IDENTIFYING NOVEL GENETIC INTERACTIONS THROUGH A STUDY OF SYNTHETIC LETHALS

Thesis submitted to the SASTRA Deemed to be University
in partial fulfillment of the requirements
for the award of the degree of

B. Tech. Biotechnology

Submitted by

LAKSHMI CHANEMOUGAM
(Reg. No.: 122010053)

July 2022

SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR | KUMBAKONAM | CHENNAI

SCHOOL OF CHEMICAL AND BIOTECHNOLOGY

THANJAVUR, TAMIL NADU, INDIA – 613 401

SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR | KUMBAKONAM | CHENNAI

SCHOOL OF CHEMICAL AND BIOTECHNOLOGY

THANJAVUR – 613 401

Bonafide Certificate

This is to certify that the thesis titled "**Identifying novel genetic interactions through a study of synthetic lethals**" was submitted in partial fulfilment of the requirements for the award of the degree of B. Tech. Biotechnology to the SASTRA Deemed to be University is a bonafide record of the work done by **Ms. Lakshmi Chanemougam** (Reg. No. 122010053) during the final semester of the academic year 2021-22, in the **School of Chemical and Biotechnology**, under my supervision. This thesis has not formed the basis for the award of any degree, diploma, associateship, fellowship or other similar titles to any candidate of any University.

Signature of Project Supervisor : Dr. Karthik Raman, Associate Professor, Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai - 600 036, India

Name with Affiliation : Dr Karthik Raman, Associate Professor, IIT-Madras

Date : 04.07.2022

Project *Viva voce* held on _____

Examiner 1 **Examiner 2**

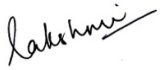
SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR | KUMBAKONAM | CHENNAI

SCHOOL OF CHEMICAL AND BIOTECHNOLOGY

THANJAVUR – 613 401

Declaration

I declare that the thesis titled "**Identifying novel genetic interactions through a study of synthetic lethals**" submitted by me is an original work done by me under the guidance of **Dr. Karthik Raman, Associate Professor, Department of Biotechnology, Indian Institute of Technology, Madras** during the final semester of the academic year 2021-22, in the **School of Chemical and Biotechnology**. The work is original, and wherever I have used materials from other sources, I have given due credit and cited them in the thesis text. This thesis has not formed the basis for the award of any degree, diploma, associate ship, fellowship or other similar titles to any candidate of any University.

Signature of the candidate(s) : 

Name of the candidate(s) : Lakshmi Chanemougam

Date : 02.07.2022

Certificate (For External Projects)

डा० कार्त्तिक रामन्
सह प्राध्यापक, भूपत और ज्योति मेहता
जैव विज्ञान विद्यालय, जैव प्रौद्योगिकी विभाग
समन्वयक, जैव तन्त्र अभियान्त्रिकी उपक्रम
सदस्य, रॉबर्ट बॉश डेटा विज्ञान एवं कृत्रिमप्रज्ञा केन्द्र
भारतीय प्रौद्योगिकी संस्थान मद्रास
चेन्नई – ६०० ०३६, भारत
✆ +९१-४४-२२५७-४९३९ 🖷 +९१-४४-२२५७-४९०२
ई-मेल / E-mail: kraman@iitm.ac.in

Dr. Karthik Raman
Associate Professor, Bhupat and Jyoti Mehta School of Biosciences, Department of Biotechnology
Co-ordinator, Initiative for Biological Systems Engineering (IBSE)
Member, Robert Bosch Centre for Data Science & Artificial Intelligence
Indian Institute of Technology Madras
Chennai - 600 036, INDIA
✆ +91-44-2257-4139 🖷 +91-44-2257-4102
http://biotech.iitm.ac.in/faculty/karthik-raman

OFFER LETTER

23rd February, 2022
Ms. Lakshmi Chanemougam
B. Tech Biotechnology (Register Number 122010053)
SASTRA University, Thanjavur.

Dear Lakshmi,

This is to confirm your position as project trainee at the Computational Systems Biology Lab at IIT Madras. Please find details of the training dates below:

Start Date: March 1st, 2022
End Date: July 1st, 2022

I look forward to welcoming you to the lab!

Regards,

Dr. Karthik Raman
Associate Professor
Department of Biotechnology
Bhupat & Jyoti Mehta School of Biosciences
Indian Institute of Technology Madras
Chennai - 600 036, India

Karthik Raman

Acknowledgements

I sincerely thank my project guide Dr Karthik Raman, Associate Professor at the Department of Biotechnology at IIT-Madras and the lab members of Computational Systems Biology Lab at IIT Madras for their invaluable support, guidances and resources offered. I also extend my gratitude to senior faculty and deans at the School of Chemical and Biotechnology, SASTRA Deemed to be University, for the exposure and knowledge I gained through this project opportunity. Furthermore, I would like to thank Mr Senthamizhan V, Project Associate at the Initiative for Biological Systems Engineering (IBSE) at IIT-Madras, for their thorough overseeing of the project and for imparting useful insights throughout the course of the project; Ms. Indhumathi P, PhD Student under Dr Karthik Raman, associated with the Computational Systems Biology Lab at IIT-Madras, for their patience in resolving many code related issues and for imparting knowledge on all the tools used in this project.

Table of Contents

List of Figures

List of Tables

Abbreviations

DGD	Double Lethal Genes
FBA	Flux Balance Analysis
GO	Gene Ontology
GPR	Gene Protein Reaction Associations
JDL	Double Lethal Reactions
JSL	Single Lethal Reactions
JTL	Triple Lethal Reactions
LP	Linear Programming
MILP	Mixed Integer Linear Programming
SGD	Single Lethal Genes
SL	Synthetic Lethals
TGD	Triple Lethal Genes
VMH	Virtual Metabolic Human
WTA	Wall Teichoic Acids

Notations

$\forall$	For all
$\in$	Belongs to
$\sum$	Summation
LB_j	Lower bound of the j^{th} reaction
S	Stoichiometric matrix
s_{ij}	Element in stoichiometric matrix
UB_j	Upper bound of the j^{th} reaction
v_{bio}	Maximum flux through the biomass reaction
v_j	Flux through the j^{th} reaction

Abstract

This thesis focuses on uncovering novel interactions amongst clinically relevant pathogens, using network based studies conducted on synthetic lethals. Synthetic lethals are pairs of non-essential genes which cause the death of the organism upon simultaneous inactivation/deletion when exposed to sub minimal nutrient levels (Guarente, 1993). The study utilizes an efficient algorithm named Fast-SL (Pratapa, Balachandran and Raman, 2015), that computes lethal gene and reaction until third order. With the rise in multidrug resistance and mechanisms of adaptations, synthetic lethality has been explored as an alternative for better understanding of complex interactions at the sub-cellular level. Therefore, post identifying synthetic lethals, the next objective was to construct protein-protein interaction networks using the STRING database. The final objective being studying the constructed networks and gathering literature evidence for some of the novel interactions observed. This study was conducted amongst 5 clinically relevant strains of pathogens: *Klebsiella pneumoniae MGH 78578, Staphylococcus aureus N315, Mycobacterium tuberculosis H37Rv, Salmonella enterica serovar Typhimurium LT2* and *Shigella dysentriae Sd197.* Programming languages Python and MATLAB (v2021b) were used for obtaining synthetic lethals, with roughly 380-400 lethal genes being generated for each of the organisms. The cutoff for lethality were taken from the Fast-SL documentation. Later, web parsing package BeautifulSoup4 was utilised for modifying the gene IDs obtained into STRING database recognizable formats. STRING database recognized each of the lethal genes with 100% sequence similarity and constructed an interaction network which also provided useful insights into the gene ontology terms. As expected amongst lethal genes, majority of them belonged to key biosynthesis pathways. Finally, the interactions with the highest levels of confidence were viewed and novel interactions between proteins with different pathway annotations were studied. Experimental evidence from other closely related proteins were gathered for such interactions to prove the utility of synthetic lethals in understanding complexity of pathways.

