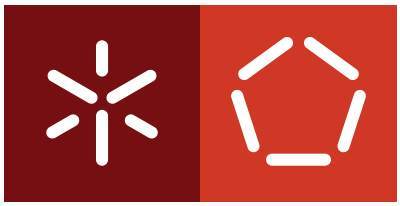
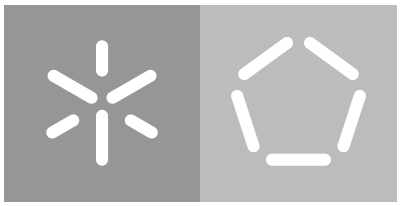


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Marta Lopes Gomes

In silico characterization of microbial communities interaction in soil samples

December 2019



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Marta Lopes Gomes

In silico characterization of microbial communities interaction in soil samples

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Oscar Manuel Lima Dias
Ulisses Nunes da Rocha

December 2019

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.
I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ACKNOWLEDGEMENTS

It would not have been possible to carry out this work without the valuable contribution of a large number of people. It is with great appreciation that I thank all of them for their dedication, their interest, but above all their support.

I start by thanking my supervisors, Professor Oscar Dias and Ulisses Rocha, for introducing me to this project, for all the advices, teachings, and words of motivation when I needed them most.

I am especially grateful to João Saraiva, who helped me at every step, listened to my doubts and ideas and always guided me in the best way.

To all the Microbial Systems Bioinformatics team residing at the Helmholtz Centre for Environmental Research, thank you for welcoming me and making me feel part of the group during my stay in Leipzig. For letting me know about their spectacular work and for listening to my ideas by actively contributing. All this experience has greatly valued this work but has essentially contributed to my personal development.

I thank all the teachers of the Master in Bioinformatics who, besides training me with all the necessary skills for the development of this project, have awakened in me the passion for this field.

I also thank my colleagues in the master's degree and all my friends for their support and company during this journey.

And of course, a big and special thank you to my parents for all the effort they have made over the years to allow me to follow my passions without restrictions. Their support has made everything I've achieved seem easier.

I also thank my boyfriend and sister for supporting me unconditionally at all times, for encouraging me to leave my comfort zone and pursue what will make me happier and more fulfilling in my professional career.

Thank you all very much!

ABSTRACT

Microbial communities, besides its many applications, can represent a solution for pollution problems with reduced costs. However, to explore them in our favor, it is necessary to understand how they work and be able to infer their potential regarding specific metabolic networks.

Because of the continuous growth of genomic data, various tools have been developed for homology and metabolic pathway inference, however new and improved strategies and algorithms still being required.

In this work, it has been developed a pipeline that makes use of clusters of orthologous data to perform the annotation of unknown sequences, and after that, the prediction of species' functional potential and microbial interactions. For that were developed two tools, OrtScraper, for the download of bulk organized data from specif pathways of interest, and OrtAn that performs the annotation on clusters of orthologous groups. The test and evaluation of the pipeline were focused on the well-known transformation of benzoate to acetyl-CoA (BTA) pathway. Two different genome sets were used, set A, from whose the annotation of the sequences was known, and set B, from whose the capacity regarding the benzoate degradation was known.

Both tools successfully performed the desired goal and for the annotation, the best cases presented an F_1 score over 0.90. The recall values of the annotation showed to be the weakest point of this pipeline, which led, possibly, to the unsatisfactory results on the prediction of the species functional potential.

Some improvements to the developed tools and pipeline were proposed to improve the annotation and species functional potential inference.

Keywords: Clustering; Orthologous; Homology; Annotation; Microbial Communities; Functional Potential.

RESUMO

As comunidades microbianas, além das suas várias aplicações, podem representar uma solução, de custos reduzidos, para problemas de poluição. No entanto, para explorá-las a nosso favor, é necessário entender como funcionam e poder inferir seu potencial em relação a redes metabólicas específicas.

Devido ao crescimento contínuo dos dados genómicos, várias ferramentas têm sido desenvolvidas para a inferência de homologia e de vias metabólicas, no entanto, estratégias e algoritmos novos e melhorados ainda são necessários.

Neste trabalho, foi desenvolvida uma pipeline que faz uso de clusters de ortólogos para a realização de anotação de sequências desconhecidas e, posteriormente, a previsão do potencial funcional das espécies e previsão de interações microbianas. Para isso foram desenvolvidas duas ferramentas, o OrtScraper, para o download de dados em massa organizados pertencentes a vias metabólicas de interesse, e o OrtAn, que realiza a anotação a partir de clusters de ortólogos. O teste e a avaliação da pipeline foram focados na bem conhecida transformação do benzoato em acetil-CoA (BTA). Foram utilizados dois conjuntos de genomas diferentes, o conjunto A, de onde se conhecia a anotação das sequências, e o conjunto B, de onde se conhecia a capacidade de degradação do benzoato.

Ambas as ferramentas realizaram com sucesso o objetivo desejado e, para a anotação, os melhores casos apresentaram pontuação F_1 acima de 0,90. Os valores de *recall* da anotação mostraram-se o ponto mais fraco desta pipeline, o que levou, possivelmente, aos resultados insatisfatórios na previsão do potencial funcional das espécies.

Foram propostas algumas melhorias nas ferramentas e pipeline desenvolvidas para melhorar a anotação e a inferência do potencial funcional das espécies.

Palavras-chave: Clustering; Ortólogos; Homologia; Anotação; Comunidades Microbiais; Potencial Funcional.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BEF	Biodiversity Ecosystem Functioning
BSS	Biosynthetic Support Score
BTA	Benzoate to Acetyl-CoA
cFBA	community Flux Balance Analyses
COG	Clusters of Orthologous Groups
ConOG	Consistent Orthogroups
DivOG	Divergent Orthogroups
EC	Enzyme Commission
FBA	Flux Balance Analyses
FN	False Negative
FP	False Positive
GPR	Gene-Protein-Reaction
IP	Integer Programming
KO	KEGG Orthology
LP	Linear Programming
MCI	Metabolic Complementary Index
MCL	Markov Clustering Algorithm
MCMC	Markov Chain Monte Carlo
TN	True Negative
TP	True Positive

1

INTRODUCTION

1.1 MOTIVATION AND GOALS

Over the years scientists have been trying to find the best way to deal with pollution problems. An effective approach with reduced costs for the treatment of polluted ecosystems is the use of microbial communities that can take advantages of industrial residues (like benzoate that can be used as a carbon source for some bacterial species)(B, 2012; Fetzer et al., 2015). However, to use microbial communities for solving these problems, it is essential that community structure and functional capabilities are understood.

In the last years, genomic data for single microorganism species or microbial communities (metagenomic) has been accumulated, mostly due to the new technologies of sequencing such as next-generation sequencing technologies (SHOKRALLA et al., 2012). However, new and improved algorithms and frameworks that enable the efficient analyses of a large amount of data are still required (Stephens et al., 2015; Schatz, 2012).

This work is included on a larger project which aims to improve our understanding of the potential response of microbial communities to introduced chemicals whose main objective is the design of scientifically based policies aimed at preventing and halting the loss of ecosystem services.

The main goal of this work is the development of a bioinformatics tool that allows characterizing metabolic networks of interest in microbial communities present in soil samples. These networks allow determining, in a specific environment, which interspecies interactions allow performing a given task. This tool will allow determining the microbial communities with the potential to degrade a given chemical compound. In the state of the art section, we will discuss current approaches used for homology inference (functional annotation), metabolic networks reconstruction/inference and microbial interactions from genomic information.

Afterwards, a bioinformatics tool will be developed with the most suitable framework. Finally, the benzoate degradation network of a well-known set of microorganisms will be used as a case of study, to evaluate the tool and developed pipeline.

1.2 STRUCTURE OF THE DOCUMENT

This document is organized in the following parts:

Chapter 2
State of the art
Role of genomics in the world of the big data. The soil and the importance of microbial communities to its health. Overview of different homology and clustering tools. Strategies and tools used for metabolic network inference and the prevision of the potential microbial interactions.

Chapter 3
Proposed Approach
Brief explanation of the proposed approach to reach the project goals.

Chapter 4
Methods
Overview of the data collected to test and evaluate the pipeline. Description of: the tool chosen for the clustering step and its evaluation; the annotation strategy and its evaluation; method for retrieve species functional potential and microbial interactions.

Chapter 5
Tools and Workflow
Description of the developed tools, to assist the pipeline on the data collection phase and annotation phase.

Chapter 6
Results and Discussion
Presentation and discussion of the results from the clustering evaluation and the various pipeline steps. Discussion of the tool's performance.

Chapter 7
Conclusions and Future work
Main conclusions of the thesis results and description of possible improvements and further work.

2

STATE OF THE ART

2.1 BIG DATA

When compared with the other major generators of Big Data (astronomy, YouTube, and Twitter), genomics is on par with these domains in terms of the acquisition, storage, distribution, and analysis requirements (Stephens et al., 2015).

Concerning the acquisition of data, the advances in next-generation sequencing technologies revolutionized this field and even made possible the emergence of new ones, like metagenomics, that corresponds to the analysis of environmental DNA (and made possible the sequencing of microorganisms that, outside of their natural environment have yet to be cultured) (SHOKRALLA et al., 2012). Considering the current rate growth on this field, Stephens et al. (2015) predict that by the year 2025 the mark of 1 zetta-bases/year of genomic data will be achieved. In fact, they estimate the sequencing of 100 million to 2 billion human genomes by this year, which would lead to an exceeding growth when compared to the other Big Data domains.

With respect to data analyses, genomics appears to be the most challenging domain. A very important and common operation, like a whole genome alignment between human and mouse, consumes approximately 100 CPU hours (Kurtz et al., 2004). To perform the whole genome alignments between all the pairs of species available in 2025, the resources allocated should allow performing these operations in a magnitude of six times faster than what is possible today (Stephens et al., 2015). These challenges must be faced by knowledge professionals of the field (quantitative biologists, bioinformaticians, and computer scientists and engineers) (Schatz, 2012) and efficient solutions (both in terms of hardware and software) and algorithms for the different problems should be created so the analysis of the available data is made possible.

2.2 SOIL

In 1996 Doran et al. (1996) defined a healthiness of a soil as the ”capacity of soil to function as a vital living system to sustain biological productivity, maintain environmental quality, and

promote plant, animal, and human health". Ensuring a healthy soil is ensuring the quality of the water we drink, of the air surrounding us and of the food we produce (Wall and Six, 2015). To maintain the benefits of the soil, this resource should be used with care and in a sustainable way since human actions could have a very negative effect on soil and the biodiversity of its communities (Wall et al., 2015).

A threat to this ecosystem, for instance, is the extensive use of chemical substances. When some chemical substances produced in between industrial processes cease to have value and become unwanted, its improper elimination could lead to possible environmental contamination. (B, 2012; Harvey et al., 2017). This is a common problem in industrial cities distributed around the world (Filippelli et al., 2015). UN-HABITAT and Ltd. (2011) show that there was an increase on the urban settings from less than 30% in 1950 to 47% in 2000 and, by 2025, is expected to increase to 60% of the percentage of urban dwellers. Thus, it is important to address this problem.

One solution for the process of soil decontamination is the use of appropriated microbial communities, as microorganisms are capable of degrading both natural and synthetic substances as a means to obtain energy and nutrients (B, 2012). Already in the 50's, the 'microbial infallibility hypothesis' was proposed, which suggests that microorganisms will be found to degrade every chemical substance synthesized by any living organism. To date, many studies (Krueger et al., 2015; Ayangbenro and Babalola, 2017; dos Santos and Maranho, 2018) point single microorganisms, as well as microbial communities, as being an effective biotechnological approach to help in pollution problems.

2.3 MICROBIAL COMMUNITIES

2.3.1 *Microbial communities as a solution for pollution problems*

Over the years, microbial communities have been pointed out as a solution for pollution or contamination problems in various ecosystems. dos Santos and Maranho (2018) discuss bioremediation as an alternative tool for recovery of petroleum-contaminated soils, focusing on a phytoremediation strategy where roots and colonies of microorganisms work together for the biodegradation of petroleum. Jiang et al. (2017) show that there are microbial communities capable of removing propazine (an s-triazine herbicide) residues from farmland soil. This process is important to ensure a safe crop production. Another study referring to bioremediation of herbicides with microorganisms is the study of Horemans et al. (2016). They showed that some bacteria were responsible for the degradation of the phenylurea herbicide linuron on agricultural soils as well as bacterial populations capable of mineralizing the downstream metabolites of linuron hydrolysis.

Debbarma et al. (2018) studied potential bacterial strains for electronic waste (e-waste) treatment, providing a protocol for screening and selection of efficient e-waste utilizing bacteria and demonstrating potential consortia ready to be used. Krueger et al. (2015) have written a review gathering the existing knowledge about the use of microbial communities on the degradation of several plastic types that pollute marine, limnic and terrestrial ecosystems. They identify microorganisms as promising candidates for bioremediation of environmental plastics. Ayangbenro and Babalola (2017) describe groups of microorganisms with biosorbent potential for heavy metal removal from the environment.

Fernández-Luqueño et al. (2011) suggested microorganisms as an effective and economical solution for PAHs (polycyclic aromatic hydrocarbons) contaminated soils and refer to better strategies to improve this process. Liu et al. (2017) investigate microbial communities related to roxarsone degradation and identified the bacteria that played important roles in this process.

2.3.2 *Microbial communities behavior and interactions*

As indicated in the section above, microbial communities can be a solution for soil pollution, but to take advantage of them or at least, recognize their capacities and limitations, we first must understand their behavior.

Inclusive fitness theory introduced by Hamilton (1964) proposes that cooperation should be common between organisms sharing the same genotype (genetically distinct set of microbial cells, in particular, groups of cells that are identical at the loci for a social phenotype) (Mitri and Richard Foster, 2013). Furthermore, this theory also indicates that organisms with different genotypes may be in competition. But these assumptions may not always be the case (West et al., 2006).

In Sieuwerts et al. (2008), the authors explain microbial interactions in mixed cultures. The classification of these interactions is based on the fitness consequences to the effector, who performs the behavior, and the target, that is affected by the behavior of the effector. See table 1. These behaviors can be divided into six main classes: mutualism, parasitism, competition, commensalism, amensalism, and neutralism. In the first two referred classes, the effector of the behavior benefit from the interaction. With respect to mutualism, both microorganisms behave like effector and target at the same time, so both benefit from the interaction. In the case of the parasitism, the target suffers detrimental effects. In competition, both microorganisms involved act like effector and target, like in mutualism, but in this case, both suffer detrimental effects (like when two species compete for the same limited carbon source or other nutrients). In the last classes, the effector is not affected for the behavior but it could be beneficial for the target microorganism (commensalism), detrimental (amensalism) or neutral too (neutralism).

Table 1.: **Classification of microbial interactions based on the effect on the participant's microorganisms fitness.** (Sieuwerts et al., 2008)

		Effect on effector		
		Beneficial	**Detrimental**	**Neutral**
Effect on target	**Beneficial**	Mutualism		Commensalism
	Detrimental	Parasitism	Competition	Amensalism
	Neutral			Neutralism

It is easy to understand why natural selection led to the cases where the effector benefit from the interaction, but the situations where only the target microorganism take benefit from it are more difficult to explain. Nevertheless, a wide range of known microbe phenotypes is consistent with this altruistic behavior (West et al., 2007). Examples are the secretion of extracellular enzymes that digest compounds making them accessible for other organisms (Wandersman, 1989) and siderophores, which allow cells to harvest poorly soluble iron (Pattus and Abdallah, 2000).

