fragments were then provided to the Pathpred server with final drug product to find out the biosynthesis pathway between fragment and drug. We were successfully able to retrieve information for 5 fragments which we will discuss in detail

Caffeine Fragment Biosynthesis

One of the enriched fragment from our results from Retropath was 1,3,7-trimethylxanthine commonly known as caffeine which is abundantly found in coffee plants (Coffea arabica) and young leaves of tea plant (Camellia sinesis). Caffeine synthesis process begins in these plants with Xanthosine as the initial substrate which is then converted into 7-methylxanthosine which is followed by second methylation step which leads to the formation of theobromine. The final product caffeine is synthesized by the enzyme caffeine synthase which converts 7-methylxanthine to theobromine and theobromine to caffeine. (Fig 5.).According to the enrichment pipeline, this fragment is significantly enriched in theophylline (DB00277) drug which is used in the treatment of asthma and bronchodilation. Pathpred provided the synthesis route between caffeine fragment and theophylline. Additionally, it also provided the synthesis pathway for theophylline derivatives - 1-methylruic acid and xanthine (Fig 5.)

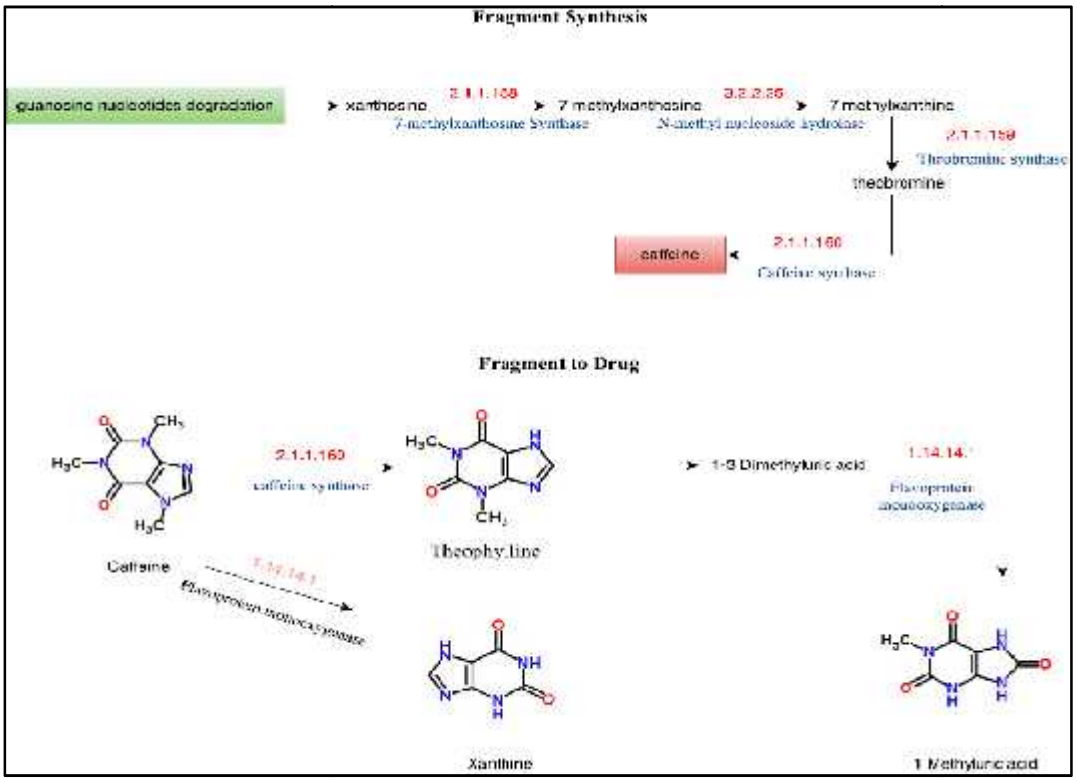

Fig. 5 . Caffeine biosynthesis pathway from plant source followed by synthesis of drugs- theophylline, xanthine and 1-methylruic acid from Caffeine fragment

Pentanal biosynthesis from 5-aminopentanal fragment

5- aminopentanal fragment has its synthesis source in algae and land plants. Primary-amine oxidase (EC 1.4.3.21) catalyzes the substrate cadaverine which leads to the production of the fragment 5-aminopentanal.This fragment is an intermediate substrate for the synthesis of pentanal (DB01919), a drug that interacts with cAMP-dependent protein kinase catalytic subunit alpha and is involved with phosphorylation in the cytoplasm. The synthesis of drug from the enriched fragment is catalysed by leucine dehydrogenase(EC 1.4.1.9) and branched chain amino acid transferase (EC 2.6.1.42) (Fig. 6.)