Specific Contribution

- Conducted literature survey into synthetic lethals and the type of studies carried out using protein-protein interaction networks
- Obtained knowledge related to all the tools used in the project; the Fast-SL documentation as well as the databases utilized.
- Prepared and executed scripts in the programming languages of Python and MATLAB to obtain the necessary results for the project.

Specific Learning

- Gained significant experience in coding in Python and MATLAB, as well as debugging the code.
- Acquired useful knowledge related to the types of interactions occurring at the genome level and especially available wet lab experimentations to uncover such interactions through the extensive literature survey performed for the project.
- Sought assistance from online learning modules such as MATLAB Onramp course and courses on network biology available on the NPTEL platform, which immensely helped understanding the background of this project.

CHAPTER 1

INTRODUCTION

1.1 SYNTHETIC LETHALITY

Network biology as a domain has opened up wide avenues for understanding organisms with far reaching depth that is not as quickly attainable with conventional experiments. Models of organisms clearly depict the state of the 'living system' in terms of interconnected metabolic reactions. This visualization of an organism helps identify several factors under varying conditions, thereby outlining its functionality with precision. Such studies, therefore, when applied to the genome scale, help identify classes of genes, notably the essential and non-essential genes. The principle behind understanding the significance of a genetic interaction is by perturbing the normal state of function, i.e genetic suppression. The cascade of effects on the rest of the system, reveals the functionality of the target gene.

Essential genes are therefore the class of genes which when suppressed, or made to be inactive under minimum nutrient levels, result in the death of the organism, and non-essential genes are those that do not bring about this result (Guarente, 1993). The death of the organism per se is visualized as zero biomass formation in network-based studies. However, not all non-essential genes can be considered insignificant to the organism's survival. Synthetic Lethals (SLs) refers to the pairs of non-essential genes, which upon deletion prove detrimental to the organism. Even if either one amongst the pair contribute to viability despite deletion, the corresponding pair does not qualify to be labelled as lethals (Kaelin, 2005). Lethals are therefore very interesting to note for drug targets, as inhibition of the products of the lethal genes can immediately cause the death of the organism. Hence, lethal genes have been explored in the backdrop of cancer progression as well, and has been proven to be very effective (Suthers, Zomorrodi and Maranas, 2009).

Synthetic lethality has therefore been able to answer for the high level of robustness observed in biological systems where the organism sustains itself despite perturbations. Lethal pairs can occur as double, triple as well as quadruple pairs. Therefore, when one amongst the pair

undergoes deletion, or is erroneous, the flux is redistributed to flow via its corresponding pair instead, maintaining viability (Sambamoorthy and Raman, 2018). As conventional methods of identification include gene deletion and cell culture experiments could be time consuming and with a high number of false positives, alternative *in silico* methods are proving to be useful. Computational methods to identify lethals include computing biomass production under random deletions via flux balance analysis (FBA) (Kauffman, Prakash and Edwards, 2003). However, computing lethals in previous methods required immense processing power and was time consuming, hence, a more recent method that not only computes all possible combinations, but in also significantly lesser time (Pratapa, Balachandran and Raman, 2015) has been used for this project.

1.2 IDENTIFICATION OF LETHALS: THE FAST-SL WAY

Fast-SL is an algorithm developed by (Pratapa, Balachandran and Raman, 2015) Raman Lab at Indian Institute of Technology, Madras. It also utilizes FBA concepts as they had been proven useful in lethal identification already (shown in yeast models by (Harrison *et al.*, 2007)). Other such algorithms like SL finder (Suthers, Zomorrodi and Maranas, 2009) have modified its approach in the form of a Mixed Integer Linear Programming (MILP) problem, which has been found to be ineffective for larger networks. Therefore, Fast-SL was developed as a means to bypass the level of computational complexity seen in previous cases, by significantly reducing the set of reactions for the algorithm to search through. This method had been proven to be extremely efficient with respect to identifying higher orders of lethals, namely, triple and quadruple lethal pairs. As typically seen in FBA, Fast-SL too utilizes a linear programming (LP) problem where an objective function is chosen to be the maximum flux possible through the reaction responsible for biomass production (denoted v_{bio}) which is then observed under the given constraints for the metabolic network (stoichiometric matrix denoted S). (Pratapa, Balachandran and Raman, 2015) uses the following FBA formulation **(Figure 1.1)** where J represents the total set of a reactions in a given model; M represents the total set of metabolites; D represents the set of reactions marked for deletion and therefore their fluxes are set to zero; v_j represents the flux of the j^{th} reaction; v_{bio} represents the flux via the reaction

responsible for biomass production; s_{ij} refers to the element in the stoichiometric matrix; LB_j and UB_j represent the bounds of the j[the] reaction.

$$\max. \upsilon_{bio}$$

s.t.

$$\Sigma_j s_{ij} \upsilon_j = 0 \qquad \forall i \in M, \forall j \in J$$

$$LB_j \leq \upsilon_j \leq UB_j \qquad \forall j \in J$$

$$\upsilon_d = 0 \qquad d \in D \subset J$$

Fig 1.1 FBA Formulation used by Fast-SL

Therefore, under the chosen constraints, Fast-SL aims to identify the specific combination of reaction deletions, that revert biomass flux to zero. The 'search space' or the set of reactions that the algorithm forms all possible combinations from, is the key to reducing computational complexity. Fast-SL aims to reduce this by significantly reducing the search space by a given logical framework. Lethality is therefore defined by this algorithm as the outcome of less than 1% wild type biomass flux. The algorithm initially identifies the l_1 norm of the flux vector, i.e, the least sum of all fluxes possible to produce the maximum biomass. This value is then taken to be the minimum solution for the LP problem and the reactions that carry non zero fluxes under this solution are made to constitute a set J_{nz}. The hypothesis behind Fast-SL is that all the single lethal reactions (Jsl) belong to the set $J_{nz.}$ Therefore, the search space is minimized to this set as J - J_{nz} would not contain any essential reactions. With this search space being utilized for computing Jsl, Jdl modifies its approach by assuming that one amongst the pair of reactions would be present in J_{nz}. This is because both components of the pair cannot have zero flux under maxium biomass production, thereby rendering both non-essential. Therefore, one amongst the pair must be in $J_{nz.}$ Once again, a solution is arrived at after deleting i[th] reaction and then a lethal reaction is identified from the set $J_{nz,i} - Jsl$, resulting in a pair of lethal reactions. A similar approach is used in identifying triple lethal reactions as well. Gene Protein Reaction (GPR) associations are then utilized by Fast-SL to arrive at the corresponding lethal genes as well (Sgd, Dgd & Tgd) (Pratapa, Balachandran and Raman, 2015).

This algorithm has been proven to be immensely time efficient, with higher orders of SLs (eg: triple lethals) take only about 8.5 minutes in a 6 core CPU using Fast-SL as opposed to exhaustive enumeration methods of the MCSEnumerator which takes almost double the time in a 12 core CPU (Pratapa, Balachandran and Raman, 2015). Implementation of Fast-SL requires MATLAB (v2016a, The MathWorks Inc.) along with the COBRA Toolbox v2.0 (Schellenberger *et al.*, 2011). The algorithm can be found at (https://github.com/RamanLab/FastSL).