If different organisms can be cooperative and live within a community, another factor that could be taken into account when trying to understand the ecosystem behavior is the biodiversity. There are studies suggesting that the biodiversity of many ecosystems is decreasing (Butchart et al., 2010) but the consequences of these changes still remain unclear (Naeem, 2002).

Frequently, species abundance and variety have a positive effect on the ecosystem behavior, but there are other factors that could have an impact on the community. The biodiversity-ecosystem functioning (BEF) relationship could be affected by the number and type of species, their evenness in the community, their functional attributes or abilities, and their interactions (Maestre et al., 2012; Mulder et al., 2001). For instance, to maximize the multi-functionality of a community, Maestre et al. (2012) indicated that particular combinations of attributes may be required. The environmental context should also not go unnoticed, as ecosystems are constantly facing multiple environmental changes that could lead to negative effects on the BEF (Cardinale, 2011). The stability of an ecosystem could be characterized by its resistance ("the degree to which microbial composition remains unchanged in the face of a disturbance" (Allison and Martiny, 2008)) and resilience ("the rate at which microbial composition returns to its original composition after being disturbed" (Allison and Martiny, 2008)).

Fetzer et al. (2015) showed that high biodiversity could benefit the community when facing environmental changes because some species could have some relevant traits or allow new interactions which would allow the community to survive. However, Pfisterer and Schmid (2002) suggested a negative correlation between biodiversity and stability.

Hence, it is important to know which are the functional capabilities of the species within a community to understand the functioning of an ecosystem and its potential.

2.4 HOMOLOGY

The need for computational tools to analyze the vast amount of data being generated was addressed in the section 'Big Data'. Regarding protein sequence data, an important step is to infer the proteins' functions via homology (common evolutionary origin) analyses (Mazumder et al., 2008).

Genes that descent from a common ancestral DNA (deoxyribonucleic acid) sequence are called homologous genes. Depending on their evolutionary relation, they can be either orthologous or paralogous. Orthologues are genes from different species that originated through speciation events whilst paralogues are genes within the same genome that originated through duplication events. Unlike paralogous genes that evolve into new functions, orthologous genes are more likely to share the same function (Fitch, 1970). In addition to these additional terms were suggested for a more specific classification when determining the evolutionary relationship: out-paralogs (not orthologous because these precede a speciation event), and in-paralogs (orthologous as these were duplicated after the speciation event) (Remm et al., 2001). Precise clusters of orthologous groups (COGs) (sets of genes/proteins able to perform the same function) are useful for having best results when predicting functions. Thus, several tools, able to assemble COGs, have been successfully developed for comparative genomics and genome annotation.

2.4.1 *Homology inference*

The sequence and/or structural similarities between proteins allow the inference of homology. Here we describe approaches and tools used for homology inference based on sequence similarities. A simple and very common way to compare proteins and search for similarities in a vast set of sequences in a database is the use of Basic Local Alignment Search Tool (BLAST)(Altschul et al., 1990). The BLAST algorithm is described in figure 1.

For the analyses of the BLAST results, it is necessary to understand some concepts and vocabulary regarding the performed alignments. These concepts and vocabulary are used not only for BLAST reports but for many other sequence search tools that follow similar strategies. Some of them will be referred to below.

Alignment is a process of trying to match up the nucleotides or amino acid residues of two (or more) sequences. The goal is to achieve maximal levels of identity to assess the degree of similarity and the possibility of homology between two sequences. There are two main types of alignments, the global ones (that attempt to align the whole sequences involved) and the local (that try to find local regions of high similarity between two sequences) (Henikoff and Henikoff, 1992), the ones that BLAST attempts to do.

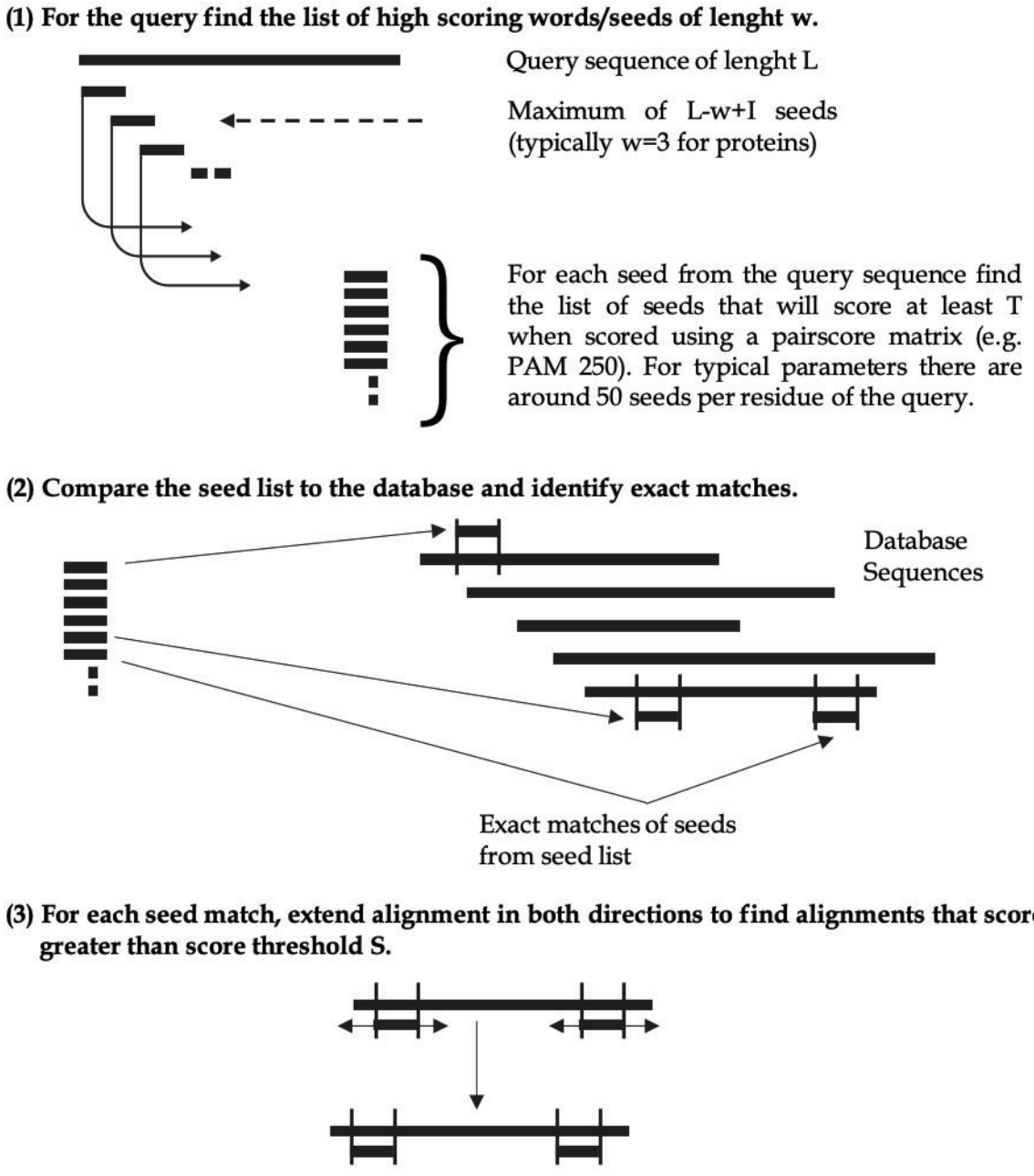


Figure 1.: **Schematic illustration of the BLAST algorithm.** Adapted from Sansom (2000).

;

In an alignment between two sequences, there are, the normally called, query sequence and subject sequence. The subject, or reference sequence, is a sequence present in the database and to which some information is known, for instance, its origin and functional annotation. The query is the sequence for which is attempted to find a match in the database, that is, a similar reference sequence. The word ”hit” is used when, for a query sequence, it is found a match in the database (Altschul et al., 1990, 1997b; Wheeler and Bhagwat, 2007).

An example of a hit/alignment resulting from BLAST is represented in figure 2. The range of alignment is the length of the section of both sequences involved in the alignment. In the range of alignment there are identical matches (represented by a ”|” in figure 2 and correspond to the positions in the alignment where the nucleotide or amino acid is the same in both query and subject sequences), the mismatches (correspond to positions in the alignment were the nucleotide or amino acid is not the same between both sequences) and gaps (represented by a ”-” in the figure 2 and corresponds to a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another) (Fassler and Cooper, 2011; Kerfeld and Scott, 2011).

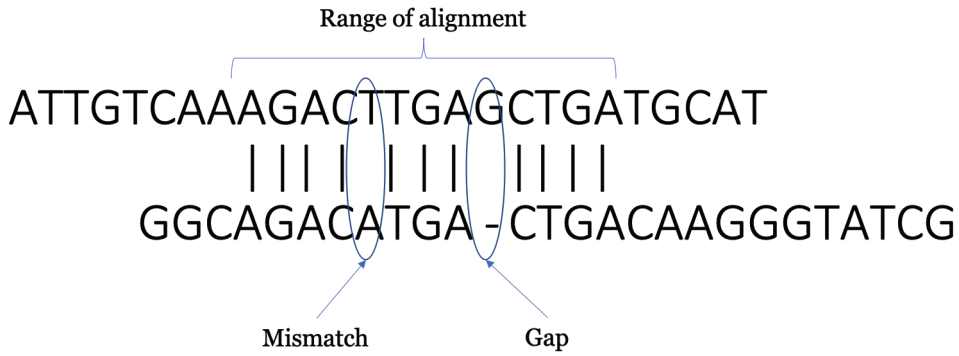


Figure 2.: **Scheme representing an example of a local alignment.** Adapted from Fassler and Cooper (2011).

To evaluate the similarity between two sequences a score is calculated, where the matches between the same nucleotide or amino acid contribute positively to the score and the miss matches and gaps contribute negatively or do not contribute. These values for the different cases of matches and miss matches are defined by a scoring matrix. This matrix contains the values that should be assigned to each possible case. They are constructed by the observations of larger samples of verified pairwise alignments and indicate the probability of, for instance, an amino acid i be substituted by an amino acid j. With this arises a new parameter that can be evaluated in the alignments, the percentage of positive matches. These are matches that have a positive value on the score matrix used to calculate the score of the alignment. These cases include identical matches, that always have positive values, and matches between

similar amino acids (Henikoff and Henikoff, 1992; Fassler and Cooper, 2011; Kerfeld and Scott, 2011).

Some other parameters that can be taken into account, besides not always retrieved directly by these tools, are the query and subject coverage. These terms correspond to the percentage of the query or subject sequence that is involved in the alignment. Another very important parameter is the Expected value (usually called e-value) that represents the likelihood of the present alignment scores or higher occurs by chance in the database. E-value is mostly important when using big databases, because, in those cases, it becomes more likely to be included matches to a query sequence that is due by change and not for homology (Kerfeld and Scott, 2011; Korf et al., 2003).

Dayhoff et al. (1975) was the first to suggest the concept of protein families (or family domains), that consists in sets of proteins grouped by similarity. Several databases store protein information based on this homology concept, such as Pfam (Finn et al., 2014) (protein families database), ProDom (Servant et al., 2002) (Protein domain families database) and PROSITE (Sigrist et al., 2013) (Database of protein domains, families and functional sites). Currently, these families (or family domains) are usually represented by multiple alignments (Mazumder et al., 2008) and there are bioinformatics representations, such as position-specific scoring matrices (PSSMs) (Gribskov et al., 1987) and hidden Markov models (HMMs) EDDY et al. (1995), wich allow the comparisons to be much more sensitive than a simple search with the BLAST algorithm.

Thus, when BLAST searches are not enough to predict a protein function, advanced sequence analyses can be performed such as profile searches (HMM and PSSM), pattern search (conserved motif analysis) and phylogenetic tree reconstruction (Mazumder et al., 2008). One of these tools is the Specific Iterative (PSI)-BLAST (Altschul et al., 1997a). PSI-BLAST is a protein sequence profile search method that generates a PSSM from a multiple alignment generated by the hits of a first run of the BLAST. In the following iterations, the search in the database is performed with the generated PSSM as a matrix of scores. The PSSM captures the conserved patterns in the multiple alignment. This allows the detection of distant relationships between proteins.

Although BLAST is an efficient algorithm, with the constant increase of genomic data the necessity of new algorithms and tools, to decrease computational time and resources whilst maintaining precision, arises.

RAPSerach2 (Zhao et al., 2012), an optimized version of RAPSerach (Ye et al., 2011), was developed with the aim of analyzing large amounts of sequences generated by transcriptomic or metagenomics samples, inferring putative functions by similarity searches. This algorithm follows the same approach as BLAST regarding the seed-and-extend paradigm, with the difference that RAPSearch2 uses a reduced amino acid alphabet (10 symbols representing groups of amino acids) when looking for flexible-length seeds (or ”words” in figure 1). With

an optimization regarding the way of the index of the protein database, RAPSearch2 became more quickly and memory efficient than RAPSearch, which was already faster than BLAST (achieving up to a 90X acceleration) while missing less than 5% of potential protein hits.

GHOSTX (Suzuki et al., 2014) is another sequence homology search tool developed for functional annotation of metagenome sequences. A workflow that summarizes the GHOSTX is shown in figure 3. It follows the seed-extension approach used in BLAST, also with the difference of flexible-length seeds and, to speed up the process, it uses suffix arrays of both queries and database sequences. After finding seeds, with a method relying on a score-based optimal seed length, GHOSTX performs alignments by extending seeds without gaps. Lastly, it makes alignments with gaps. With the optimization of the seed search step (one of the most computationally intensive parts of BLAST) GHOSTX can be faster than BLAST searches with similar levels of sensitivity. Compared to RAPSearch2, GHOSTX can also be faster (up 1.4 times) and achieve higher levels of accuracy.