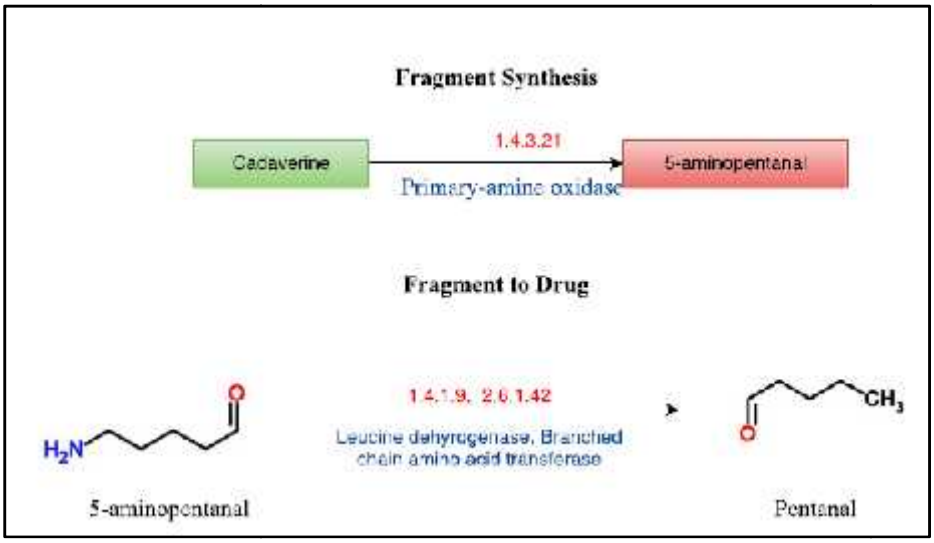

Fig. 6. The Pentanal biosynthesis pathway from fragment 5-aminopentanal

Glycine Biosynthesis Pathway

Glycine was found to be significantly enriched between compounds from the DrugBank and supernatural database. Interestingly, Glycine (DB00145) is a drug in itself with its known ability in anti-spastic and anti-inflammatory activities. The biosynthesis pathway for Glycine exists in plant Arabidopsis *thaliana* with enzyme serine Hydroxymethyl transferase (EC 2.1.2.1) which catalyses the initial substrate l-serine into glycine. Alternative synthesis pathways for Glycine also exists in other natural sources like bacteria (E coli.) and yeast (Saccharomyces *cerevisiae*).

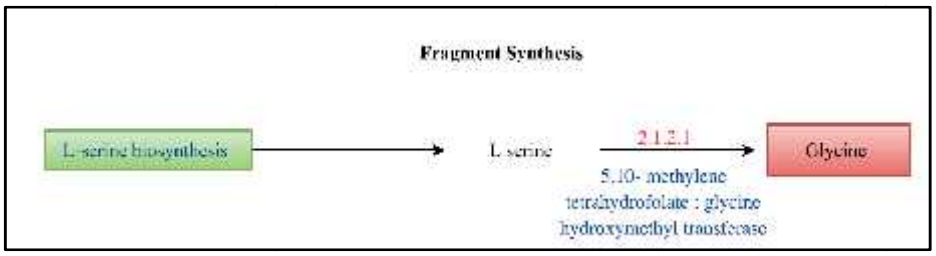

Fig. 7. Glycine biosynthesis from Arabidopsis thaliana plant source

Fragment Biosynthesis from Non-Plant Sources

For two enriched fragments we could not find the biosynthesis pathway in plants but they had their source from other natural source. Styrene fragment was found to have a biosynthesis pathway in Sacchromyces Cerevisiae. The enzyme Ferulic acid decarboxylase (EC Number: 4.1.1.M2) catalyses the substrate trans-cinnamate to produce styrenene. This fragment results in the synthesis of drugs: Eugenol (DB09086) and Coumarin (DB04665) with the catalyzing enzymes, p-Coumarate 3 monooxygenase (EC: 1.14.18.1) acts on p-anol and Caffeic acid 3-O-methyltransferase (EC: 2.1.1.68) acts on 4-[(1E)-Propen-1-yl]-1,2-benzenediol (Fig 8.) .Eugenol has an anesthetic property and is used as a sedative. It also prevents oxidative changes in membrane and acts as an antioxidant.While coumarin is a vasoactive agent and used in antivaricose therapy(a condition of abnormally enlarged dilated veins). It also has anti-fungicidal and anti-tumor activities.