1.3 THE NEED FOR A LETHALS DATABASE: CASTLE

Functional redundancies are by far the most interesting phenomenon uncovered alongside lethality. Some insights into these revealed pairs of lethals with participation in highly distant pathways, thus making it clear how little is known about nature's complex designs of adaption. Synthetic lethals as discussed previously hold the key to unraveling complex networks, and identify the key metabolic pathways of an organism. (Sambamoorthy and Raman, 2018) discusses how functionally distant genes also exist as synthetic lethal pairs and can be called non intuitive SLs. Though available models of organisms do not necessarily constitute all the happening reactions, it is expected that as understanding lethals would pave way to identifying novel interactions and 'compensation mechanisms' as coined by (Sambamoorthy and Raman, 2018). Therefore, a comprehensive database for hosting all information on synthetic lethals is under development at Raman Lab, IIT-M. The database currently hosts over 100 organisms and their synthetic lethals (reactions as well as genes) which were computed using Fast-SL. The current organisms' models were taken from databases such as Virtual Metabolic Human (VMH) and BiGG. Therefore, the results of this project aids in the argument for this database, its validity as well as its possible applications.

1.4 MOTIVATION

In the wake of multidrug resistance amongst pathogens, and discovery of unexplained defense mechanisms such as persistence, studies related to essentiality prove to be highly useful in selecting drug targets. Not just as drug targets, but lethality is also being utilized as a means of

treatment against some types of cancer. Recent cancer studies have uncovered the existence of the synthetic lethal pair of BRCA1/2 and PARP genes, and clinical trials of PARP inhibitor drugs have also shown positive response amongst those with BRCA gene mutations (Lord, Tutt and Ashworth, 2015). From the perspective of microbiology, they have also helped reveal novel interactions behind clinically relevant strains such as methicillin resistant *Staphylococcus aureus* (MRSA). (Campbell *et al.*, 2011) is one such study where the lethal link between wall teichoic acids (WTA) and a peptidoglycan transpeptidase (PBP2A) has shown to incorporate reduced susceptibility to β-lactam antibiotics. Such is the potential of synthetic lethals, and therefore the focus of this project.

CHAPTER 2

OBJECTIVES

As the thesis title suggests, this project aims to understand in depth the interactions amongst lethal genes of a few microbes, and identify novel interactions amongst the same. Therefore, the fundamental aims of this project are as follows:

- Compute synthetic lethals for select organisms using the Fast-SL algorithm (Pratapa, Balachandran and Raman, 2015)
- Construct protein-protein interaction networks using the STRING database
- Study the constructed network and survey literature as support for novel interactions

Therefore, the first step involves obtaining biochemical reaction models from established databases such as Virtual Metabolic Human (VMH) and BiGG. Over the progress of the project, the databases used were narrowed to just the BiGG database, reasons for which will be elaborated later in the report. The reaction models were exported in both SBML (Systems Biology Markup Language) and XML (Extensible Markup Language) formats. With the given models, Fast-SL code from the GitHub repository (URL mentioned previously) was run on MATLAB v2016a. For the chosen organisms already present on the Raman Lab maintained

CASTLE database (https://github.com/RamanLab/CASTLE), the list of synthetic lethals (single, double and triple) were exported directly from the database in the .csv format.

STRING was chosen for the second step of network studies as it has been shown to have produced networks with the most experimentally verified interactions (Bajpai *et al.*, 2020). This recent study compared STRING against 16 other popular interaction databases such as UniHI, hPRINT, APID, HIPPIE etc. and found that STRING supported almost 84% of its projected interactions with experimental evidence (Bajpai *et al.*, 2020). With excellent coverage and confidence scores, STRING was chosen as it aids the aims of the project, with respect to finding support for the novel interactions identified.

Lastly, additional interactome databases (IntAct, maintained by EMBL-EBI) and research publication search engines (Google Scholar, NCBI, PubMed etc.) were scourged to find experimental support for the identified interactions. The identified interactions were amongst the query proteins (compiled list of single, double and triple lethals) and with the highest confidence scores. The interactions were also chosen on the basis of its corresponding gene ontology. Overall, this project hopes to provide strong arguments for research focus on synthetic lethals using network-based approaches.

CHAPTER 3

METHODOLOGY

3.1 TARGET ORGANISMS

The project initially aimed to perform the above-mentioned objectives for the entire list of microbial models available on the Virtual Metabolic Human database (also referred to as the AGORA database). The entire database was exported and the list of organisms exceeded 800. Due to the large number of models, MATLAB (v2021b) on local computer could not be used to support Fast-SL. Instead, a cluster server of suitable capacity hosted at the Computational

Systems Biology Lab at IIT-Madras was used. The AGORA database was chosen as it was more expansive than BiGG database, and lethal based studies were not popularly carried out using AGORA models. In such a scenario, this project hoped to pioneer such studies using all the models available on AGORA and submit the obtained results to the developing CASTLE database for credibility. However, the project faced setbacks when the cluster server encountered memory allocation issues which could not be resolved in time for the completion of this project. Thus, the focus of this project was significantly reduced to five clinically relevant pathogens available on the BiGG database. Use of models from AGORA also had to be discontinued due to gene ID issues which will explained in the later sections.

Once the project workflow was decided, it was tested using a sample organism model of *Escherichia coli* 536 strain from the BiGG database. This sample was used as it is a well-established model and species with high available literature. After the workflow was finalized, the target organisms were narrowed down to 5 of the most clinically relevant pathogens available on the BiGG database, namely:

- *Klebsiella pneumoniae MGH 78578*
- *Staphylococcus aureus N315*
- *Mycobacterium tuberculosis H37Rv*
- *Salmonella enterica* serovar Typhimurium *LT2*
- *Shigella dysentriae Sd197*

The specific strain of MGH 78578 was chosen amongst other clinical strains of the UTI pathogen *K. pneumoniae* due to its highest number of antibiotic resistance genes (almost 28 genes) and significant efflux pump genes (Kumar *et al.*, 2011). *Staphylococcus aureus*, the most notorious nosocomial pathogen, responsible for skin infections and in severe cases endocarditis and toxic shock syndrome, was also chosen for this study. Specifically, the methicillin resistant endemic causing N315 strain was chosen due to the higher degree of resistance (Lindsay and Holden, 2004). The pathogen responsible for tuberculosis in humans, *M. tuberculosis* and its most well studied strain H37Rv had also been chosen due to the well characterized nature of the strain (Borrell *et al.*, 2019). *S. enterica* serovar Typhimurium LT2

strain was chosen for its enhanced ability in biofilm formation as opposed to other strains (Fornefeld *et al.*, 2017). Lastly, *S. dysentriae* responsible for most diarrheal diseases in humans, and for a third of infant deaths was also chosen for its clinical relevance (Jalal *et al.*, 2022).