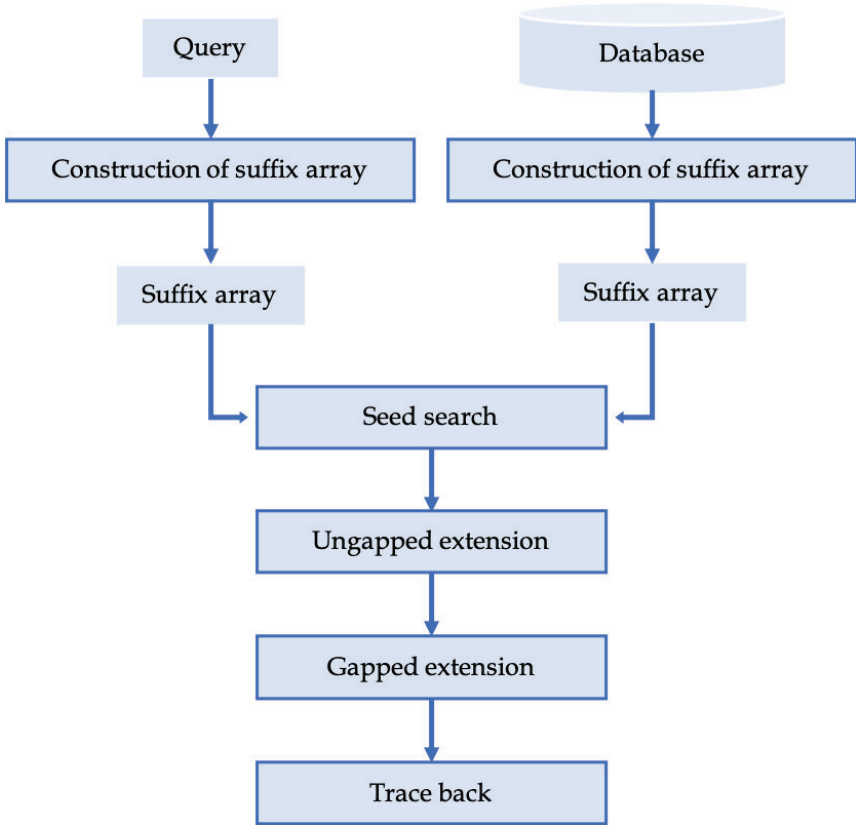


Figure 3.: **The workflow of GHOSTX.** Adapted from Suzuki et al. (2014).

DIAMOND (Buchfink et al., 2015) is another tool inserted in this group of tools that uses the seed-and-extend paradigm that is described in figure 1 (when explaining the BLAST algorithm functioning). The DIAMOND algorithm only works with protein sequences and

claims to be 20,000 times faster than BLAST aligning short reads while maintaining a similar level of sensitivity. Regarding the seed-and-extend strategy, it consists of a first phase where there is a search for matching seeds in the database, and then, the second phase is responsible for the 'extend' or alignment between the sequences. This requires the programs to store the index of the seeds found in the sequences from the database. The usual approach after that is to scan the query sequences linearly and match their seeds to the seeds from the reference sequence accessing then the created index in a randomly way. The way DIAMOND approaches up this process is by the used of double indexing, where all the seeds, from both queries and reference sequences, have their seeds and locations listed. Both lists are sorted and passed through at the same time to determine all the seeds that match and their locations. This strategy reduces the demands on the main memory bandwidth. To reduce the time compared to the alignment search tools referred to before, DIAMOND made other modifications to the strategy normally used. Normally the seeds used are single consecutive short seeds because the use of longer seeds causes a decrease in the sensitivity. However, short seeds slow down the computation (Ma et al., 2002). What DIAMOND does to increase speed and do not lose sensitivity is to use spaced seeds, having the positions of these a specific number and layout carefully chosen to fulfill the needs (Ilie et al., 2011). Besides that, DIAMOND also uses a reduced amino acid alphabet but composed of 11 letters instead of the 10 symbols used by RAPSearch2. It was already shown that the information lost when using a reduced alphabet that doesn't go from less than 10 letters is very little (Murphy et al., 2000) and does not compromise the homology search. Thus, using a reduced alphabet, both RAPSearch2 and DIAMOND can perform the comparisons faster without compromising too much sensitivity. DIAMOND also uses simple exact match criteria to decide which seeds are passed to the extension phase, which consists of a computation of a Smith-Waterman alignment (Smith and Waterman, 1981). DIAMOND achieved impressive results in terms of the time needed to compare big quantities of sequences. However, to use the default or "fast" mode of DIAMOND, that can be 20,000 times faster than BLAST, is necessary to abdicate some sensitivity. Only the DIAMOND "sensitive" mode, 2,000 times faster than BLAST, is capable of achieving sensitivity results close to the ones obtained with BLAST (the loss of recovered matches is less than 10%)(Buchfink et al., 2015).

The homology search problem can also be seen as a statistical inference problem underlying two hypothesis: First, the target sequence is a homologue of the query sequence. Second, the null hypothesis, the target sequence is a "random" (not homologous) sequence (Eddy, 2009). Hereupon, tools for homology inference were developed based on this strategy.

HMMER is a homology search tool based on probabilistic inference using profile HMMs in its implementation. This tool could be used for single sequence search (e.g. BLAST) or for iterative/profile search (e.g. PSI-BLAST), normally using profile databases such as

Pfam. HMMER can work at the same speed as BLAST while being able to detect remote homologues, relying on the strength of its underlying probability models (Eddy, 2009).

Lately, researchers are trying to take advantage of the promising field of machine learning in genetics and genomics (Libbrecht and Noble, 2015). Zou et al. (2017) present an ensemble learning framework called EnMIMLNN that uses RBF (radial basis function) neural networks (Bishop, 1995) which are able to learn from Multi-Instance Multi-Label examples (objects used to train have multiple labels) to address the prediction task. For the same prupose, but in the branch of deep learning, Zou et al. (2017) propose deep restricted Boltzmann machines (DRBM). The two approaches present promising results; however, these are limited to GO (Gene Ontology) terms (Ashburner et al., 2000) classification.

2.4.2 *Clustering of orthologous tools*

As previously mentioned, clustering of orthologues tools could be advantageous in the process of functional annotation. Examples of algorithms that perform best in identification of orthologues, whilst maintaining the balance of sensitivity and specificity (Hulsen et al., 2006; Chen et al., 2007), are INPARANOID (Remm et al., 2001) and OrthoMCL (Li et al., 2003).

INPARANOID is able to distinguish orthologous and in-paralogous genes from out-paralogous genes, though this approach is limited to comparisons between two species. The algorithm starts with the detection of sequence pairs with mutually best hits between the two species in the analysis, which result from all-versus-all BLAST searches. The matched area from the BLAST result is forced to be longer than 50% of the longer sequence, thus avoiding clustering sequences that share only short domains. After the definition of a cluster and its main sequences, additional orthologous genes are added to the cluster. In-paralogous genes are only assigned if their sequence is more similar to the main orthologous gene than to any sequence from other species. In the end, to solve overlapping cases, depending on the type and extent of the overlap, overlap groups can be merged, deleted or separated (Remm et al., 2001).

To overcome INPARANOID's limitations, OrthoMCL was developed. Unlike the former, OrthoMCL allows performing the identification of orthologous groups between multiple species (Li et al., 2003). This approach is similar to INPARANOID, though it uses the Markov Cluster algorithm (MCL; based on probability and graph theory and allows simultaneous classification of global relationships in a similarity space) (van Dongen, 2000) which allows solving the problem relative to multi-genome comparisons. In figure 4, a workflow that summarizes the OrthoMCL algorithm approach is shown.

Emms and Kelly (2015) describe a problem common to tools such as INPARANOID and OrthoMCL that use BLAST to measure pairwise sequence similarity and lack the consideration of the length of the sequences in analyses. The problem is that in BLAST short

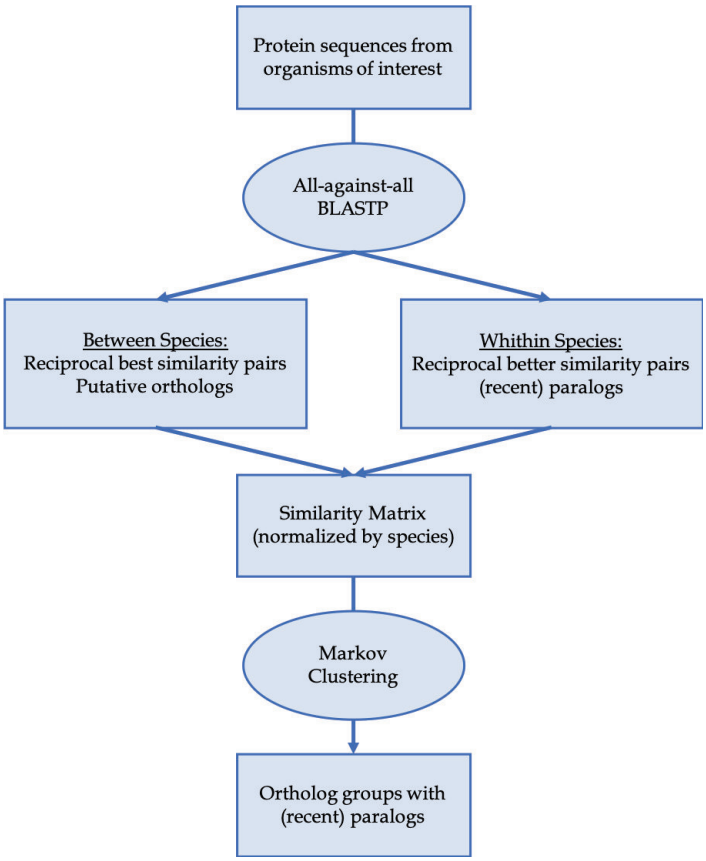


Figure 4.: **Flow chart of the OrthoMCL algorithm for clustering orthologous proteins.** Adapted from Li et al. (2003).

sequences alignments do not result in large bit scores or low e-values, while long sequences result in several hits with better scores than best hits involving short sequences. Thus, the results from these tools could contain orthologous groups missing short genes and orthologous groups with long genes that should not be clustered.

OrthoFinder (Emms and Kelly, 2015) aims to solve this problem. It uses an approach similar to OrthoMCL but includes a score transformation to eliminate gene length bias in the detection of orthologous groups. In figure 5 an overview of the steps of OrthoFinder algorithm is shown.

All-versus-all BLAST searches require intensive computational resources, and consequently, specially when using large data sets for clustering, leads to long running times (Li et al., 2012). To overcome this challenge, new methods for clustering sequences were developed, such as CD-HIT(Li and Godzik, 2006) and Uclust (Edgar, 2010). CD-HIT's first step is the ordering of the sequences by length, and then, set the longest sequence as the seed of the first cluster. After that, the remaining sequences are compared with the seeds of the existent clusters and the sequence in question is grouped into a COG if the similarity with the seed meets

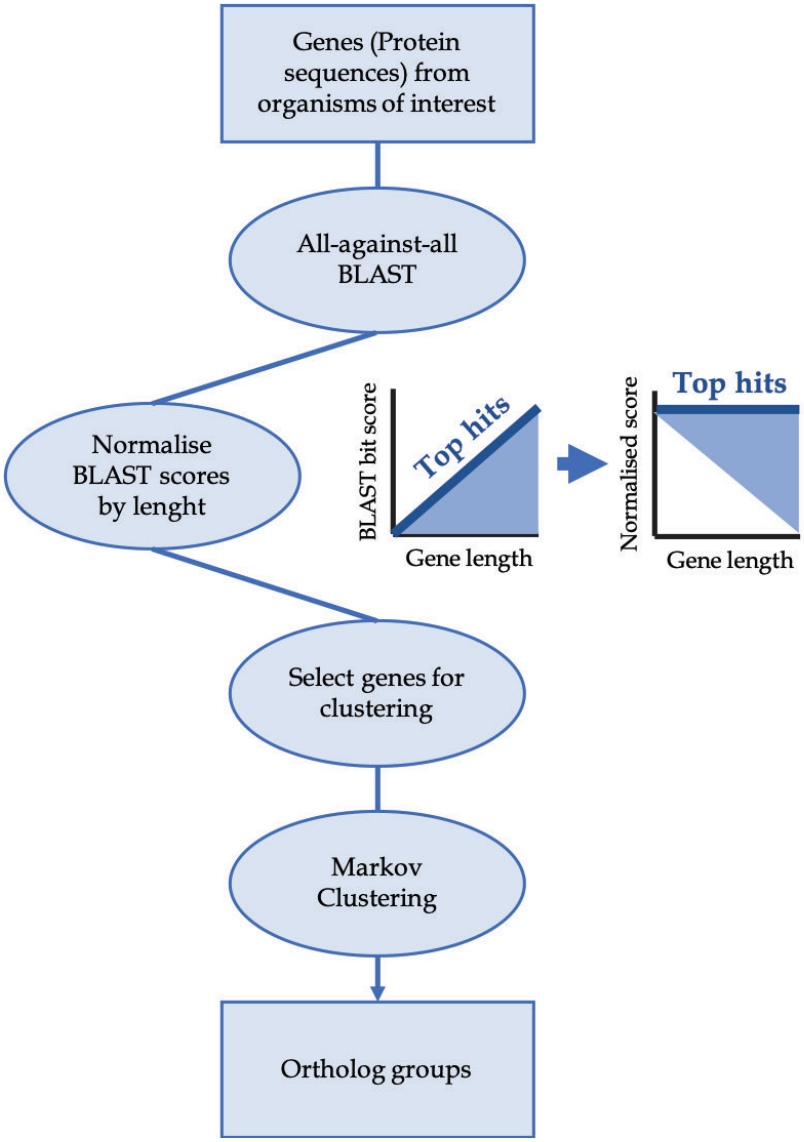


Figure 5.: **Flow chart of the OrthoFinder algorithm for clustering orthologous genes.** Adapted from Emms and Kelly (2015).

Table 2.: **Comparative table of some important features of the clustering tools.**

	INPARANOID	OrthoMCL	OrthoFinder	CD-HIT	Uclust
Multiple species allowed	No	Yes	Yes	Yes	Yes
Inparalogs detection	Yes	Yes	Yes	No	No
Need of all against all comparison	Yes	Yes	Yes	No	No
Comparisons and Measure of Similarities between sequences	BLAST	BLAST	BLAST (but includes a score transformation to eliminate gene length bias).	The similarities are estimated by common word counting.	Usearch
Clustering Algorithm	Uses bi-directionally best hits to find the main pair of each group, where additional orthologouss (or in-paralogs) are clustered latter.	MCL	MCL	Greedy incremental algorithm	Greedy incremental algorithm
Main advantages	High sensitivity and specificity when clustering orthologous.	High sensitivity and specificity when clustering orthologous.	High sensitivity and specificity when clustering orthologous; An easy command that uses as input a multiFASTA file (one per species); Minimizing the bias of the length of the sequences.	Ultrafast	Ultrafast
Main disadvantages	Needs a lot of dependencies including BLAST; Limited to two species.	Needs a lot of dependencies including BLAST and MCL algorithms to run; Difficult to use (a lot of commands needed).	Needs a lot of dependencies including BLAST and MCL algorithms to run.	Only highly similar sequences are grouped in the same cluster.	Only highly similar sequences are grouped in the same cluster.
Year of realease	2001	2003	2015	2001	2010

a pre-defined cut-off value. Unclustered sequences becomes the seed of a new clusters (Li et al., 2012). Uclust also follows a greedy incremental approach like CD-HIT, but for fast sequence comparison, it uses a heuristic called Usearch. It gains speed by comparing a few top sequences, the ones which have shorter words in common, instead of the full databases (Li et al., 2012; Edgar, 2010).

All the algorithms described above in this subsection are able to cluster sequences based on their similarity, but could lead to slight differences in their results. OrthoFinder is the most accurate between all the algorithms previously described followed by OrthoMCL which exhibits more potential for accurate functional annotation of unknown protein sequences compared to INPARANOID (Chen et al., 2007; Emms and Kelly, 2015). The referred algorithms are not specifically focused on finding orthologous genes, but in group similar sequences. They could be viable methods, for instance, to transform large redundant data sets into non-redundant data set ones (Li et al., 2012).