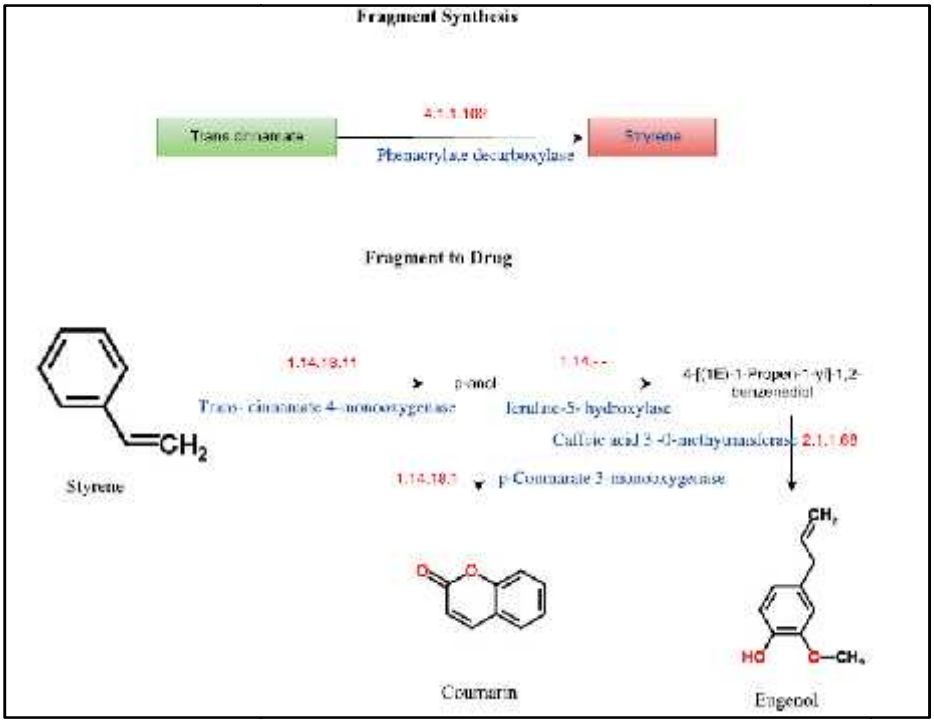

Fig. 8. Production of styrene fragment followed by synthesis of coumarin and eugenol

Fragment Biosynthesis from Non-Plant Sources

Thymine fragment can be biosynthesised from E coli where thymidine is the initial substrate and the enzymes catalysing the reaction are pyrimidine nucleoside phosphorylase (EC 2.4.2.2) and thymidine phosphorylase (EC 2.2.2.4) (Fig. 9.). Thymine fragment can be used to synthesize drug

hydroxytryptophan commonly known as Oxitriptan (DB02959) which is an antidepressant,appetite suppressant, and sleep aid. It is a one step reaction catalysed by enzyme tyrosinase (EC 1.14.18.1).

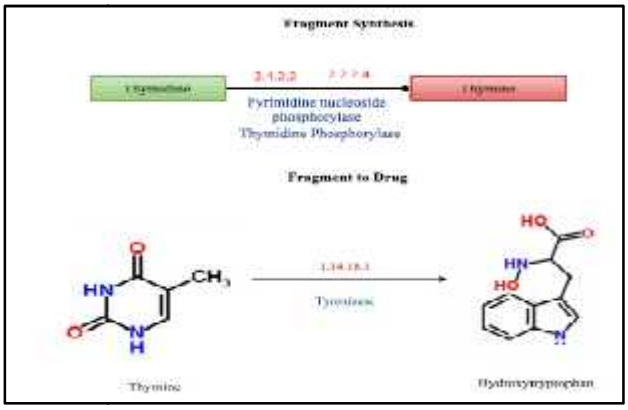

Fig. 9. Thymine biosynthesis pathway and synthesis of hydroxytryptophan drug

Retrieval Data

Fragment (Smiles)	No. of Drugs in which the Fragment is enrichment	No. of Drugs which are biosynthesized from the fragment	% of Retrieval
O=c1[nH]c(=O)c2c(n1C)nc[nH]2	11	3	27.27
O=C[C@H](CC(C)C)N	4	1	25
NCC(=O)O	3	1	33.33
C/C=C/c1ccccc1	2	2	100
Cc1c[nH]c(=O)[nH]c1=O	3	1	33.33

Fig. 10. The above table shows the quantitative analysis of the drugs obtained from the fragments. The above five fragments are shown to give rise to drugs which are registered in DrugBank and are categorized as small molecule drugs.

Conclusion

Natural products not only complement synthetic molecules, they also exhibit drug relevant features unsurpassable by any synthetic compound which is probably one of the reasons why metabolic engineering in plants sources has been the basis of modern medicine and is seen to pave the way for future of pharmaceutical industry. The possibility of using an enriched fragment as an intermediate to improve the drug yield have opened exciting perspectives for the exploitation of biosynthetic capacity of plants cells. Fragment based enrichment methods can be significant for metabolic engineering especially for identification of novel drug leads from the highly diverse plant sources. Our current results validate the utilitarian significance of fragment enrichment across drug and natural compounds in two ways: First, it can be used to identify the biosynthesis pathway which can be used alternatively for the production of a drug compounds instead of conventional methods used in synthetic chemistry. Second, an enriched fragment can predict the ability of a natural source to be a potential therapeutic agent. Hence, providing a rationale for the identification of novel drug leads with biosynthetic ability.