3.2 COMPUTING SYNTHETIC LETHALS USING FAST-SL

As mentioned previously, Fast-SL requires either XML or MAT files of organisms as input to identify the lethal reactions and genes. As the project required analysis involving only the lethal genes, the function (fastSL_tg) from the Fast-SL source code was used. The outputs of this function produced 3 array structures containing the single, double and triple lethal genes respectively. Each of these lists were then made to convert to a csv file as the final output. As local computer version of MATLAB (v2021b) differed from that on the server (v2016a), some of the functions used had to be modified (eg: writecsv() and readCbModel() functions). **(Figure 3.1)** shows the MATLAB code used to compute synthetic lethals on the cluster server.

```
addpath('/data/lakshmi/FastSL-master/Genes');
addpath('/data/lakshmi/cobratoolbox');
addpath('/data/lakshmi');
initCobraToolbox
%solverOK = changeCobraSolver('ibm_cplex');
folder = '/data/lakshmi/mat/';
file_list = dir(fullfile(folder,'*.mat'));
problem_orgs = {};
for k = 2:length(file_list)
   tic
   orgname = file_list(k).name;
   load(orgname);
   toc
   try
       solution = optimizeCbModel(model);
       maxGrowth(k,1) = solution.f;
       mkdir('/data/lakshmi/results');
       tic
       [sgd,dgd,tgd] = fastSL_tg(model,0.01,1);
       toc
       tic
       orgname = strrep(orgname,'.mat','');
       sgdpath = strcat('/data/lakshmi/results/',orgname,'_sgd.csv');
       dgdpath = strcat('/data/lakshmi/results/',orgname,'_dgd.csv');
       tgdpath = strcat('/data/lakshmi/results/',orgname,'_tgd.csv');
       toc
       cell2csv(sgdpath,sgd);
       cell2csv(dgdpath,dgd);
       cell2csv(tgdpath,tgd);
       disp(k);
   catch
       problem_orgs = [problem_orgs, orgname];
   end
   cell2csv('/data/lakshmi/problem_orgs.csv',problem_orgs);
end
fprintf('\n Finished Fast-SL on all 818 organisms of AGORA...');
```

Fig 3.1: MATLAB Script for computing synthetic lethals

As per Fast-SL code documentation, the default value for lethality cutoff was taken as 0.01 and the same was used for this project. While the code used .xml files as input initially, due to error (empty fieldnames while reading the model) encountered with the function readCbModel() from the COBRA Toolbox (Schellenberger *et al.*, 2011), the alternate version of .mat files and the load() function was used. Fast-SL documentation (Pratapa, Balachandran and Raman, 2015) also noted the most compatible solver to be the ibm_cplex solver, hence, the default glpk and gurobi solvers were replaced. The SGE job scheduler pre-installed on the server was used to run all MATLAB code. **(Figure 3.1)** represents the code used for running Fast-SL on all organisms of AGORA, however, due to issues with the server, the code was modified to be compatible with the version installed on local computer and looping was removed.

3.3 RESOLVING GENE IDS

One of the primary differences encountered in models from AGORA and BiGG were the format in which gene IDs were listed. While AGORA resembled Patric Database IDs the most, BiGG provided Entrez IDs which were universally recognized and interconvertible. The PATRIC database IDs example is 'fig|592020.peg.4.1818' and AGORA provided IDs without the prefix of 'fig|' (Gillespie *et al.*, 2011). While the project initially focused on AGORA models, these IDs were attempted at converting to alternate universally recognized protein IDs such as UniProt which could be provided as input to STRING database for the next step. While the only database in which these IDs could be recognized was AGORA and Patric, the latter had an ID mapping tool for ID conversion. However, this method too did not provide the expected results and could not translate these IDs into either UniProt KB Accession IDs or any other. The tool worked only for conversion into RefSeq IDs (hosted by NCBI database), however the corresponding gene entries on NCBI did not provide further external database links, and entries themselves were removed from NCBI due to being outdated. Even UniProt ID mapping tools failed to recognize these RefSeq IDs due to the entries being removed. Lastly, AGORA URL for each gene ID consisted of the metabolite it translates into, the reaction it participates in and finally the nucleotide sequence of the gene. As the NT sequence was the only possible way for them to be recognized by STRING database, it still needed to be converted to protein sequence, which was decided at the time to be too time consuming. Hence, work was stopped related to AGORA based models.

Fortunately, with BiGG database, the URL for each gene contained the corresponding amino acid sequence, external database links for UniProt, InterPro and Gene Ontology. Hence the project utilized the Python programming language to develop two web parsing scripts which were tested and run-on Google Colab. Web parsing is extracting information from a website using its HTML script. Parsing was performed using the Python package Beautiful Soup (documentation available at https://beautiful-soup-4.readthedocs.io/en/latest/). While the first parsing script extracted only the external database links after constructing URLs for each synthetic lethal gene ID **(Figure 3.2)**, the second parsing script extracted the amino acid sequence directly **(Figure 3.3)**. This was done as the external links section was not available

for a significant number of lethal genes for even *E. coli* 536 model, hence we expected more absence for the other relatively lesser studied models.

Though STRING API did not have provisions for searching multiple inputs based on amino acid sequence, the web interface did. Important to note that the web interface only accepted input files as 'multi-fasta' formats, and hence the second script also had additions where the result were the corresponding AA sequences of all single, double and triple lethal genes of a given model, in the multi fasta format **(Figure 3.4)**. This file was readily recognizable by STRING database, and almost every sequence was mapped with a 100% sequence similarity score. Alternatively, the sequences were also mapped using bi-directional blast available on Patric database (the proteome comparison tool). The results were found to be the same as STRING mapping and hence the default mapping done by STRING was proceeded with to create the network.

```
base_url = "http://bigg.ucsd.edu/models/iEC042_1314/genes/"

import pandas as pd
import requests
from bs4 import BeautifulSoup

gene_data = pd.read_csv('/content/Escherichia coli 042.csv')
gene_data.head()

gene_data = gene_data.fillna('dummy')

for column in gene_data.columns:
  gene_data[column] = gene_data[column].apply(lambda x: x.strip("'"))
gene_data.head()

gene_data = gene_data.drop(['A','B','C','D','E','F'], axis=1)
gene_data.head()

"""http://bigg.ucsd.edu/models/iEC042_1314/genes/{gene_id}"""

for column in gene_data.columns:
  gene_ids = list(gene_data[column])
  gene_ids = list(filter(('dummy').__ne__, gene_ids))

  gene_names = []
  uniprot = []
  interpro = []
  goa = []

  #cnt = 0
  for gene_id in gene_ids:
    # cnt += 1
    # if cnt == 5:
    #   break
    full_url = base_url + gene_id
    content = requests.get(full_url)
    soup = BeautifulSoup(content.text, 'html.parser')
    u = []
    i = []
    g = []

    for link in soup.find_all('a'):
      if str(link.get('href')).startswith("http://identifiers"):
        url_parts = str(link.get('href')).split('/')
        if url_parts[-2] == 'uniprot':
          u.append(url_parts[-1])
        elif url_parts[-2] == 'interpro':
          i.append(url_parts[-1])
        elif url_parts[-2] == 'goa':
          g.append(url_parts[-1])

    gene_names.append(gene_id)
    uniprot.append(','.join(u))
    interpro.append(','.join(i))
    goa.append(','.join(g))

  mydf = pd.DataFrame(list(zip(gene_names,uniprot,interpro,goa)), columns = ['gene_id','uniprot','interpro','goa'])
  mydf.to_csv(column+'.csv')
```

Fig 3.2 Python external links parsing script: Extracts external database links from each BiGG URL for lethal genes.

```
base_url = "http://bigg.ucsd.edu/models/iSDY_1059/genes/"

"""Import the BeautifulSoup Package"""

import pandas as pd
import requests
from bs4 import BeautifulSoup

"""Import file containing all SLs identified in an organism"""

gene_data = pd.read_csv('/content/Shigella_dysenteriae_Sd197.csv')
gene_data.head()

gene_data = gene_data.fillna('dummy')

"""Remove '' from gene IDs to make it recognisable and usable in URL"""

for column in gene_data.columns:
  gene_data[column] = gene_data[column].apply(lambda x: x.strip("'"))
gene_data.head()

"""Drop A-F if analysis only for SL genes and not reactions"""

gene_data = gene_data.drop(['A','B','C','D','E','F'], axis=1)
gene_data.head()

"""Loop will create a prot.txt file in multi FASTA format and also protein_sequence.csv
containing all sequences corresponding to gene ID.
'Skip:' output denotes genes with no sequences on their BiGG URLs."""

#Run this;

dfs = []
f = open("prot.txt","w")

for column in gene_data.columns:
  gene_ids = list(gene_data[column])
  gene_ids = list(filter(('dummy').__ne__, gene_ids))

  gene_names = []
  protein_seq = []

  for gene_id in gene_ids:
    #cnt += 1

    full_url = base_url + gene_id
    content = requests.get(full_url)
    soup = BeautifulSoup(content.text, 'html.parser')
    try:
      seq = soup.find_all("p", {"class": "sequence"})[1].getText()
      protein_seq.append(seq)
      gene_names.append(gene_id)
      f.write(">")
      f.write(gene_id+'\n')
      f.write(seq+'\n')
    except:
      print("skip: ",gene_id)
  mydf = pd.DataFrame(list(zip(gene_names,protein_seq)), columns = [column+' gene_id',column+' protein'])
  dfs.append(mydf)

f.close()
protein_seq_data = pd.concat(dfs, axis=1)
protein_seq_data.to_csv("protein_sequence.csv",index=False)
```