In the table 2 the tools previously mentioned are compared regarding some important features and their main advantages and disadvantages.

2.5 METABOLIC NETWORKS INFERENCE

Knowledge on which metabolic pathways present in an organism or microbial community allow to understand what can happen in a given environment. Therefore, following the functional annotation, one can perform metabolic network reconstruction / inference. Several strategies have been developed to determine which metabolic networks are present in a given community and databases containing metabolic networks data play a major role in this process. Two projects that contain pathways information that could be taken as reference when performing functional annotation and pathway reconstruction of large data sets are KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2017) and MetaCyc (Caspi et al., 2018) (Altman et al., 2013).

KEGG is a database with functional information about genes and genomes. In KO (KEGG Orthology), molecular function data is organized in functional orthologous groups. The KO identifiers, k numbers, are used across KEGG, including the KEGG pathway maps. These maps include diagrams of networks of molecular interactions/reactions.

MetaCyc contains a large curated collection of metabolic pathways, including information about reactions and involved components, enzymes and chemical compounds. Its data has been utilized for pathway prediction in the BioCyc database collection (Caspi et al., 2016).

The strategy commonly used when trying to perform the pathways reconstruction of a genome or metagenomics sample consists in the simple mapping of genes/proteins, based on their homology, to a database containing pathways information. Next, a pathway is considered to be present if one or more of the previewed functions in the pathway are identified. For instance, KAAS (KEGG Automatic Annotation Server) (Moriya et al., 2007) uses KEGG database to perform the annotation of a data set. It first finds the homologues of the query sequences and assigned them a k number (based on which KO group the gene belongs to) after which, the functions identified are mapped to the reference pathways (KEGG pathway maps). The output is a list of all the pathway maps with a link in which k numbers found are identified. According to Ye and Doak (2011), this simple approach could lead to a problem of over-estimation of the number of pathways, in which the same protein is linked to multiple pathways.

Hence, MinPath (Ye and Doak, 2011) was developed to address this issue. Unlike the previously mentioned approach for pathway reconstruction (given a set of functions, reconstruct the complete pathways encoded or identify the pathways that have at least one function associated), this approach is based on a parsimony problem.

The algorithm consists of a linear programming (LP) problem, in which all the variables assume integer values, making it an integer programming (IP) problem. The objective function of this called minimal pathway reconstruction problem is in equation 1. The algorithm aims at finding the minimal set of pathways that can be performed by all assigned functions

(inferred from the sequence data), that may come from complete or incomplete genomes or metagenomes. n is the number of functions annotated in a data set.

The minimal set of pathways that can be performed by all the given functions is composed by the pathways with P_j=1, where:

$$\begin{aligned} & min \sum_{j=1}^{p} P_j \\ s.t. \quad & \sum_{j=1}^{p} M_i j P_j \geq 1 \quad \forall i \in [1, n] \end{aligned} \tag{1}$$

p is the number of putative pathways which have at least one component function annotated;

M is the mapping of protein functions to the pathways, M_{ij}= 1 if function i is involved in pathway j, otherwise 0;

P_j indicates if the pathway j is on the final list or not (1 if selected, 0 otherwise);

Although MinPath uses a sensible approach for inferring pathways represented in a set of sequences, it still not perfect and there is room for improvements (Ye and Doak, 2011).

Jiao et al. (2013) proposed a probabilistic approach to infer the reactions available in a community. The approach uses a Markov Chain Monte Carlo (MCMC) algorithm for sampling potential and valid subnetworks (networks possible to occur in the community) taking as input the list of annotated reactions in the community and a global network to analyze. After the MCMC sampling, the probability of the occurrence of each reaction is calculated. This method takes into account a problem normally dismissed in pathway inference methods, the promiscuous enzymes. Promiscuous enzymes are those mapped to more than one reaction. However, given certain environmental conditions like pH and temperature, there is a higher likelihood of the enzyme to catalyze a given reaction than the others. Thus, when inferring about metabolic networks, it should not be assumed that all the reactions are catalyzed equally Nobeli et al. (2009). The algorithm also indirectly favors highly connected networks, reducing the number of terminal metabolites, a fact that is also taken into account when reconstructing metabolic networks (Feist et al., 2009).

However, this approach has limitations, such as the lack of consideration of compartmentalization (although other studies (Greenblum et al., 2012) also ignore these boundaries when studying a whole microbial community), the assumption that the reactions are reversible (which does not normally occur in the cellular environment) and not taking into account the enzymes abundance.

2.6 POTENTIAL OF MICROBIAL INTERACTIONS

Approaches used to determine the function of genome/metagenome sequences, followed by several strategies used to construct/infer the metabolic pathways from the annotated reactions have been described. Next, the prediction of potential interactions between microorganisms within a community will be discussed.

The approaches vary from mathematical modeling (Song et al., 2014) to text mining of scientific literature (Freilich et al., 2010). Regarding mathematical modeling, Song et al. (2014) describe in their review a wide range of approaches varying on their main focus, the necessary input, and in problem formulations. One way to classify these approaches is dividing them by the modeling unit that is considered in the problem, in other words, the entities of which the interactions are evaluated. These modeling units can be individual cells, species, functional guilds (groups of organisms sharing similar traits) or even a community as a whole. Supra-Organismal is the name of those approaches where the chosen modeling unit is the community. Here instead of treat a community as a set of species, the community is seen as a set of genes/reactions from which interactions are calculated. Also, cell boundaries are not considered in the problem. This technique is used in comparative metagenome analysis eliminating the need to identify genes origin species (Tringe, 2005). Two approaches possible to be used with the supra-organismal concept are Stoichiometric Model-Based Analysis and Metabolic Function-Based Dynamic Modeling, first developed for analyzing single organisms (Song et al., 2014). Stoichiometric models are given by the mass balances in conjunction with reactions flux boundaries. To apply the super-organismal concept in this approach, a metabolic network representative of the whole community has to be reconstructed. This networks could then be used with flux balance analysis (FBA) (Orth et al., 2010), a stoichiometric based-model that obtains an optimal pathway regarding biomass (or metabolites such as ATP) production through linear programming (LP) problem. While stoichiometric models evaluate the flux distributions of the community in a specific environment, dynamic modeling adjusts to the environment, so the response to environment variations could be studied. Dynamic models are very complex, so there are some strategies created with the goal of reducing this complexity. One is to focus only on some key metabolic functions for the network. An example following this strategy is Gene-centric approach (Reed et al., 2014) that also takes into account the dependencies of the network reaction on the functional genes and their dynamic responses.

Population-based models are the most used ones to study the communities dynamics. The entities or modeling units taking into account in these type of models are either species or functional guilds. These models assume homogeneity in the cells phenotypic behavior within a population. To include on the modeling the heterogeneity observed in populations, individual cells have to be the modeling units (Song et al., 2014).

To infer microbial interactions we can simply assume that, for instance, a species A has a positive effect on species B if this one grows better on the presence of species A. If species A is not affected by species B, this could be a case of commensalism. In the case of mutualism, both species have a positive effect on each other, and in the case of competition, this effect is negative (Faust and Raes, 2012). These basic interactions could be verified by analyzing the growth rates of these species alone and together in a specific environment. But doing this experimentally is difficult, so there's the necessity of theoretical tools capable of preview these interactions (Song et al., 2014).

It is possible to extract information about community composition and species abundance from metagenomic samples (Mande et al., 2012). Through this abundance data, microbial interactions are inferred based on the correlation of their abundance patterns. If the patterns are not correlated they are competitive species, but if the patterns are similar it could be a case of cooperation (e.g. mutualism) (Faust and Raes, 2012).

Stoichiometric models can also be used to preview interactions between species or functional guilds in an environment, in addition to flux distributions. A variant of the FBA approach, called community flux balance analysis (cFBA), is used to study microbial communities (Khandelwal et al., 2013). Zomorrodi and Maranas (2012) created a framework called OptCom implementing cFBA where both the community-level and individual cell-level are under optimization. Zomorrodi et al. (2014) also developed a dynamic version of OptCom, d-OptCom. However, these are complex problems and are difficult to execute with a large community where both the number of species/guilds or reactions are really big, making the memory requirements huge. Thus, they are limited to simple consortia. But there are a lot of other techniques to infer these relationships ranging from Nonlinear Regression (Elith and Leathwick, 2009) to Thermodynamically-Based Models (LAROWE et al., 2008) and Trait-Based Modeling (Boon et al., 2014) where the dynamic of the community is analyzed focusing on the traits as entities that mediate microbial interactions with each other and the environment.

Also, studies where interactions in a community were previewed with the simple analysis of metagenomics data, are available. In Li et al. (2018), interactions between the species of *Microcystis* microbiome with the information achieved with the reconstruction of some specific pathways (after the processes of genome reconstruction and annotation), were predicted.

NetCooperate (Levy et al., 2015) is a tool that aims to determine the cooperative potential between microorganisms (and between host-microbe). This tool is based on the *reverse-ecology* framework (Levy and Borenstein, 2012), which focuses on obtaining information about microbial interactions with its ecosystem making use of a large amount of genomic data. The input of NetCooperation tool are the metabolic networks of two species, each one encoded as a directed graph where the nodes represent compounds and the edges represent reactions.

The first step is to determine the seed set, a concept derived from *reverse-ecology* framework that refers to the minimal set of compounds acquired exogenously from the microorganism that enables the production of the all the other compounds of the network, providing indications of what should be the habitat of the microorganism.

Based on the calculated seed sets for each species, two metrics are then used: Biosynthetic Support Score (BSS) (refers to compatibility between host and parasite) (Borenstein and Feldman, 2009) and Metabolic Complementarity Index (MCI) (refers to compatibility between two microbial species) (Levy and Borenstein, 2013). The latter is calculated based on the fraction of seeds of the first species that can be found in the network of the second species, but not in its seed set (see figure 6). Both scores range from 0 (no potential for cooperation) to 1 (full cooperation). The purpose of the scores is the comparison between pairs of species, allowing to predict each one present more potential for cooperation.

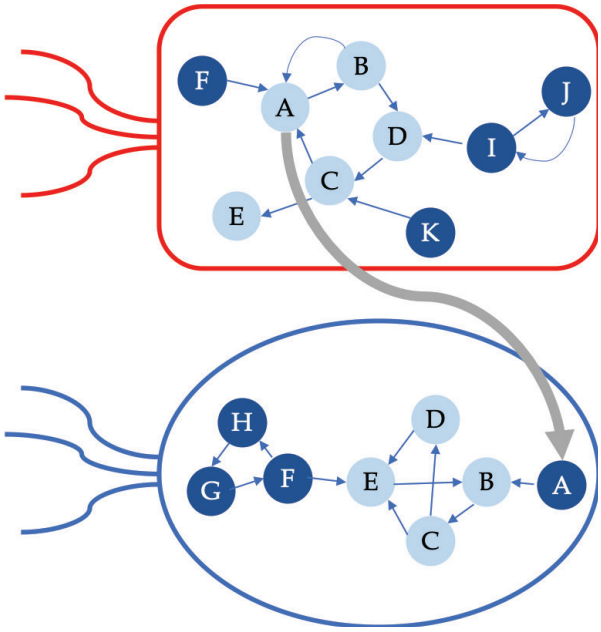


Figure 6.: **Simplification scheme of the measurement of potential cooperation between two different species (MCI).** Here are represented two microbes each one with simple networks represented. It possible to see that the seed A (blue species) is a product in red species network, allowing a cooperation (represented with the grey arrow). On the other hand, compound F is a seed in both species, and so the red species cannot complement F for the blue species. The resultant MCI of the red species on the blue species is 0.5. Adapted from Levy et al. (2015).

3

PROPOSED APPROACH

As discussed in the previous chapter, the goal of this work is to develop a simple and easy to use framework for functional annotation of microbes based on any specific pathway. Further, this framework also allows to infer potential synergistic microbial interactions. The pipeline starts with the assembly of COGs from a set of genomes from a microbial community of interest. Next, annotation of the generated clusters is performed based on a set of reactions that compose a pathway of interest. Finally, potential interspecies interactions are inferred based on the combined genetic potential of microbial clusters to encode all proteins necessary for a complete pathway.

Tools and strategies used for the clustering of genomic sequences and functional annotation are determined based on their efficiency, performance, simple use and easy integration into the pipeline.

All the developed work and the resulting pipeline is described in detail in the following chapters.

4

METHODS

In the present chapter, the datasets and their sources, the clustering and annotation strategy, and pipeline evaluation method, is described.

4.1 DATA

The development and evaluation of a pipeline that goes from clustering of genomes to sequence annotation and metabolic pathways inference, requires high amounts of data and prior information concerning specific pathways and genome annotations. Thus, data regarding a well-known model pathway was collected on which the whole evaluation was focused on. The evaluation of the clustering and annotation methods was based on a set of annotated genomes belonging to model organisms. Furthermore, an additional set of genomes, whose behavior regarding the selected pathway was known, was used for characterizing their functional potential.

4.1.1 *Data resources*

KEGG database was the main resource of the data used in this work. The reason to select this database was its simplicity, the fact that is one of the most used biological databases and the way that genes and gene products (enzymes/reactions involved in various pathways) are linked, facilitating the collection of bulk data in an organized way (Kanehisa, 2019).

A big component of KEGG is KEGG PATHWAY, a collection of graphical diagrams, called pathway maps. In these pathway maps it is easy to visualize the links between enzymes/reactions and compounds in a specific metabolic pathway. Another valuable trait of these maps is that, through the reactions in the metabolic pathway, it is possible to access and retrieve KEGG Orthology (KO) groups. These KO groups are manually defined functional orthologous, and KO identifiers are a cluster of genes from annotated genomes that can be accessed through KEGG. The KO identifiers allow accessing gene identifiers from various organisms, as well as their nucleotide and amino acid sequences (Kanehisa, 2019). These features allow

a simple visualization of the connections between the elements of a metabolic pathway. Furthermore, it is also possible to download bulk data in an organized manner, which facilitates the analysis and processing of the data.

4.1.2 *Metabolic Pathway*

For the evaluation of the pipeline a set of reactions of the well-known transformation of benzoate to acetyl-CoA (BTA) pathway (KEGG pathway identifier - map00362) were selected. The selected set of reactions are shown in Figure 7, where 3 alternative paths were considered:

- **path 1 (P1)** - represented in red, is composed of 12 reactions/enzymes and 32 associated KO groups;
- **path 2 (P2)** - is an alternative to the path 1, were the reactions R05579 and R03028 are replaced by the reaction R02488 (in yellow), composed with 11 reactions/enzymes and 31 KO groups;
- **path 3 (P3)** - represented in green, is composed of 7 reactions/enzymes and 14 KO groups.

In total there are 20 reactions, involving 20 enzymes and 47 related KOs. The gene-protein-reaction (GPR) rules were manually analyzed to enable a better evaluation and discussion of the results.

4.1.3 *Databases*

All amino acid sequences from the genes, belonging to the previously identified KO groups, were downloaded from the KEGG database. The database format used in this pipeline consists of a folder containing FASTA format files, each corresponding to one function (in this case, one KO). All sequences belonging to the selected model organisms used in this study (described in subsection 4.1.4) were not included in the database, to prevent bias of the results. Finally, 48755 sequences associated to 47 KO groups were downloaded. The database was created using the OrtScraper (ort, b) a tool specifically developed to download bulk data from KEGG database, explained in detail in the next chapter. The created database was divided into three databases, each one to represent each of the alternative paths referenced before. These databases were used for the annotation pipeline. The information regarding the number of reactions, enzymes, KO and sequences represented in each one of the created databases (complete BTA, BTA P1, BTA P2, and BTA P3) are shown on table 3.