References

[1] Hajduk, P. J., & Greer, J. (2007). A decade of fragment-based drug design: Strategic advances and lessons learned. *Nature Reviews Drug Discovery Nat Rev Drug Discov, 6*(3), 211-219. doi:10.1038/nrd2220

[2] Hajduk, P. J., & Greer, J. (2007). A decade of fragment-based drug design: Strategic advances and lessons learned. *Nature Reviews Drug Discovery Nat Rev Drug Discov, 6*(3), 211-219.

[3] Gerwick, W., & Moore, B. (2012). Lessons from the Past and Charting the Future of Marine Natural Products Drug Discovery and Chemical Biology. *Chemistry & Biology, 19*(12), 1631. doi:10.1016/j.chembiol.2012.12.004

[4] Pickens, L. B., Tang, Y., & Chooi, Y. (2011). Metabolic Engineering for the Production of Natural Products. *Annual Review of Chemical and Biomolecular Engineering Annu. Rev. Chem. Biomol. Eng., 2*(1), 211-236. doi:10.1146/annurev-chembioeng-061010-114209

[5] Beal, J. L. (1981). Natural products as medicinal agents: Plenary lectures of the International Research Congress on Medicinal Plant Research, Strasbourg, July 1980. Stuttgart: Hippokrates Verlag.

[6] Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., & Dobchev, D. A. (2010). Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chemical Reviews Chem. Rev., 110*(10), 5714-5789. doi:10.1021/cr900238d

[7] Amedei, A., & Niccolai, E. (2014). Plant and Marine Sources. Natural Products Analysis Instrumentation, Methods, and Applications, 43-113.

[8] Erlanson, D. A., Ballinger, M. D., & Wells, J. A. (2006). Tethering. *JAHNKE: FRAGMENT-BASED APPROACHES IN DRUG DISCOVERY O-BK Methods and Principles in Medicinal Chemistry Fragment-based Approaches in Drug Discovery*, 285-310. doi:10.1002/3527608761.ch14

[9] Balunas, M. J., & Kinghorn, A. D. (2005). Drug discovery from medicinal plants. *Life Sciences, 78*(5), 431-441. doi:10.1016/j.lfs.2005.09.012

[10] Cytotoxic natural products. (1997). *Drug Discovery Today, 2*(2), 79. doi:10.1016/s1359-6446(97)83294-1

[11] Daniel, M. (2006). *Medicinal plants: Chemistry and properties*. Enfield, NH: Science.

[12] Hoelder, S., Clarke, P. A., & Workman, P. (2012). Discovery of small molecule cancer drugs: Successes, challenges and opportunities. *Molecular Oncology, 6*(2), 155-176. doi:10.1016/j.molonc.2012.02.004

[13] Organizing Committee. (1977). Low Molecular Weight Sulphur Containing Natural Products, Ii.

[14] Discovery and resupply of pharmacologically active plant-derived natural products: A review. (2015). *Draw Science*. doi:10.18516/0001

[15] Drug Discovery from Natural Products. (2012). *RSC Drug Discovery*. doi:10.1039/9781849734950

[16] Müller-Kuhrt, L. (2003). Putting nature back into drug discovery. *Nat Biotech Nature Biotechnology, 21*(6), 602-602. doi:10.1038/nbt0603-602

[17] Balasubramanya, A., Thapa, I., Bastola, D., & Ghersi, D. (2015). A novel approach to identify shared fragments in drugs and natural products. *2015 IEEE International Conference on Bioinformatics and Biomedicine(BIBM)*.doi:10.1109/bibm.2015.7359746

[18] Carbonell, P., Planson, A., Fichera, D., & Faulon, J. (2011). A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Systems Biology BMC Syst Biol, 5(1), 122.

[19] Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27-30.

[20] Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., & Kanehisa, M. (2010). PathPred:An enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Research, 38 (Web Server).

[21] Ghersi, D., & Singh, M. (2014). MolBLOCKS: Decomposing small molecule sets and uncovering enriched fragments. *Bioinformatics, 30*(14), 2081-2083. doi:10.1093/bioinformatics/btu173

[22] Balasubramanya, A., Thapa, I., Bastola, D., & Ghersi, D. (2015). A novel approach to identify shared fragments in drugs and natural products. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi:10.1109/bibm.2015.7359746

[23] Carbonell, P., Parutto, P., Baudier, C., Junot, C., & Faulon, J. (2014). Retropath: Automated Pipeline for Embedded Metabolic Circuits. *ACS Synth. Biol. ACS Synthetic Biology, 3*(8), 565-577. doi:10.1021/sb4001273