Fig 3.3 Python sequence parsing script: Extracts the amino acid sequence from BiGG URL and writes the output into a file in the multi-fasta format

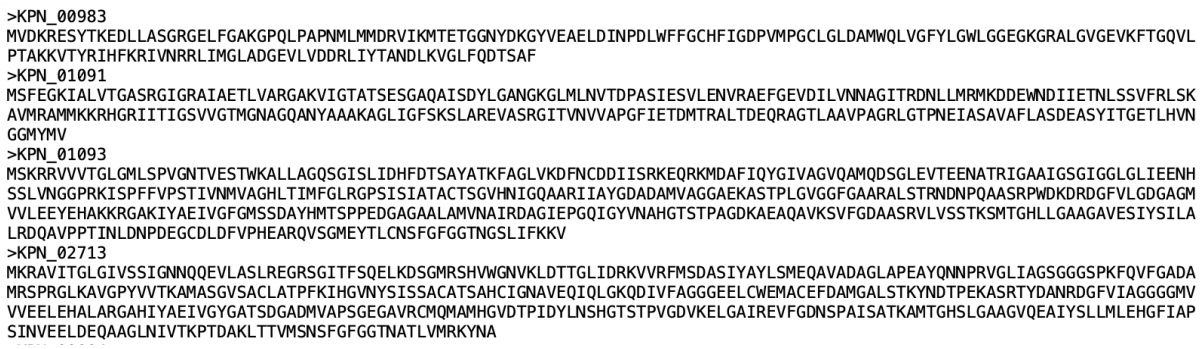


Fig 3.4 Preview of output multi-FASTA file: produced by the above script for *K. pneumoniae*

3.4 CONSTRUCTING PPI NETWORKS VIA STRING

As STRING web interface was used for the project, the multi-fasta files obtained in the previous step was uploaded as input under the 'multiple proteins by sequence' query option. The database initially maps each sequence to the corresponding protein available already in its database for the given organism and strain. Post mapping, it automatically selects those proteins identified with >99.99% similarity scores and proceeds to construct the interaction network. The database constructs the network using experimental sources as well as using its prediction tools, from other interactions of homologous proteins (Szklarczyk *et al.*, 2011). The results contain nodes of proteins and each interaction having a confidence score. The database also provides accessory information such as protein domain and 3D structures in the interactive network image that it generates. STRING also provides insights into Gene Ontology based clustering, and has user friendly options to modify the view of the network. These features were utilized for the project in arriving to necessary conclusions.

3.5 SEARCH FOR INTERACTION EVIDENCE

The final step in the project workflow was analyzing those lethal interactions belonging to the first few significant GO terms projected by STRING and text mining for publication evidence. It was also noted that the evidence needs to be in the form of studies that did not use network concepts to predict such interactions, instead the study needs to perform *in vitro* biochemical assays and other experimentations. Since STRING also provided a list of node-node

interactions under each protein family and domain with a descending order of confidence scores, these lists were probed to find interactions with high scores as well as clear experimental evidence. The interaction was also searched under similar interactome databases such as IntAct, which provided publications as well as publication confidence scores. Though most of these evidences did not constitute studies using the exact same organism and strain for their experiments, they constituted closely related species and highly similar biochemical pathways.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 DECONSTRUCTING THE PPI NETWORK

The network clustering coefficient for all 5 organisms ranged between 0.355 to 0.448 which indicate a moderate level of clustering and less dense network. Each network also provides for results in the tsv format for functional annotation of each node as well as its corresponding network coordinates, for the purpose of future reconstruction of network using tools such as Cytoscape etc. Gene ontology studies map the query proteins to their corresponding functional annotations. These studies have 3 primary components-biological, molecular and cellular components. While biological component annotates the protein in terms of the biochemical pathways it participates in, molecular component elaborates on the functions of the protein and cellular component groups proteins based on their position of activity within the cell. STRING also provided insights along these components for all 5 organisms, a sample of the results for *S. aureus* have been attached below **(Figure 4.1)**. The count in network column denotes the number of nodes having the given GO term vs the total number of proteins including the background that have the same term annotation. Strength denotes the effect of enrichment, which is calculated using a log ratio of observed proteins with a given annotation divided by expected proteins with the same annotation in a

randomly generated network. Lastly, the false discovery rate is calculated by STRING using the Benjamini Hochberg procedure with corrected p-values. These gene ontologies are obtained by STRING using annotations available on the AmiGO database.

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0006189	De novo imp biosynthetic process	12 of 12	0.99	4.69e-06
GO:0009206	Purine ribonucleoside triphosphate biosynthetic process	10 of 10	0.99	3.66e-05
GO:0000162	Tryptophan biosynthetic process	8 of 8	0.99	0.00029
GO:0046654	Tetrahydrofolate biosynthetic process	7 of 7	0.99	0.00084
GO:0009423	Chorismate biosynthetic process	7 of 7	0.99	0.00084
				(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0046933	Proton-transporting atp synthase activity, rotational mechanism	8 of 8	0.99	0.00055
GO:0016774	Phosphotransferase activity, carboxyl group as acceptor	7 of 7	0.99	0.0016
GO:0003989	acetyl-CoA carboxylase activity	6 of 6	0.99	0.0049
GO:0016682	Oxidoreductase activity, acting on diphenols and related substanc...	5 of 5	0.99	0.0135
GO:0046912	Transferase activity, transferring acyl groups, acyl groups convert...	4 of 4	0.99	0.0365
				(more ...)

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0045259	Proton-transporting atp synthase complex	8 of 8	0.99	0.00070
GO:0045261	Proton-transporting atp synthase complex, catalytic core f(1)	5 of 5	0.99	0.0131
GO:0045239	Tricarboxylic acid cycle enzyme complex	5 of 5	0.99	0.0131
GO:0009317	acetyl-CoA carboxylase complex	4 of 4	0.99	0.0333
GO:1990204	Oxidoreductase complex	11 of 20	0.73	0.00089

Fig 4.1: Functional Enrichment observed in *S. aureus N315* lethal genes

The most significantly enriched pathway in both *S. dysentriae* and *S. enterica* was the chorismate biosynthetic process, which is the formation of unsymmetrical ether from intermediates in the biosynthesis of aromatic amines. The enriched pathway for *M. tuberculosis* was however the Glucose-6-phosphate metabolic process with the chorismate biosynthetic process closely following as the fourth most enriched pathway term. The chorismate biosynthetic process remains in the top 5 enriched terms for *S. aureus* as well, but the topmost in this case was found to be de novo IMP biosynthesis, which is the formation of inosine monophosphate from the purine ring of ribose-5-phosphate. Moreover, ATP synthesis associated processes as expected were significantly enriched amongst all 5 organisms. It is also in line with previous findings that lethal genes encode proteins responsible for synthesis of key compounds in metabolic pathways.