More information regarding the associations between reaction IDs, Enzyme Commission (EC) numbers and KO identifiers can be found in Supplementary tables 21,22,23 and 24.

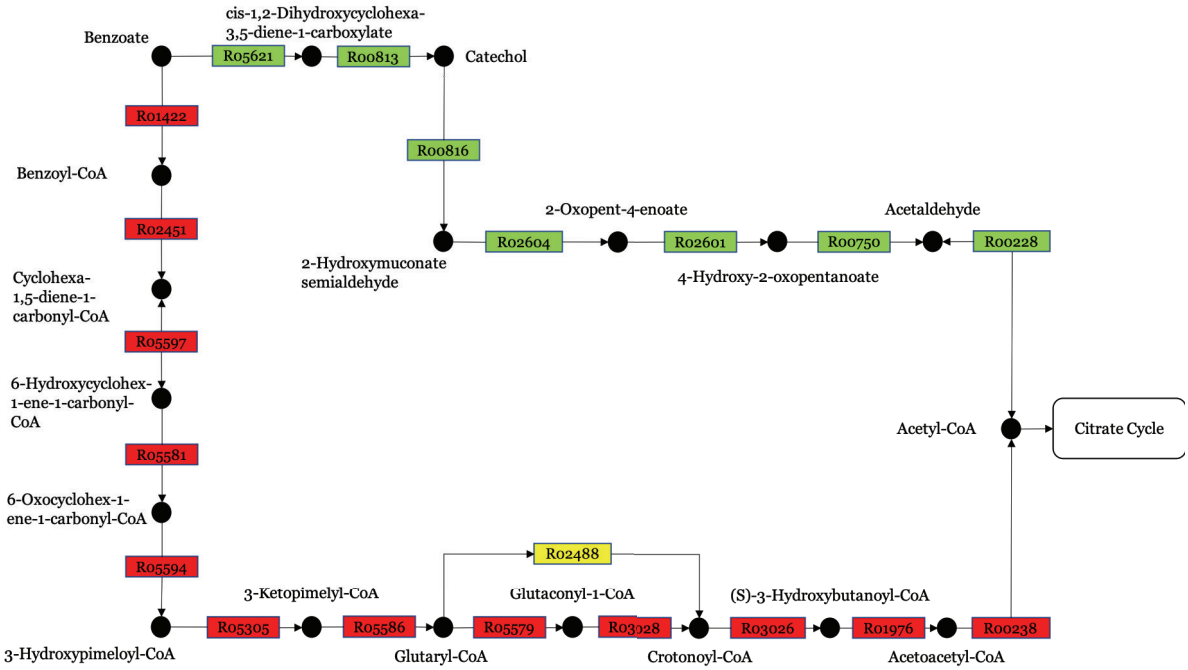


Figure 7.: **BTA pathway map.** Three alternative paths are represented: path 1 (red), path 2 (yellow, differing from path 1 in the transformation from Glutaryl-CoA to Crotonoyl-CoA), path 3 (green).

Table 3.: **Number of reactions, enzymes, KO groups and sequences represented in each alternative path.**

	BTA	BTA P1	BTA P2	BTA P3
Total Reactions	20	12	11	7
Total Enzymes	20	12	11	7
Total KO groups	47	32	31	14
Total number of sequences	48755	37976	40431	8253

4.1.4 *Genomes*

Two different sets of genomes were collected: Set A and Set B.

Set A encompassed a total of 18 species (Table 4) from who's the annotated genomes were downloaded from KEGG. Species were selected based on their genetic ability to encode proteins in at least one reaction involved in Benzoate to Acetyl-CoA conversion. An additional subset of 3 genes associated to KO IDs K05783, K07537 and K07538 from *Burkholderia vietnamiensis* G4, *Azoarcus sp.* CIB and *Aromatoleum Aromaticum* EbN1, respectively, were artificially mutated in their coding sequences at the rates of 0.01, 0.03, 0.05, 0.1, 0.15 and 0.25. Each rate of mutation resulted in a new genome. The mutated genes were used to determine how clustering of orthologous was affected by different sequence similarities. From

Table 4.: **Name, taxonomic code, genome entry code and complete paths from the species represented in Set A of this study.**

Name and strain	Taxonomic code	Genome entry code (KEGG)	Complete paths
Acinetobacter defluvii WCHA30	adv	T05474	P3
Arabidopsis thaliana	ath	T00041	
Azoarcus sp. KH32C	aza	T02502	P2
Azoarcus sp. DN11	azd	T05691	P2
Azoarcus sp. CIB	azi	T04019	P2
Burkholderia cepacia DDS 7H-2	bced	T03302	P3
Burkholderia vietnamiensis G4	bvi	T00493	P3
Cycloclasticus sp. P1	cyq	T02265	P3
Cycloclasticus zancles 78-ME	cza	T02780	P3
Desulfosporosinus orientis DSM 765	dor	T01675	
Aromatoleum aromaticum EbN1	eba	T00222	P2
Latimeria chalumnae (coelacanth)	lcm	T02913	
Magnetospirillum sp. XM-1	magx	T04231	P2
Paraburkholderia aromaticivorans BN5	parb	T05169	P3
Rhodococcus ruber P14	rrz	T05142	P3
Sulfuritalea hydrogenivorans sk43H	shd	T03591	P2
Staphylococcus sciuri FDAARGOS_285	sscu	T05176	
Thauera sp. MZ1T	tmz	T00804	P2, P3

the manually retrieved GPR rules it was possible to calculate which species had the genetic potential to encode all proteins necessary for one of the complete benzoate degradation pathways. A total of 7 and 8 species had the genetic potential to perform the complete pathways 2 and 3, respectively. No single species had the genetic potential to completely implement pathway 1. This information was used as a reference when evaluating the pathway inference step of the pipeline with set A. The information regarding the annotated genes was used as a reference on the evaluation of the annotation step of the pipeline.

Set B was assembled according to the species used by Fetzer et al. (2015)(Table 5). Genome recovery was performed as follows. Bacterial cryo-cultures were revived on LB agar plates. Single colonies were picked and grown overnight in 2 ml LB medium at 37°C. The cells were pelleted by centrifugation, cells were lysed and genomic DNA was extracted using a Nucleospin Tissue Kit (Machery and Nagel). Approx. 150 to 1000 ng of DNA were used for fragmentation (insert size: 300 – 700 bp) and sequencing library preparation following the NEB Ultra II FS Kit protocol (New Enland Biolabs). Libraries were quantified using a JetSeq Library Quantification Lo-ROX Kit (Bioline) and quality-checked by Bioanalyzer (Agilent). Libraries were sequenced on an Illumina MiSeq Instrument with a final concentration of 8 pM using the v3 600 cycles chemistry and 5% PhiX.

Gene prediction and protein sequences were calculated using Prodigal (Hyatt et al., 2010).

For each species, it was known the capability of benzoate degradation (present in the last column of table 5). Only 5 of the 12 species are considered benzoate degraders. This information was used as a reference when evaluating the pathway inference step of the pipeline with set B.

Table 5.: **Name, code and information regarding benzoate degradation capabilities from the species represented in Set B.**

Name and Strain	Code	Benzoate Degrader
Bacillus subtilis ATCC 6633	A	No
Paenibacillus polymyxa ATCC 842	B	No
Brevibacillus brevis ATCC 8246	C	No
Comamonas testosteroni ATCC 11996	D	Yes
Cupriavidus necator JMP 134	E	Yes
Variovorax paradoxus ATCC 17713	H	No
Acidovorax facilis Isolate UFZ	J	No
Pseudomonas putida ATCC 17514	F	Yes
Pseudomonas fluorescens DSM 6290	G	Yes
Rhodococcus sp. Isolate UFZ	I	No
Rhodococcus ruber BU3	K	Yes
Sphingobium yanoikuyae DSM 6900	L	No

4.2 CLUSTERING

Clustering, as mentioned in chapter 3, is the first step of the pipeline. Here, COGs are generated based on the provided genomes of interest. This process allows assessing the degree of similarity sequences must share to be grouped together.

4.2.1 *OrthoFinder*

OrthoFinder was selected for clustering of genomic sequences. OrthoFinder provides information regarding orthologous genes detection, inference of rooted gene trees and species trees, identification of gene duplication events and comprehensive statistics for further comparative genomics analyses. The clusters created with this tool comprise sequences for multiple species, called orthogroups (Nichio et al., 2017).

4.2.2 *Clustering evaluation*

To assess whether sequences clustered together would share the same function, and how the sequence aligner (DIAMOND or BLAST) used by OrthoFinder would affect results, the clustering evaluation was performed by pair-wise precision and recall. The attributes necessary for calculating performance are as follows:

- true positives (TP) are pairs of sequences that share the same function and were indeed clustered together;

- true negatives (TN) are pairs of sequences that do not share the same function and were not clustered together;

- false positives (FP) are pairs of sequences that were wrongfully clustered together, though not sharing the same function;

- false negative (FN) are pairs of sequences sharing the same function but were not clustered together.

The precision, recall, and F_1 were calculated following the equations:

$$precision = \frac{TP}{FP + TP} \tag{2}$$

$$recall = \frac{TP}{FN + TP} \tag{3}$$

$$F_1 = 2 \cdot \left(\frac{precision \cdot recall}{precision + recall} \right) \tag{4}$$

In which, precision is the fraction of the pairs of sequences clustered together that should effectively be clustered together. Recall is the fraction of sequences that should have been clustered together, successfully clustered. F_1 is the harmonic mean between them.

4.3 ANNOTATION

The goal for the annotation strategy is to make use of the clustering information to perform a precise and rapid annotation of the sequences. The simplest way to achieve this goal is to select a few random representatives from each cluster, compare the representative sequences against the known sequences from the database, and annotate all the sequences from a cluster to the function that the representative sequence was annotated for. A potential limitation with this strategy is the propagation of precision errors from the clustering to the annotation phase. That is, if sequences clustered together do not share the same function that means that a cluster should not be annotated with only a single function.

Thus, a new strategy (described in 4.3.1) was employed to make use of the clustering information, reduce the search space for higher efficiency whilst not compromising the precision of annotation.

4.3.1 *Annotation strategy*

The annotation phase of this pipeline takes as inputs the clusters of sequences obtained during the clustering phase and a user-defined database organized into groups of annotated sequences sharing the same function.

Annotation is divided into two main steps: relaxed search and restrictive search, explained in detail below.

■ Relaxed Search

The goal of relaxed search is to decrease the number of alignments required to assign function to COGs. Here, for each COG, one random sequence per each ten sequences is aligned to the groups of sequences in the database. Clusters, where these sequences share a predefined identity percentage value to sequences in the database, are selected for the restrictive search. Notice that one or more functions might be assigned to each cluster.

■ Restrictive Search

In the restrictive search a more stringent set of parameters are employed to assign function to individual sequences. The main steps of restrictive search are as follows:

1. Search between clusters and associated functions - compare all the sequences from the clusters against all the sequences from the database with the function that the cluster was associated with during the relaxed search.

2. Annotation – filtering from the obtained results the best hits for each sequence and assignment of function are based on the following parameters:
   - Percent of identity;
   - Percent of positive matches;
   - Query coverage;
   - Subject coverage.

Since the definition of the best threshold to be used for each one of the parameters could be difficult, a score in which each one of the parameters has the same weight is used. This score is calculated as follows:

$$score = \frac{\%identity + \%pos + \%qcov + t\%scov}{4} \tag{5}$$

The %identity is referent to the percent of identity, %pos to the percent of positive matches, % qcov to the percent of the query coverage and %scov to the percent of the subject coverage.

4.3.2 *Annotation evaluation*

The evaluation of sequence annotation was also performed by calculating precision and recall. However, here TP, TN, FP, and FN have different definitions:

- TP is a sequence that is assigned to the correct function;
- TN is a sequence that is correctly not assigned with any function from the database;
- FP is a sequence assigned to a wrong function (where the correct result should have been another function or none of the database);
- FN is a sequence not assigned to any function when it should have been.

In this case, the precision is the fraction of the annotated sequences that were assigned to the correct function. Recall is the fraction of sequences that should have been annotated to some function, that were successfully annotated. F_1 is the harmonic mean between them.

4.4 species functional potential and microbial interactions

The assessment of the functional potential of individual species, present in the input genome, set to implement a pathway of interest is based on GPR rules of said pathway. Furthermore, the potential synergistic interactions between species to implement a complete pathway is based on their combined genetic content (e.g., if a microbe is missing a single gene to theoretically perform the complete pathway, all other microbes with that gene after considered potential interacting partners).

5

TOOLS AND WORKFLOW

In the present chapter the developed tools and pipeline construction is described. All code is written with Python 3.6 and compatible with UNIX systems. To make the distribution and installation easy for users the Python library *setuptools* (set) is used. For the setup of the virtual environment Python 3 *virtualenv* tool (ven) is recommended. The *argparse* Python module (arg) is employed to allow the creation of a user-friendly command-line interface. The developed tools, OrtScraper and OrtAn, are available on the GitHub platform, along with README files in which the installation and usage of the tools is described.

5.1 ORTSCRAPER

OrtScraper (ort, b) is a tool developed to retrieve data in bulk from the KEGG database, and create, with the download data, a customized and organized database.

5.1.1 *Input*

The tool accepts different types of inputs, which can be:

- A single pathway map ID from KEGG;
- A text file containing reaction IDs from KEGG (one per line);
- A text file containing EC numbers (one per line);
- A text file containing KO group IDs from KEGG (one per line);

5.1.2 *Output*

The output of the tool is:

- Organized database, where each FASTA file corresponds to a KO group and contains all the sequences from the KEGG database assigned to that KO group;

- A text file containing the associations between reactions IDs or EC numbers and the selected KO to download (this file does not exist in case of the input being a pathway map ID or a list of KO groups, since in case the input being a pathway ID all the KO groups download are associated to that pathway ID).

5.1.3 *Implementation*

OrtScraper consists of 3 scripts:

- download_kos.py

The download_kos.py is the main script, which executes the OrtScraper pipeline. First, KEGG is accessed to retrieve the list of KO groups associated with a given list of identifiers. This procedure is skipped in the cases where the input provided by the user already consists of a list of KO groups. Once the list of KO groups is defined, the next task, still with this script, is to retrieve the list of all the genes in the KEGG database assigned to each KO group. Finally, a FASTA file e generated for each KO, where the sequences of all the obtained genes are stored. By default, OrtScraper downloads amino acid sequences for each genome but the user also has the option to request nucleotide sequences.

- MultipleRequests.py

The MultipleRequests.py script is responsible for performing the requests. It uses a Python library called *grequests* (gre) that helps to make asynchronous HTTP requests simple. Asynchronous HTTP requests uses non-blocking I/O, i.e., the program is not blocked while waiting for an answer, allowing for multiplied requests to be made at the same time (using different threads), accelerating the entire process of getting all the requested responses from the KEGG database (htt).