4.2 NOVEL INTERACTIONS

The network was then filtered to constitute only those interactions with the highest confidence scores, and the interactions were depicted with different coloured lines to depict the type of evidence supporting it **(Figure 4.2)**; where blue indicates curated databases, pink indicates experimentally determined, dark green indicates gene neighborhood, blue indicates gene co-occurrence, light green indicates available text mining, black indicates co-expression data and lilac indicates protein homology.

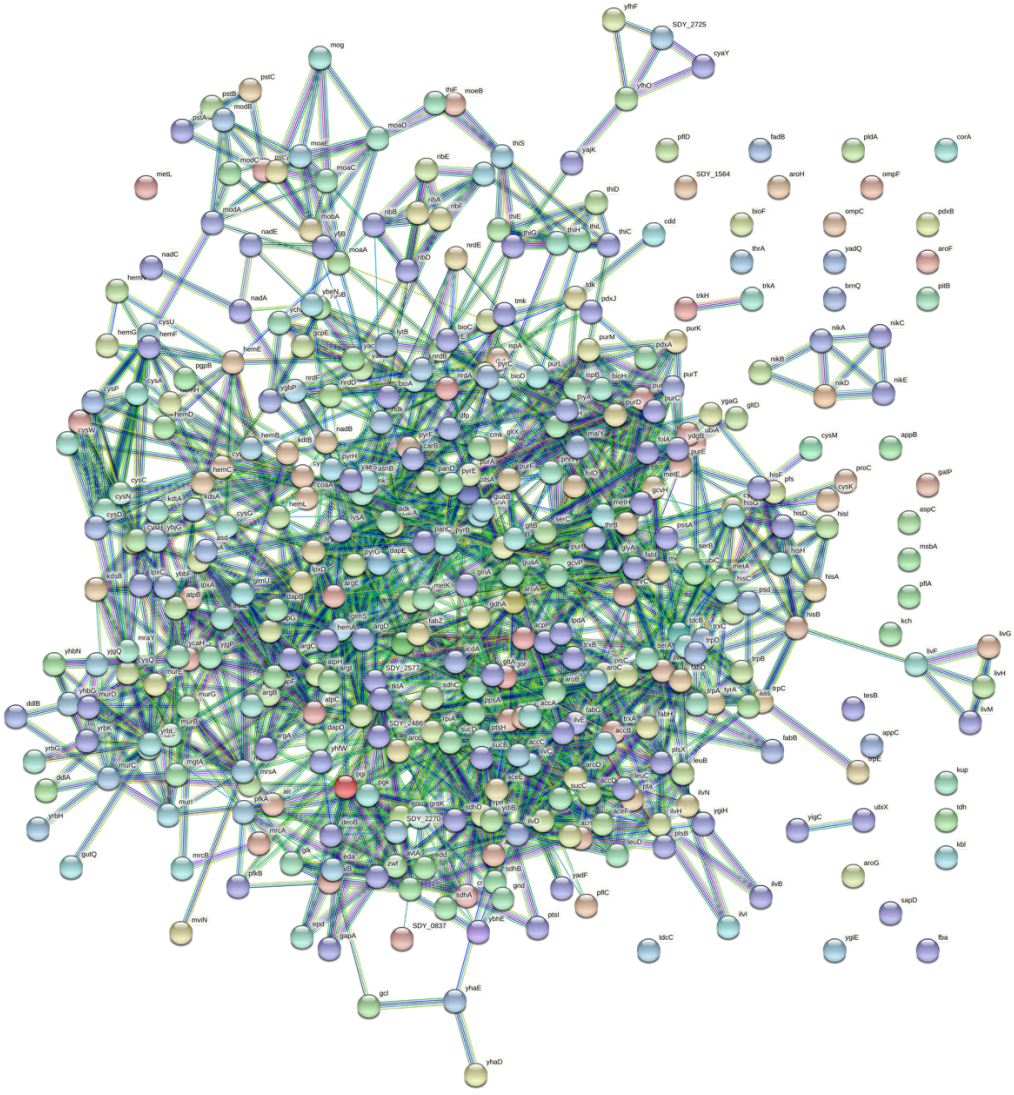


Fig 4.2: PPI Network Image of *Shigella dysentriae*: represents the high confidence interactions amongst lethals of *Shigella dysentriae*

Upon probing the list of interactions with descending confidence scores, the majority of highest confidence values (>0.900) had interactions between lethal genes encoding for aminotransferases and key enzymes in the glucose pathways (phosphoglycerate kinase, transaldolase/ketolase, 6-phosphofructokinase etc.). This was observed in *S. aureus, S. enterica* and *S. dysentriae*. Whereas, in *K. pneumoniae* and *M. tuberculosis,* Shikimate dehydrogenase and chorismate synthase (both implicated in chorismate biosynthetic process) was observed. Amongst this list, the few interactions with the most diverse functions and implicated pathways, which had experimental support were chosen and compiled into **(Table 4.1)**.

Table 4.1: Compiled list of interactions with publication evidence

Organism	**Interaction**	**Interaction Confidence**	**Supporting Literature DOI**	**Literature Confidence**	**Evidence Type**
Klebsiella pneumoniae subsp. MGH 78578	KPN_00192 (LpxD) - KPN_00942 (lpxK)	0.996	https://doi.org/10.3390/biom10020266	0.198	In-silico structure based study (Bhaskar *et al.,* 2020)
Klebsiella pneumoniae subsp. MGH 78578	KPN_01125 (ywlF) - KPN_02238 (prsA)	0.901	https://doi.org/10.1186/s12934-020-01302-7	0.203	Review on pathways (Liu *et al.,* 2020)

Table 4.1: Compiled list of interactions with publication evidence (contd.)

Klebsiella pneumoniae subsp. MGH 78578	KPN_01127 (tkt) - KPN_03757 (rpe)	0.98	https://doi.org/10.1074/mcp.m115.057117	0.286	Quantitative co-purification assay (Shatsky *et al.*, 2016)
Staphylococcus aureus subsp. N315	SAOUHSC_00999 (qoxD) - SAOUHSC_01002 (qoxA)	0.999	https://doi.org/10.1038/82824	0.442	Oxidoreductase assay (Abramson *et al.*, 2000)
Staphylococcus aureus subsp. N315	SAOUHSC_01001 (qoxB) - SAOUHSC_01002 (qoxA)	0.999	https://doi.org/10.1038/s41594-018-0172-z	0.999	Mass spectrometry assay (Hartley *et al.*, 2018)
Staphylococcus aureus subsp. N315	SAOUHSC_00148 (argJ) - SAOUHSC_00150 (argD)	0.999	https://doi.org/10.1074/mcp.M115.057117	0.281	Quantitative co-purification assay (Shatsky *et al.*, 2016)
Mycobacterium tuberculosis H37Rv	Rv1310 (atpD) - Rv1311 (atpC)	0.999	https://doi.org/10.1083/jcb.201404118	0.994	Biochemical assay (Rahman *et al.*, 2014)

Table 4.1: Compiled list of interactions with publication evidence (contd.)

Mycobacterium tuberculosis H37Rv	Rv2763c (dfrA) - Rv2764c (thyA)	0.999	https://doi.org/10.1186/s12864-019-5615-3	0.455	Genome analysis (Oppong *et al.*, 2019)
Mycobacterium tuberculosis H37Rv	Rv1286 (cysN) - Rv2392 (cysH)	0.999	https://doi.org/10.1038/s41598-019-45652-8	0.597	Gene expression analysis (Gample, Agrawal and Sarkar, 2019)
Salmonella enterica subsp. LT2	STM0985 (lpxK) - STM3724 (kdtA)	0.998	https://doi.org/10.1038/s41579-019-0201-x	0.955	Review on pathways (Simpson and Trent, 2019)
Salmonella enterica subsp. LT2	STM2947 (cysl) - STM3477 (cysG)	0.95	https://doi.org/10.1038/s41467-020-14722-1	0.837	Mutagenesis experiments (Pennington *et al.*, 2020)
Salmonella enterica subsp. LT2	STM2430 (cysK) - STM3699 (cysE)	0.997	https://doi.org/10.1007/bf02513025	0.962	Gene expression (Sivaprasad *et al.*, 1992)

Table 4.1: Compiled list of interactions with publication evidence (contd.)

Shigella dysentriae Sd197	SDY_3720 (purH) - SDY_3721 (purD)	0.999	https://doi.org/10.1093/gbe/evz035	0.972	Review on pathways (Cruz *et al.*, 2019)
Shigella dysentriae Sd197	SDY_3720 (purH) - SDY_4624 (pyrB)	0.942	https://doi.org/10.3389/fmicb.2019.03101	0.878	Proteomic analysis (Hirschfeld *et al.*, 2020)
Shigella dysentriae Sd197	SDY_1353 (ribA) - SDY_1888 (ribE)	0.933	https://doi.org/10.1038/s41598-020-62890-3	0.759	Biochemical assay (Anam, Nasuno and Takagi, 2020)

The interactions as compiled above, indicate lethal genes encoded proteins participating in moderately distant biochemical pathways. For instance, the proteins argJ and argD in *S. aureus* does not have direct evidence from a study on *S. aureus* itself, but a quantitative tagless co-purification assay performed on *Desulfovibrio vulgaris* has shown direct interaction amongst the two proteins in the same L-ornithine biosynthesis pathway. This evidence was considered after evaluating the similariy between the *S. aureus* proteins and *D. vulgaris* proteins. In both the species, the implicated nodes argJ functioned as an acyl donor part of the LpxD subfamily of proteins and argD was an aminotransferase implicated in ATP phosphate transfer reactions. Similar evidences were collected for the other organisms, after evaluating the homology of function and pathway terms in the species upon which experimental evidence is based upon, and the species of interest. Therefore, over 15 such references were collected for all 5 organisms in total, which depict interactions in the species, which have not yet been directly implicated. These predictions of interactions, have been done using similar experiments on closely related species as well as highly homologous proteins.

CHAPTER 5

CONCLUSION

With the obtained results, this project collected evidence to support the likelihood of the constructed network. As there are not many studies revolving around these lists of lethal genes, it is difficult to gather experimental support for even lesser-known pathogens. However, the primary purpose of this project was firstly, to validate the need for a study on lethals and second, to prove that network studies based on lethals can act as preliminary evidence of direct interaction. This can be observed in the final results of the project as well, where the gathered evidence was not from experiments on the same pathogen but another pathogen with highly homologous proteins. The networks were also predicted and constructed by STRING using many such parameters, such as PubMed abstracts implicating the same proteins but in other closely related species, the neighborhood of the said genes in the whole genome of the target species, co-occurrence of these genes across other closely related species and finally available literature data. Other interactome databases, such as IntAct uses databases such as ChEBI and Ensembl as their reference resource when scoring an interaction. IntAct too has a similar scoring scheme where interaction detection method (Biochemical/Protein complementation assay etc.) is given a specific 'weight' (Kerrien *et al.*, 2012). The combined confidence score taking into consideration all the said parameters is what is displayed in **(Table 4.1)** as well. The chosen interactions, therefore, are those types of interactions which were discovered in the target species only via network prediction using lethals as query proteins, and experiments depicting those interactions have not been performed yet. This implies the novel nature of interaction predictions done using lethals. Lethals therefore have the potential to enhance antimicrobial research by exploring efficient metabolic targets and this project hopes to sustain this in the near future.

REFERENCES

Abramson, J. *et al.* (2000) "The structure of the ubiquinol oxidase from Escherichia coli and its ubiquinone binding site," *Nature Structural Biology 2000 7:10*, 7(10), pp. 910–917. Available at: https://doi.org/10.1038/82824.

Anam, K., Nasuno, R. and Takagi, H. (2020) "A Novel Mechanism for Nitrosative Stress Tolerance Dependent on GTP Cyclohydrolase II Activity Involved in Riboflavin Synthesis of Yeast," *Scientific Reports 2020 10:1*, 10(1), pp. 1–10. Available at: https://doi.org/10.1038/s41598-020-62890-3.

Bajpai, A.K. *et al.* (2020) "Systematic comparison of the protein-protein interaction databases from a user's perspective," *Journal of Biomedical Informatics*, 103, p. 103380. Available at: https://doi.org/10.1016/J.JBI.2020.103380.

Bhaskar, B.V. *et al.* (2020) "Structure-Based Virtual Screening of Pseudomonas aeruginosa LpxA Inhibitors using Pharmacophore-Based Approach," *Biomolecules 2020, Vol. 10, Page 266*, 10(2), p. 266. Available at: https://doi.org/10.3390/BIOM10020266.

Borrell, S. *et al.* (2019) "Reference set of Mycobacterium tuberculosis clinical strains: A tool for research and product development," *PLOS ONE*, 14(3), p. e0214088. Available at: https://doi.org/10.1371/JOURNAL.PONE.0214088.

Campbell, J. *et al.* (2011) "Synthetic lethal compound combinations reveal a fundamental connection between wall teichoic acid and peptidoglycan biosyntheses in staphylococcus aureus," *ACS Chemical Biology*, 6(1), pp. 106–116. Available at: https://doi.org/10.1021/CB100269F/SUPPL_FILE/CB100269F_SI_001.PDF.

Cruz, D.C.B. *et al.* (2019) "Different Ways of Doing the Same: Variations in the Two Last Steps of the Purine Biosynthetic Pathway in Prokaryotes," *Genome Biology and Evolution*, 11(4), pp. 1235–1249. Available at: https://doi.org/10.1093/GBE/EVZ035.

Fornefeld, E. *et al.* (2017) “Persistence of Salmonella Typhimurium LT2 in soil enhanced after growth in lettuce medium,” *Frontiers in Microbiology*, 8(APR), p. 757. Available at: https://doi.org/10.3389/FMICB.2017.00757/BIBTEX.

Gample, S.P., Agrawal, S. and Sarkar, D. (2019) “Evidence of nitrite acting as a stable and robust inducer of non-cultivability in Mycobacterium tuberculosis with physiological relevance,” *Scientific Reports 2019 9:1*, 9(1), pp. 1–12. Available at: https://doi.org/10.1038/s41598-019-45652-8.

Gillespie, J.J. *et al.* (2011) “Patric: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species,” *Infection and Immunity*, 79(11), pp. 4286–4298. Available at: https://doi.org/10.1128/IAI.00207-11/SUPPL_FILE/FIGURE_S4.PDF.

Guarente, L. (1993) “Synthetic enhancement in gene interaction: a genetic tool come of age,” *Trends in Genetics*, 9(10), pp. 362–366. Available at: https://doi.org/10.1016/0168-9525(93)90042-G.

Harrison, R. *et al.* (2007) “Plasticity of genetic interactions in metabolic networks of yeast,” *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp. 2307–2312. Available at: https://doi.org/10.1073/PNAS.0607153104.

Hartley, A.M. *et al.* (2018) “Structure of yeast cytochrome c oxidase in a supercomplex with cytochrome bc1,” *Nature Structural & Molecular Biology 2018 26:1*, 26(1), pp. 78–83. Available at: https://doi.org/10.1038/s41594-018-0172-z.