- aux.py

The aux.py script is responsible for calling the methods for parsing the responses retrieved from the requests and makes use of the library BeautifulSoup (bfs) that simplifies the parsing of HTML pages.

5.1.4 *Usage*

After the installation, facilitated by the use of the *virtualenv* tool and the *setuptools* library, OrtScraper is ready to use through a command-line interface. The only command available in the tool is `download_kos`. The available arguments and their respective actions in this command are:

- `h` - print in the console information regarding the tool and all the possible arguments;
- `o` - path to the output directory, i.e., the folder where the database will be stored;
- `m` - pathway map ID from where the user pretends to download the associated KO groups;
- `r` - path to a text file containing the reaction IDs from where the user pretends to download the associated KO groups;
- `e` - path to a text file containing the EC numbers from where the user pretends to download the associated KO groups;
- `k` - path to a text file containing the KO group IDs that the user pretends to download;
- `p` - used to indicate that the user pretends to download amino acid sequences (default option);
- `g` - used to indicate that the user pretends to download nucleotide sequences;
- `s` - defines the number of requests that are made to the KEGG database at the same time (default: 5);
- `v` - used to set loglevel to DEBUG, allowing the user to see debug information while the command is running.

Some examples of how the tool can be used are as follows:

1. `download_kos -o /path/to/output/folder/ -m map map00362`
2. `download_kos -o /path/to/output/folder/ -k kos.txt`

In the first example, the tool will download all the gene sequences (amino acid) assigned to the KO groups associated with the pathway with the ID map00362. In the second example, the tool will create a database with all gene sequences (amino acid) assigned with the KO groups listed in the file kos.txt. The created database always corresponds to a directory with the individual FASTA files for each KO group.

5.2 ORTAN

OrtAn (ort, a) is a tool developed to perform genome annotation via the orthogroups generated by OrthoFinder. OrtAn uses DIAMOND as sequence aligner. The workflow of OrtAn is shown in Figure 8.

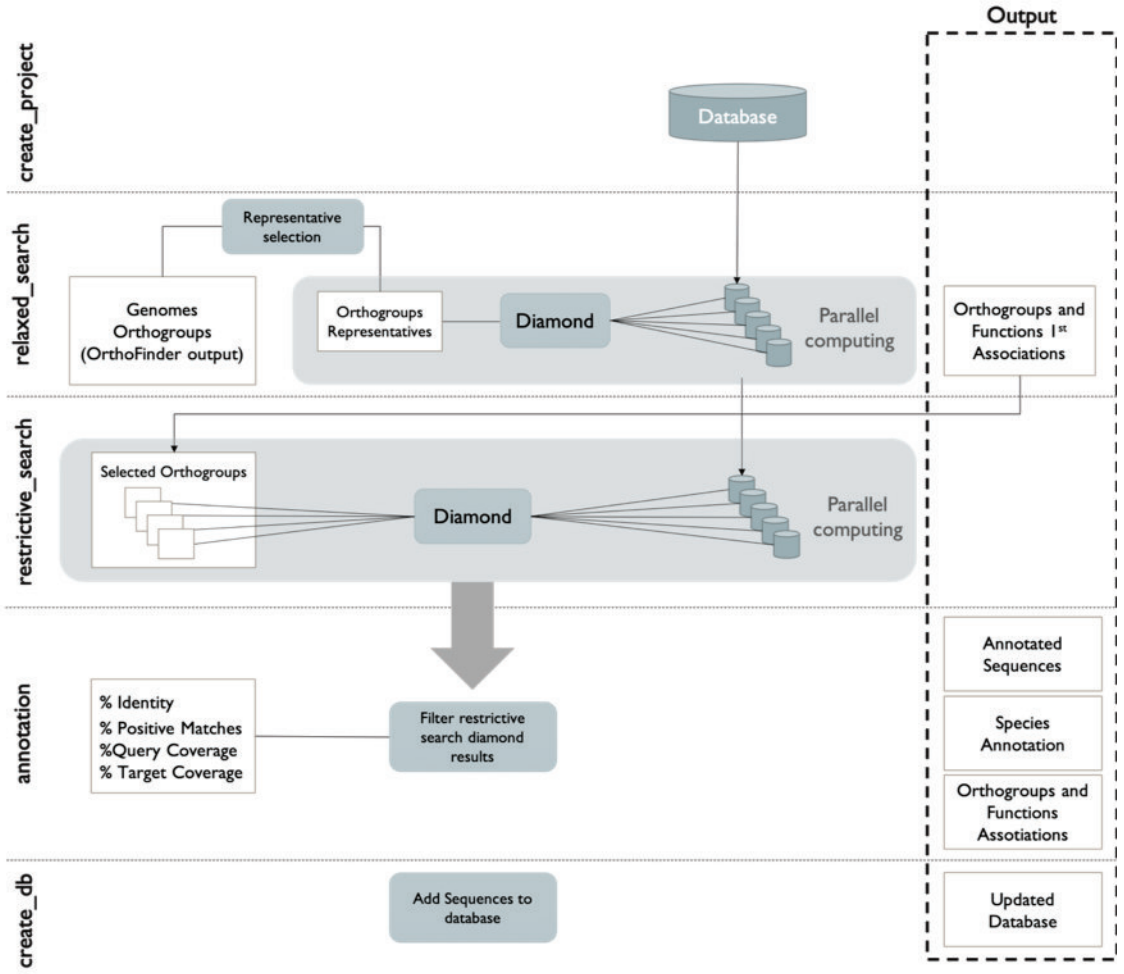


Figure 8.: **OrtAn workflow.** The workflow of OrtAn is divided into five sections separated by dotted lines, each corresponding to one of the main steps/commands of the tool. It is possible to visualize which information is necessary for every command to run, which information is generated by each command and which are the outputs of each command that constitutes the final results for the user (inside the dashed box). The first step is `create_project`, and only consists of the input database (in the OrtScraper output format). The second step is the `relaxed_search`, which calculates associations between the clusters generated by OrthoFinder and the functions in the database. The main actions of this step are representative selection (that consists of selecting representative sequences from each orthogroup) and DIAMOND search that will find the first associations between orthogroups and functions represented in the database - the output of this step. The 'relaxed' name comes from the relaxed threshold used to filter these associations (should be between 40% and 70%). The `restrictive_search` step consists in doing a second DIAMOND search with stricter parameters. Here, each search corresponds to all the sequences of an orthogroup as queries and all the sequences of the functions that the orthogroup was associated with in the `relaxed_search` step. The results obtained in the `restrictive_search` step are then used in the `annotation` step. The `annotation` consists of filtering the DIAMOND search results (using the values of identity, positive matches, query and subject coverage percent combined in a unique score). At the end of the filtering, the kept hits are used to register the annotation of the individual sequences (and not the orthogroups as a whole) and extract information about the functional potential of each species. Further, an overview of the annotated sequences in each orthogroup are calculated and returned as an output to the user. The last and optional step is `create_db`, whose only purpose is to build a new database combining the initial database with the recently annotated sequences. In both search steps, the DIAMOND searches are possible to be parallelized, since the used database is organized in different FASTA files and, in case of the `restrictive_search` step, different query files are used.

5.2.1 *Input*

To perform the annotation the tool needs two inputs:

- OrthoFinder results obtained with the genomes that the user pretends to annotate;
- Organized database (OrtScraper returned format) containing the functions that the user pretends to find in the genomes.

5.2.2 *Output*

From the relaxed search step, described in the previous chapter, the output is a text file (/Results/Associations.txt) containing the associations between the orthogroups and the functions represented in the database.

From the restrictive search/annotation step are outputted 6 different text files:

- **/Results/Annotation_Function_Protein.txt** - Shows in the first column the functions and in the second the sequences annotated to that function (one association per line);
- **/Results/Annotation_Protein_Function.txt** - Shows in the first column the sequences and the second the functions for which the sequences were annotated to (one association per line);
- **/Results/ConOG.txt** - Consistent Orthogroups (ConOGs), orthogroups where all the sequences were annotated to with same function. The function that was assigned to the orthogroup is also indicated in the second column;
- **/Results/DivOG.txt** - Divergent Orthogroups (DivOGs), orthogroups where not all the sequences were annotated to the same function. This means that the orthogroup could have sequences that were not annotated to any function or sequences annotated to different functions. The functions that were associated with the DivOGs are also indicated in the file, second column;
- **/Results/Orthogroups_Annotation.csv** - This file shows a table indicating how many sequences in each orthogroup were annotated and to which function;
- **/Results/Species_Annotation.csv** - This file shows a table indicating which functions are present in which species (1 - at least one sequence annotated to the function, 0 - no sequences annotated to the function);
- **/Results/Overview.csv** - This file shows a table containing an overview of the values obtained through all the process, comparing the values from where the tool started to

the relaxed search and restrictive search. The values presented in this table and its meaning are the following:

- **Total orthogroups**: total number of orthogroups obtained with OrthoFinder (input data);
- **KOs in the database**: Number of different KO groups/ functions represented in the database;
- **Selected orthogroups**: number of orthogroups that had at least one association with some of the functions represented in the database in the relaxed search step (one of the randomly selected representatives presented at least one alignment against a sequence from the database with an identity value higher than the used cut-off);
- **% Selected orthogroups**: percent of selected orthogroups (the number of selected orthogroups divided by the number of total orthogroups);
- **Associated KOs**: number of KO groups/functions represented in the database that were associated with at least one orthogroup;
- **% Associated KOs**: percent of associated KO groups (the number of associated KO groups divided by the number of total KO groups in the database, multiplied by 100);
- **Orthogroups with annotated sequences**: number of orthogroups that had at least one of its sequences annotated to some KO group/function;
- **% of Orthogroups with annotated sequences**: percent of orthogroups with at least one annotated sequence (the number of orthogroups with annotation divided by the number of total orthogroups, multiplied by 100);
- **KOs with assigned sequences**: number of KO groups/functions that were assigned to at least one sequence between the orthogroups;
- **% KOs with assigned sequences**: percent of KO groups with assigned sequences (the number of assigned KO groups divided by the number of total KOs in the database, multiplied by 100);
- **ConOG**: number of consistent orthogroups (orthogroups were all the sequences were annotated to the same function);
- **DivOG**: number of divergent orthogroups (orthogroups were not all the sequences were annotated to the same function, could be not annotated to any or annotated to another one);
- **DivOG with more than one KO**: number of divergent orthogroups sequences annotated to different KO groups/functions;

- **Lost orthogroups**: number of orthogroups that were associated with some KO group/function in the relaxed search step but in the restrictive search step, none of its sequences was annotated to any KO group/function;
- **% Lost orthogroups**: percent of lost orthogroups (the number of selected orthogroups (relaxed search) minus the number of orthogroups annotated sequences (restrictive search), divided by the number of selected orthogroups (relaxed search), multiplied by 100);
- **Lost KOs**: number of KO groups/functions that were associated with some orthogroup in the relaxed search step but were not assigned to any sequence in the restrictive search/annotation step;
- **% Lost KOs**: percent of lost KO groups/functions (the number of associated KO groups (relaxed search) minus the number of assigned KO groups (restrictive search), divided by the number of associated KO groups (relaxed search), multiplied by 100).

5.2.3 *Implementation*

The OrtAn pipeline is divided into 5 steps, assured by 5 main scripts, which are also responsible for the 5 possible commands involved in each step. The commands, that must be performed sequentially, are described as follows:

- `create_project`

This command takes as input the path where the user wants to store all the information generated by the tool (working directory) and the path to the database. Some of the tasks performed by this command are common to all the others (validation of the input format; creation of a log file (if desired by the user) to store the messages regarding the command runs; and the storage of information regarding the actions successfully performed). The main task of this command is the creation of the structure for the working directory, where some temporary data (for instance query and database files to use with DIAMOND as well as its results files) will be stored, as well as some persistent data to use between the different commands. The directory where all the results will be stored is also created in this step. A dictionary variable that is stored in a JSON file with the purpose to share data between the different commands regarding the actions that were successfully performed, and other relevant aspects is also generated during this step.

- `relaxed_search`

This step takes as input the path to the working directory of the project (generated in the `create_project` step) and the path to the OrthoFinder results directory. Here the tool

performs the relaxed search referred in the previous chapter (with DIAMOND) between the orthogroups (resulted from OrthoFinder) and the given database. Only 1 in 10 sequences from each orthogroup are randomly selected and used as queries to find an association between the orthogroups and the functions from the database. All the representative sequences of the orthogroups are used as queries in the DIAMOND search, and the used database is composed by the sequences from the input database. In this step, only the identity percentage is used to filter the DIAMOND results and to retrieve the associations between orthogroups and functions. If at least one of the sequences used to represent an orthogroup has a hit with an identity percent equal or higher than the selected threshold, the orthogroup will be associated with the respective function of the hit subject sequence. The user can indicate the identity threshold to be used. Theoretically, the lower the identity threshold, the higher the number of associations that will be found between the orthogroups and the database functions. A higher number of associations means a higher number of sequences to be included in the DIAMOND searches of the next step, which could affect negatively the time performance of the tool. However, a lower identity threshold could lower the number of FN in the final annotation results (i.e., minimizing the chances of missing annotations). The default value is 50%. The output of this step is the established associations between the orthogroups, and the functions represented in the database.

- `restrictive_search`

The sole input of this step is the working directory of the project (that contains all the necessary information generated until here). Here, the tool uses the calculated associations in the relaxed search. Considering each one of the associations, DIAMOND is used to search for the best hits between all the sequences from an orthogroup (queries) against all the sequences from the database (subjects) annotated to the function of which the orthogroup was associated.

- `annotation`

The input here is, as before, the working directory. The default value for the score to be used to filtering the DIAMOND hits (used to perform the annotation) is 90%. All the parameters (% identity, % positive matches, % subject coverage, % query coverage) contribute equally to the score (see equation 5). However, the option to limit the annotation with a specific parameter is also offered. For instance, the user can exclude hits with values lower than 95% for the query coverage, and, this way, even in cases where the score threshold is fulfilled, if the query coverage value is under 95%, the hit is not kept for the annotation.

From the DIAMOND results calculated in the restrictive search, the hits are filtered according to the following:

- In case there were thresholds defined to each parameter, all the hits that do not meet the threshold are eliminated;
- All hits below the defined score threshold are eliminated;
- If the same query sequence has hits corresponding to subject sequences annotated to different functions, only the hit with the highest score is kept.

In the end, the kept hits are used to perform the annotation. All the query sequences among these results are annotated with the function of the subject sequence that they were aligned with.

The goal of this step is to provide a very accurate annotation of orthogroup sequences. Thus, the used thresholds to make the selection from the DIAMOND search results and minimize the presence of FP, performed in the restrictive search, should be more stringent. The output of this step is 1) a list of sequences from the genomes that were annotated to any of the functions represented in the database, 2) information regarding the functions that were found in each species/genome, 3) information regarding the annotation of the orthogroups (ConOGs, DivOG, and the exact number of annotated sequences) and 4) an overview of all the values obtained through the OrtAn steps.

- `create_db`

With the `create_db` command (optional) the user has the opportunity to update the given database with the recently annotated sequences from the input genomes. The user also has the option to maintain the original database and create a new one that includes all the sequences (which requires the user to provide the working directory and the directory where the new database is to be stored).