Hirschfeld, C. *et al.* (2020) “Proteomic Investigation Uncovers Potential Targets and Target Sites of Pneumococcal Serine-Threonine Kinase StkP and Phosphatase PhpP,” *Frontiers in Microbiology*, 10, p. 3101. Available at: https://doi.org/10.3389/FMICB.2019.03101/BIBTEX.

Jalal, K. *et al.* (2022) “Identification of vaccine and drug targets in Shigella dysenteriae sd197 using reverse vaccinology approach,” *Scientific Reports 2022 12:1*, 12(1), pp. 1–19. Available at: https://doi.org/10.1038/s41598-021-03988-0.

Kaelin, W.G. (2005) “The Concept of Synthetic Lethality in the Context of Anticancer Therapy,” *Nature Reviews Cancer 2005 5:9*, 5(9), pp. 689–698. Available at: https://doi.org/10.1038/nrc1691.

Kauffman, K.J., Prakash, P. and Edwards, J.S. (2003) “Advances in flux balance analysis,” *Current Opinion in Biotechnology*, 14(5), pp. 491–496. Available at: https://doi.org/10.1016/J.COPBIO.2003.08.001.

Kerrien, S. *et al.* (2012) “The IntAct molecular interaction database in 2012,” *Nucleic Acids Research*, 40(D1), pp. D841–D846. Available at: https://doi.org/10.1093/NAR/GKR1088.

Kumar, V. *et al.* (2011) “Comparative genomics of Klebsiella pneumoniae strains with different antibiotic resistance profiles,” *Antimicrobial Agents and Chemotherapy*, 55(9), pp. 4267–4276. Available at: https://doi.org/10.1128/AAC.00052-11/SUPPL_FILE/TABLES_S1_S2_S7_S8_KUMAR_ET_ALA.ZIP.

Lindsay, J.A. and Holden, M.T.G. (2004) “Staphylococcus aureus: Superbug, super genome?,” *Trends in Microbiology*, 12(8), pp. 378–385. Available at: https://doi.org/10.1016/J.TIM.2004.06.004.

Liu, S. *et al.* (2020) “Production of riboflavin and related cofactors by biotechnological processes,” *Microbial Cell Factories 2020 19:1*, 19(1), pp. 1–16. Available at: https://doi.org/10.1186/S12934-020-01302-7.

Lord, C.J., Tutt, A.N.J. and Ashworth, A. (2015) “Synthetic Lethality and Cancer Therapy: Lessons Learned from the Development of PARP Inhibitors,”

https://doi.org/10.1146/annurev-med-050913-022545, 66, pp. 455–470. Available at: https://doi.org/10.1146/ANNUREV-MED-050913-022545.

Oppong, Y.E.A. *et al.* (2019) "Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance," *BMC Genomics*, 20(1), pp. 1–15. Available at: https://doi.org/10.1186/S12864-019-5615-3/FIGURES/3.

Pennington, J.M. *et al.* (2020) "Siroheme synthase orients substrates for dehydrogenase and chelatase activities in a common active site," *Nature Communications 2020 11:1*, 11(1), pp. 1–11. Available at: https://doi.org/10.1038/s41467-020-14722-1.

Pratapa, A., Balachandran, S. and Raman, K. (2015) "Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks," *Bioinformatics*, 31(20), pp. 3299–3305. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTV352.

Rahman, M. *et al.* (2014) "Drosophila Sirt2/mammalian SIRT3 deacetylates ATP synthase β and regulates complex V activity," *Journal of Cell Biology*, 206(2), pp. 289–305. Available at: https://doi.org/10.1083/JCB.201404118.

Sambamoorthy, G. and Raman, K. (2018) "Understanding the evolution of functional redundancy in metabolic networks," *Bioinformatics*, 34(17), pp. i981–i987. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTY604.

Schellenberger, J. *et al.* (2011) "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0," *Nature Protocols 2011 6:9*, 6(9), pp. 1290–1307. Available at: https://doi.org/10.1038/nprot.2011.308.

Shatsky, M. *et al.* (2016) "Quantitative tagless copurification: A method to validate and identify protein-protein interactions," *Molecular and Cellular Proteomics*, 15(6), pp. 2186–2202. Available at:

https://doi.org/10.1074/MCP.M115.057117/ATTACHMENT/C80E37AE-D604-4188-B53A-F774C1B5DCFB/MMC1.ZIP.

Simpson, B.W. and Trent, M.S. (2019) “Pushing the envelope: LPS modifications and their consequences,” *Nature Reviews Microbiology 2019 17:7*, 17(7), pp. 403–416. Available at: https://doi.org/10.1038/s41579-019-0201-x.

Sivaprasad, A. v. *et al.* (1992) “Coexpression of thecys E andcys M genes ofSalmonella typhimurium in mammalian cells: a step towards establishing cysteine biosynthesis in sheep by transgenesis,” *Transgenic Research 1992 1:2*, 1(2), pp. 79–92. Available at: https://doi.org/10.1007/BF02513025.

Suthers, P.F., Zomorrodi, A. and Maranas, C.D. (2009) “Genome-scale gene/reaction essentiality and synthetic lethality analysis,” *Molecular Systems Biology*, 5. Available at: https://doi.org/10.1038/msb.2009.56.

Szklarczyk, D. *et al.* (2011) “The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored,” *Nucleic Acids Research*, 39(suppl_1), pp. D561–D568. Available at: https://doi.org/10.1093/NAR/GKQ973.

Thesis-122010053.pdf

ORIGINALITY REPORT

11%	7%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to SASTRA University Student Paper	4%
2	Submitted to University of Hong Kong Student Paper	1%
3	academic.oup.com Internet Source	1%
4	Submitted to Indian Institute of Technology, Madras Student Paper	<1%
5	www.nature.com Internet Source	<1%
6	link.springer.com Internet Source	<1%
7	Pratapa, Aditya, Shankar Balachandran, and Karthik Raman. "Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks", Bioinformatics, 2015. Publication	<1%

8	Aditya Pratapa, Shankar Balachandran, Karthik Raman. "Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks", Bioinformatics, 2015 Publication	<1 %
9	ouci.dntb.gov.ua Internet Source	<1 %
10	Hongyu Chen, Xiaoqin Lai, Yihan Zhu, Hong Huang, Lingyan Zeng, Li Zhang. "Quantitative proteomics identified circulating biomarkers in lung adenocarcinoma diagnosis", Research Square Platform LLC, 2022 Publication	<1 %
11	vtechworks.lib.vt.edu Internet Source	<1 %
12	www.scribd.com Internet Source	<1 %
13	Michael Carter, Sophie Casey, Gerard W. O'Keeffe, Louise Gibson, Louise Gallagher, Deirdre M. Murray. "Maternal Immune Activation and Interleukin 17A in the Pathogenesis of Autistic Spectrum Disorder and Why It Matters in the COVID-19 Era", Frontiers in Psychiatry, 2022 Publication	<1 %
14	academic.hep.com.cn Internet Source	<1 %

15	tudr.thapar.edu:8080 Internet Source	<1 %
16	acr.iitm.ac.in Internet Source	<1 %
17	genomebiology.com Internet Source	<1 %
18	Researchonline.lshtm.ac.uk Internet Source	<1 %
19	doaj.org Internet Source	<1 %
20	ibse.iitm.ac.in Internet Source	<1 %
21	ir.uitm.edu.my Internet Source	<1 %
22	dspace.plymouth.ac.uk Internet Source	<1 %
23	nature.com Internet Source	<1 %
24	Jennifer Campbell, Atul K. Singh, John P. Santa Maria, Younghoon Kim et al. " Synthetic Lethal Compound Combinations Reveal a Fundamental Connection between Wall Teichoic Acid and Peptidoglycan Biosyntheses in ", ACS Chemical Biology, 2010 Publication	<1 %