The auxiliary scripts to help in the main commands are:

- aux.py

This script is composed of various auxiliary functions that go from preparing query files or databases to parsing/filtering results files from the DIAMOND searches.

- diamond_mp.py

This script is responsible for calling the DIAMOND commands and manage the parallel processes. Here the commands used to run DIAMOND are *makedb* (with the default options) which is used to create the databases in the required format which will be used to perform the searches, and *blastp* that performs the searches between the amino acid sequences. The options used in the *blastp* command are set to return all the hits found (unless the user has

defined a specific threshold to be respected for one of the parameters referred before, %identity, %positive matches, %query coverage, and %subject coverage). The default filtration by the e-value that DIAMOND employs is disabled. This is necessary because the e-value is dependent on the database size and sequences, and, since different databases are used, the e-value cannot be compared between the different searches (Kerfeld and Scott, 2011). Since in each step, there are normally various independent searches to be made, the code is prepared to run the DIAMOND searches in parallel (depending on the number of cores available in the machine). The number of parallel processes can also be defined by the user.

5.2.4 *Usage*

Similarly, to OrtScraper, the OrtAn installation is facilitated by the *virtualenv* tool and the *setuptools* library. The communication of the user with the tool is also made through a user-friendly command-line interface. As already referred (in the subsection 5.2.3), the 5 commands of the tool must be performed sequentially. The parameters in each command that can be defined by the user are described in detail below.

- `create_project`
  - `h` - displays information regarding the tool usage and all the possible arguments;
  - `out` - path to the folder that will be the working directory;
  - `db` - path to the database directory;
  - `l` - used to send log messages to a file in the output directory;
  - `v` - used to set loglevel to DEBUG, allowing the user to see debug information while the command is running.
- `relaxed_search`
  - `h` - displays information regarding the tool usage and all the possible arguments;
  - `wd` - working directory;
  - `of` - path to OrthoFinder results directory. This directory must contain the Orthgroups, Orthogroup_Sequences and WorkingDirectory folders;
  - `ident` - identity percentage threshold used to filter the DIAMOND results;
  - `t` - the number of processes to run in parallel. By default, it uses all CPUs available on the machine. Setting to '1' will make the processes run sequentially;
  - `del` - used to delete the results stored from DIAMOND. This option should be used to save memory space, between steps;
  - `l` - used to send log messages to a file int the output directory;

- `v` - used to set loglevel to DEBUG, allowing the user to see debug information while the command is running.

- `restrictive_search`
  - `h` - displays information regarding the tool usage and all the possible arguments;
  - `wd` - working directory;
  - `t` - the number of processes to run in parallel. By default, it uses all CPUs available on the machine. Setting to '1' will make the processes run sequentially;
  - `ident` - identity percent threshold used to filter the DIAMOND results.
  - `l` - used to send log messages to a file in the output directory;
  - `v` - used to set loglevel to DEBUG, allowing the user to see debug information while the command is running.

- `annotation`
  - `h` - displays information regarding the tool usage and all the possible arguments;
  - `wd` - working directory;
  - `s` - score threshold used to filter the DIAMOND results;
  - `ident` - identity percent threshold used to filter the DIAMOND results;
  - `qc` - query sequence coverage percent threshold used to filter the DIAMOND results;
  - `sc` - subject sequence coverage percent threshold used to filter the DIAMOND results;
  - `ppos` - positive matches percent threshold used to filter the DIAMOND results (should be equal or higher to the identity percent threshold);
  - `l` - used to send log messages to a file in the output directory;
  - `v` - used to set loglevel to DEBUG, allowing the user to see debug information while the command is running.

- `create_db`
  - `h` - displays information regarding the tool usage and all the possible arguments;
  - `wd` - working directory;
  - `o` - path to the output directory to create a new database (initial database + new annotated sequences);
  - `up` - used to update the initial database with the new annotated sequences. By using this option, the initial database will be changed permanently;

- `l` - used to send log messages to a file in the output directory;
- `v` - used to set loglevel to DEBUG, allowing the user to see debug information while the command is running.

An example of how OrtAn is run after the installation is shown below. First, `$work_dir` is used to refer to the path of the desired working directory. Second `$database` is set as the path of the input database. Third, `$orthof` is used to set the path where the OrthoFinder results are stored and last, `$new_db` is used to set the desired path to store the new database.

1. `create_project -out $work_dir -db $database`
2. `relaxed_search -wd $work_dir -of $orthof -t 1 -ident 60`
3. `restrictive_search -wd $work_dir -t 2`
4. `annotation -wd $work_dir -s 95`
5. `create_db -wd $work_dir -o $new_db`

In this example, a new OrtAn project will be created with a previously prepared database (1). The relaxed search step will run sequentially (-t 1) and the results will be filtered by an identity percent threshold of 60 (2). The restrictive search step will then run with two DIAMOND search processes running in parallel (3). The annotation will be performed using the score threshold of 95 (4). Lastly, a new database containing the recently annotated sequences will be created, maintaining the initial one intact (5).

6

RESULTS AND DISCUSSION

In this chapter, all the results obtained are shown as well as the evaluation of the different steps of the pipeline with the different sets of data (described in chapter 4, section 4.1). Also, the performance of the developed tools will be discussed.

First, the results of the clustering evaluation, performed with the genome set A containing the mutated genomes as well, will be presented. The pipeline, performed with the genomes set A and set B, starts with the clustering step. After that, the annotation. The annotation evaluation (with the calculation of the precision, recall and F_1 values) is only performed with set A since the annotation for the sequences composing the genomes of set B is not known. Then, it is performed the metabolic network inference with the annotation results (for both genome sets) for assessment of species functional potential and possible microbial interactions. To end this chapter, the performance of both tools, OrtScraper and OrtAn, will be discussed.

6.1 CLUSTERING EVALUATION

The clustering evaluation was made using the genome set A with the extra mutated genomes. The purpose was to focus only on the sequences that were known to belong to the BTA pathway and analyze if the sequences sharing the same function were effectively clustered together. For the clustering evaluation, as was already referred in chapter 4, was used a pair-wise precision and recall method. The pool of all the possible pairs included only the sequences involved in the BTA pathway. Each pair of functions clustered together would count as a TP, if they shared the same function, or as a FP, if otherwise. The pairs that were not clustered together and shared the same function would count as a FN. Based on these values, the precision, recall, and F_1 were calculated. This evaluation method was employed for both OrthoFinder results (using BLAST or DIAMOND as an alignment search tool), and the results are shown in Table 6.

Although the DIAMOND aligner is recommended by OrthoFinder developers (D.M. and S., 2018) due to a better trade-off between execution time and sensitivity, both options were independently used to evaluate the influence of the sequence aligner in the final results. The

Table 6.: **Pair-wise precision and recall method evaluation results of OrthoFinder clusters using BLAST and DIAMOND as an alignment search tool.** The OrthoFinder input was the genome set A + mutated genomes.

	OrthoFinder - BLAST	**OrthoFinder - DIAMOND**
TP	47270	51848
FP	27479	28668
FN	13602	8917
Precision	0.63	0.64
Recall	0.77	0.85
F_1	0.69	0.73

difference observed between both aligners were not noteworthy, thus DIAMOND was maintained as the default method. Another observation when analyzing the obtained orthogroups when using DIAMOND was the fact that all the mutated sequences, even those with a mutation rate of 25% were clustered together with the original sequence.

6.2 PIPELINE

This section shows the results obtained running the different steps of the pipeline: clustering, annotation and metabolic pathway inference. Here, only genome set A (without mutated genomes) and set B were used. The use of the genome set A was crucial on the evaluation of the OrtAn sequence annotation since the sequences could be mapped to the reactions (and KO groups) involved in the BTA pathway. Genome set B was used to evaluate how the pipeline could be used for the inference of functional potential of microbial species and provide information of potential synergistic microbial interactions. All the created databases, BTA, BTA P1, BTA P2, and BTA P3 (described in chapter 4, section 4.1) were used to test the pipeline and to understand the effect that different sets of data can have in the adjustment of the parameters.

6.2.1 *Clustering: Set A*

Genomes set A consisted of a total of 157788 genes of which 90787 (57,5%) were assigned to orthogroups. A total of 10692 orthogroups were generated of which 558 were species-specific (i.e., orthogroups were all sequences originated from a single species). A total of 4126 (2.6%) of the genes were assigned to species-specific orthogroups. Further, a total of 228 orthogroups were obtained that were composed of sequences from all the species represented in set A and 7 single-copy orthogroups which were composed by exactly one single gene from each species. Orthogroups were, on average and median, composed of 8.5 and 4 sequences, respectively.

6.2.2 *Clustering: Set B*

Genome set B consisted of 69210 genes of which54856 (79.3%) were assigned to orthogroups. The assigned genes were distributed across 8362 orthogroups of which 45 were species-specific, containing 120 genes (0.2%). A total of 546 and 196 orthogroups were obtained that contained genes belonging to all species present in the genome set and that contained a single gene from each species, respectively. The mean and the median of genes per orthogroup was 6.6 and 3.0, respectively.

6.2.3 *Annotation: Set A*

Genome set A was used to test for different thresholds for the relaxed and restrictive search parameters. The databases tested were the BTA, BTA P1, BTA P2, and BTA P3. For the relaxed search, identity cut-off values of 40% and 70% were used to test the effects of identity stringency in the first associations. For the restrictive search step, score cut-off values of 90% and 95% were employed. The results obtained for the different cases are presented in table 7 (score cut-off of 90%) and table 8 (score cut-off of 95%).

As it can be seen in both tables, 7 and 8, in the section referent to the results obtained in the relaxed search step, the percentage of selected orthogroups was very low. This was already expected since the used databases represent only a very small fraction of the functions composing a genome, in this study the BTA pathway. This outcome becomes increasingly striking when using only one of the alternative paths for benzoate to acetyl-CoA conversion due to the reduction in the number of reactions.

The number of selected orthogroups in the relaxed search step is always smaller when using an identity cut-off of 70%. This is expected since, using a more stringent threshold, more of the aligned hits do not meet the threshold and, therefore, are not used to calculate the associations between orthogroups and functions, i.e., the selected orthogroups. The same behavior can also be seen regarding the number of selected KO groups. For the same reason, more hits that don't meet the threshold, using an identity cut-off of 70% leads to a lower number of KO groups that were associated with orthogroups. This is an important aspect to be aware of because, if a KO group is not associated with any orthogroup in the relaxed search, its sequences will not be compared with the genome sequences in the further steps. Therefore, this function/KO group won't be annotated to any sequence.

Comparing the values obtained in the relaxed search section of the tables 7 and 8 is possible to observe that the results slightly differ even when using the same data and identity cut-off. The reason for that is the fact that they correspond to different OrtAn runs and, since the orthogroups representative selection is made randomly, different representatives can lead to different results regarding the associations between orthogroups and KO groups.

Table 7.: **Overview of OrtAn results with Genome set A for the databases BTA, BTA P1, BTA P2, and BTA P3, relaxed search identity cut-off of 40% and 70% and restrictive search score cut-off of 90%.** The meaning of the values in each line is described in chapter 5, subsection 5.2.2. (OrtAn Overview.csv output file).

Genome Set	**Set A**							
Database	**BTA**		**BTA P1**		**BTA P2**		**BTA P3**	
Relaxed Search identity cut-off	40	70	40	70	40	70	40	70
Annotation score cut-off	90							
Total orthogroups	10692							
KOs in the database	47		32		31		14	
Relaxed Search								
Selected orthogroups	62	27	50	20	47	19	11	9
% Selected orthogroups	0.6	0.3	0.5	0.2	0.4	0.2	0.1	0.1
Associated KOs	47	21	31	13	29	14	12	12
% Associated KOs	100.0	44.7	96.9	40.6	93.5	45.2	85.7	85.7
Restrictive Search/Annotation								
Orthogroups with annotated sequences	34	25	26	18	25	17	8	9
% of Orthogroups with annotated sequences	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.1
KOs with assigned sequences	40	21	26	12	24	14	11	12
% KOs with annotated sequences	85.1	44.7	81.2	37.5	77.4	45.2	78.6	85.7
ConOG	5	5	4	4	4	4	1	1
DivOG	29	20	22	14	21	13	7	8
DivOG with more than one KO	12	5	6	2	5	3	3	3
Relaxed Search to Restrictive Search								
Lost orthogroups	28	2	24	2	22	2	3	0
% Lost orthogroups	45.2	7.4	48.0	10.0	46.8	10.5	27.3	0.0
Lost KOs	7	0	5	1	5	0	1	0
% Lost KOs	14.9	0.0	16.1	7.7	17.2	0.0	8.3	0.0

Table 8.: **Overview of OrtAn results with Genome set A for the databases BTA, BTA P1, BTA P2, and BTA P3, relaxed search identity cut-off of 40% and 70% and restrictive search score cut-off of 95%.** The meaning of the values in each line is described in chapter 5, subsection 5.2.2. (OrtAn Overview.csv output file).

Genome Set	**Set A**							
Database	**BTA**		**BTA P1**		**BTA P2**		**BTA P3**	
Relaxed Search identity cut-off	40	70	40	70	40	70	40	70
Annotation score cut-off	95							
Total orthogroups	10692							
KOs in the database	47		32		31		14	
Relaxed Search								
Selected orthogroups	61	26	51	18	46	18	14	6
% Selected orthogroups	0.6	0.2	0.5	0.2	0.4	0.2	0.1	0.1
Associated KOs	45	24	32	13	31	13	13	8
% Associated KOs	95.7	51.1	100.0	40.6	100.0	41.9	92.9	57.1
Restrictive Search/Annotation								
Orthogroups with annotated sequences	29	20	22	11	22	11	8	6
% of Orthogroups with annotated sequences	0.3	0.2	0.2	0.1	0.2	0.1	0.1	0.1
KOs with assigned sequences	34	21	22	11	22	11	12	8
% KOs with annotated sequences	72.3	44.7	68.8	34.4	71.0	35.5	85.7	57.1
ConOG	0	0	0	0	0	0	0	0
DivOG	29	20	22	11	22	11	8	6
DivOG with more than one KO	10	4	5	3	5	2	4	2
Relaxed Search to Restrictive Search								
Lost orthogroups	32	6	29	7	24	7	6	0
% Lost orthogroups	52.5	23.1	56.9	38.9	52.2	38.9	42.9	0.0
Lost KOs	11	3	10	2	9	2	1	0
% Lost KOs	24.4	12.5	31.2	15.4	29.0	15.4	7.7	0.0

Analysis of results obtained during the restrictive search and annotation steps reveal that a complete assignment of KO groups to a sequence is never reached. Among others, factors that can contribute to the results are:

- Restrictive search cut-off score too restrictive:

  By using a score cut-off of 95%, the number of KO groups will only be associated with any highly similar sequences, leading to improved precision. However, more restrictive thresholds lead to a higher number of FN which decreases recall.

- KO groups poorly represented in the database:

  For instance, K04105 is only represented by 19 sequences in the database, K07547 is only represented 5 and K07548 only 6. This makes it harder to identify sequences having the same function in different genomes due to the decreased number of reference sequences of the function in the analysis.

- KO groups that, besides separated, correspond to the same function.

  This issue arises due to the annotation of KO groups in the sources used to retrieve data. An example case of that is a set of 10 KO groups (K07515, K01825, K07514, K07511, K01782, K15016, K01692, K13767, K01715 and K10527) that are assigned to the enzyme 4.2.1.17 (present in path 1 and 2) and where the product is the same for all, enoyl-CoA hydratase. The number of sequences in the database for each one of these KO groups goes from 130 to 6554. The KO group represented with more sequences will probably lead to more alignments with sequences in the genome set that have the same function, making it easier to find a hit with best values. Because OrtAn only allows the annotation of each sequence to the function of the best hit, it is more probable that all the sequences that the product corresponds to enoyl-CoA hydratase will be more easily associated with the KO group that has more reference sequences representing it in the database.

Another interesting outcome of the restrictive search step, is the higher number of DivOG than ConOG. This is an indicator that an annotation strategy of annotating all the sequences from an orthogroup to a function would frequently lead to misleading results, especially considering that some of the DivOG have sequences annotated to different KO groups. It is also interesting that when using a restrictive search score cut-off of 95% no ConOG was obtained. A possible explanation for this scenario is that due to a very restrictive score the less likely that all the sequences from a KO will be annotated to a sequence in the database.

The number of orthogroups at the end of the restrictive search phase using different relaxed search parameters shows a direct correlation to identity cutoffs. When using a lower identity cut-off on the relaxed search step, a higher number of orthogroups analyzed in the restrictive search end up not having any sequence annotated to any of the functions to which

the orthogroup was associated before. The number of lost KO groups behaves the same way. This shows that using a very low cut-off of the relaxed search can lead to extra work in the further steps, i.e., more comparisons to be made that were unnecessary. Conversely, if a very high identity cut-off is used, that can translate in a loss of opportunity to identify sequences related to the functions in the database. The goal of presenting these values in the OrtAn results is to show if the used thresholds are adjusted to the data.

Regarding the cases of DivOGs with sequences annotated to different KO groups, most of the cases correspond to KOs that correspond to the same function (like the case of 10 KO groups explained before). Some cases showed to correspond indeed to KO groups that don't correspond to the same function, reinforcing the importance of not assign the same function to all the sequences in an orthogroup to avoid some misleading results.

The evaluation of the annotation using OrtAn is shown in tables 9 (annotation score cut-off of 90%) and 10 (annotation score cut-off of 95%). The reference results used for this evaluation were the annotation present in KEGG for the genomes composing set A.

Table 9.: **Annotation evaluation (Set A).** Presenting TP, FN, FP, precision, recall and F_1 values for genome set A annotation evaluation for different relaxed search identity cut-off (40% and 70%), restrictive search score cut-off of 90% and for 4 different databases (BTA, BTA P1, BTA P2 and BTA P3).

Relaxed Search identity cut-off	**40**	**70**	**40**	**70**	**40**	**70**	**40**	**70**
Restrictive search/Annotation score cut-off	**90**							
Database	**BTA**		**BTA P1**		**BTA P2**		**BTA P3**	
TP	436	376	296	254	307	273	126	129
FN	168	228	151	193	149	183	16	13
FP	36	31	36	29	32	32	0	0
Precision	0.923	0.923	0.891	0.897	0.905	0.895	1.0	1.0
Recall	0.721	0.622	0.662	0.568	0.673	0.598	0.887	0.908
F_1	0.810	0.743	0.759	0.695	0.772	0.717	0.940	0.952

Table 10.: **Annotation evaluation (Set A).**Presenting TP, FN, FP, precision, recall and F_1 values for genome set A annotation evaluation for different relaxed search identity cut-off (40% and 70%), restrictive search score cut-off of 95% and for 4 different databases (BTA, BTA P1, BTA P2 and BTA P3).

Relaxed Search identity cut-off	**40**	**70**	**40**	**70**	**40**	**70**	**40**	**70**
Restrictive search/Annotation score cut-off	**95**							
Database	**BTA**		**BTA P1**		**BTA P2**		**BTA P3**	
TP	299	259	187	147	200	169	106	84
FN	305	345	260	300	256	287	36	58
FP	16	12	16	16	15	14	0	0
Precision	0.949	0.956	0.921	0.902	0.930	0.923	1.0	1.0
Recall	0.495	0.429	0.418	0.329	0.439	0.371	0.746	0.596
F_1	0.651	0.592	0.575	0.482	0.596	0.529	0.855	0.743

The differences in the results obtained with the different databases show the impact that the database can have on the annotation. Here, the best annotation results were obtained for BTA P3 database. The lesser performance of databases BTA P1 and P2 can be attributed to KO

groups with low numbers of representative sequences, that, as referred before in this section, can difficult the identification of sequences having the same function. This contributed to high numbers of FN and consequently to low recall.

When looking at cases were only the relaxed search identity cut-off was changed, it is possible to observe that the precision values do not fluctuate meaningfully but the recall values do. Again, this highlights the impact that the identity cut-off on the relaxed search can have on the recall values of annotation. F_1 values showed better when using the identity cut-off on the relaxed search of 40%. The only exception to this is the database BTC P3 when using 90% as score cut-off on the restrictive search. In this case, the results obtained when using an identity cut-off of 70% were slightly better. An explanation for this can be the randomly orthogroups representatives' selection. In the case of the run where was used the identity cut-off of 40%, the selected representative for an orthogroup could have led it to not be associated with the function that it should.

Considering only the score cut-off (90% on table 9 or 95% on table 10) as a factor shows that both precision and recall values are affected. The precision values are better when using a high score cut-off, however, the recall also suffers a higher decrease. This leads to F_1 values worse than the ones obtained when using 90% as score cut-off on the restrictive search step.

In general, for the tested datasets, was observed that the best combination of thresholds used are 40% for the identity cut-off in the relaxed search, and 90% for the score cut-off in the restrictive search.

6.2.4 *Annotation: Set B*

In the table 11 (score cut-off of 90%) and 12 (score cut-off of 95%), are presented the results obtained with OrtAn for set B. The test of the pipeline with this set was similar to the set A. It was tested for the 4 different databases (BTA, BTA P1, BTA P2, and BTA P3), for an identity cut-off on the relaxed search step of 40% and 70%, and a score cut-offs of 90% and 95% for the restrictive search.

Most of the observations made with set A can be seen in set B results as well. In tables 11 and 12, on the section regarding the relaxed search results, it is possible to observer that there was a low percentage of the selected orthogroups, because the databases used represent only a small fraction of the functions presented in a genome. This percent was even smaller when using an identity cut-off of 70%, as expected. Here, contrary to the set A results, there wasn't any case where 100% of the KO groups would have at least one association with an orthogroup. Since the genomes here are different, belonging to different species and the annotation of the amino acid sequences is not known is difficult to justify why this happened. It could be due to the quality of the databases used, or simply by the fact that there are no sequences with

Table 11.: **Overview of OrtAn results with Set B for the databases BTA, BTA P1, BTA P2, and BTA P3, relaxed search identity cut-off of 40% and 70% and restrictive search score cut-off of 90%.** The meaning of the values in each line is described in chapter 5, subsection 5.2.2. (OrtAn Overview.csv output file).

Genome Set	**Set B**							
Database	**BTA**		**BTA P1**		**BTA P2**		**BTA P3**	
Relaxed Search identity cut-off	40	70	40	70	40	70	40	70
Annotation score cut-off	90							
Total orthogroups	8362							
KOs in the database	47		32		31		14	
Relaxed Search								
Selected orthogroups	71	38	47	32	54	31	16	8
% Selected orthogroups	0.8	0.5	0.6	0.4	0.6	0.4	0.2	0.1
Associated KOs	38	19	24	10	26	13	26	8
% Associated KOs	80.9	40.4	75.0	31.2	83.9	41.9	83.9	57.1
Restrictive Search/Annotation								
Orthogroups with annotated sequences	38	34	29	29	29	28	11	8
% of Orthogroups with annotated sequences	0.5	0.4	0.3	0.3	0.3	0.3	0.1	0.1
KOs with assigned sequences	23	18	11	9	11	12	13	8
% KOs with annotated sequences	48.9	38.3	34.4	28.1	35.5	38.7	92.9	57.1
ConOG	19	19	15	15	16	16	3	3
DivOG	19	15	14	14	13	12	8	5
DivOG with more than one KO	8	3	5	3	4	3	3	1
Relaxed Search to Restrictive Search								
Lost orthogroups	33	4	18	3	25	3	5	0
% Lost orthogroups	46.5	10.5	38.3	9.4	46.3	9.7	31.2	0.0
Lost KOs	15	1	13	1	15	1	1	0
% Lost KOs	39.5	5.3	54.2	10.0	57.7	7.7	7.1	0.0

these functions present in any of the genomes analyzed. Also, the gene prediction carried out with Prodigal could have not identified some of these genes.

Regarding the restrictive search step, the number of ConOG and DivOg are more balanced.

Another aspect that differs from the results obtained with set A is that the number of ConOG obtained with set B does not change significantly when using different score cut-offs of 90% and 95%. From this, it can be assumed that the ConOG that were maintained when using a score cut-off of 95% correspond to orthogroups with more similar or conserved sequences.

The losses of orthogroups and KO groups from the relaxed search to the restrictive search step presented the same pattern regarding the comparison between both identity cut-offs used, i.e., small losses when using higher identity cut-offs.

6.3 METABOLIC NETWORK INFERENCE

The metabolic network inference strategy combines the information regarding the GPR rules (in this case, manually retrieved from the KEGG database for the BTA pathway) and the

Table 12.: **Overview of OrtAn results with Set C for the databases BTA, BTA P1, BTA P2, and BTA P3, relaxed search identity cut-off of 40% and 70% and restrictive search score cut-off of 95%.** The meaning of the values in each line is described in chapter 5, subsection 5.2.2. (OrtAn Overview.csv output file).

Genome Set	**Set B**							
Database	**BTA**		**BTA P1**		**BTA P2**		**BTA P3**	
Relaxed Search identity cut-off	40	70	40	70	40	70	40	70
Annotation score cut-off	95							
Total orthogroups	8362							
KOs in the database	47		32		31		14	
Relaxed Search								
Selected orthogroups	69	40	49	31	52	29	14	11
% Selected orthogroups	0.8	0.5	0.6	0.4	0.6	0.3	0.2	0.1
Associated KOs	39	21	25	10	25	11	13	9
% Associated KOs	83.0	44.7	78.1	31.2	80.6	35.5	92.9	64.3
Restrictive Search/Annotation								
Orthogroups with annotated sequences	37	35	26	26	27	25	10	11
% of Orthogroups with annotated sequences	0.4	0.4	0.3	0.3	0.3	0.3	0.1	0.1
KOs with assigned sequences	22	19	9	9	11	10	12	9
% KOs with annotated sequences	46.8	40.4	28.1	28.1	35.5	32.3	85.7	64.3
ConOG	17	17	13	13	14	14	3	3
DivOG	20	18	13	13	13	11	7	8
DivOG with more than one KO	8	4	5	4	5	3	3	0
Relaxed Search to Restrictive Search								
Lost orthogroups	32	5	23	5	25	4	4	0
% Lost orthogroups	46.4	12.5	46.9	16.1	48.1	13.8	28.6	0.0
Lost KOs	17	2	16	1	14	1	1	0
% Lost KOs	43.6	9.5	64.0	10.0	56.0	9.1	7.7	0.0

results obtained with OrtAn, more specifically the ones present in the file Species_Annotation.csv. This file indicates which functions/KO groups were found in each species of the genome set. By combining this information, it is possible to calculate if a species has a minimum combination of genes that would allow the complete performance of one of the alternative paths (P1, P2 or P3) used in this work.

In a more detailed way, to determine if a species has the potential to perform a complete pathway, in each reaction it is verified if a minimum combination of genes/KO groups that enable the performance of the reaction (recurring for that to the GPR rules and the OrtAn annotation) is present in that species genome. For genome set A, the reference results are based on the annotation data from the KEGG database (the same data used to validate the annotation). For set B, the reference results correspond to the information regarding if the species are considered benzoate degraders or not.

The results presented here are only the ones obtained with OrtAn output that originated, in general, the best results in the annotation evaluation (an identity cut-off of 40% for the relaxed search, and a score cut-off of 90% for the restrictive search).

6.3.1 *Set A*

Since the results of the annotation using a score cut-off on the restrictive search of 95% were not satisfactory, in this step of the pipeline, it was decided to evaluate the pathway inference only with the cases using a score cut-off of 90%.

Also, for each one of the alternative paths inference, the OrtAn results obtained using the database containing the functions necessary to the respective path were used.

In Tables 13, 14, and 15, it is possible to compare, for path 1, 2 and 3 respectively, the expected reactions and those that were identified with OrtAn annotation pipeline. With the green boxes representing the correct results, it is possible to observe that most of the reactions were correctly previewed. Most of the errors (red boxes) are related to reactions that were not found in the species with the OrtAn annotation results but should have been. The only case where a reaction was found when it shouldn't have been, occurred in path 1, reaction R01422, and species code parb. These tables also allow us to easily understand which species were previewed to perform the complete path, that is, the lines/species were all the columns/reactions are filled with an x. It also allows to easily observe the cases were a reaction failed to be found in most of the species. A case to notice is that corresponding to the reaction R02451 (path1 and 2 only), which was not found in any of the 7 species where it should have been. According to the GPR rules, for this reaction to be performed 4 genes are required corresponding to the functions from the KO groups K04114, K04113, K04112, and K04115. These KO groups are represented in the database from only 18, 28, 24 and 18 sequences respectively. This could be the reason for this reaction not being found in any of the

Table 13.: **Reactions inference of BTA alternative path 1 in set A.** The calculations were made using the GPR rules relative to the BTA pathway and the annotation results obtained with OrtAn (database – BTA P1, identity cut-off - 40%, score cut-off - 90%). The reference results had into consideration the annotation present in KEGG. x – represents a reaction that was correctly found recurring to OrtAn annotation; o – represents a reaction not found but that should be. The green and red boxes help to visualize the correct and wrong assumptions, respectively.

Species code	Reaction code											
	R00238	R01422	R01976	R02451	R03026	R03028	R05305	R05579	R05581	R05586	R05594	R05597
adv	x		o		x		x			x		
ath	x		x		x		x			x		
aza	x	o	x	o	x		x		o	x	o	o
azd	x	x	x	o	x		x		o	x	o	o
azi	x	o	x	o	x		x		o	x	o	o
bced	x		x		x		x			x		
bvi	x		x		x		x			x		
cyq	x				x		x			x		
cza	x				x		x			x		
dor	x		x		x	o	x	x		x		
eba	x	o	x	o	x		x		o	x	o	o
lcm	x				x		x			x		
magx	x	x	x	o	x		x		x	x	x	x
parb	x	*	x		x		x			x		
rrz	x		x		x		x			x		
shd	x	o	x	o	x		x		o	x	o	o
sscu	o				o		o			x		
tmz	x	o	x	o	x		x		x	x	x	x

species, because a poor representation of a function in the database makes it more difficult for the identification of sequences with the same function in the genomes. The reaction R03028 (path 1) was not identified in any of the species. Other reactions, despite being identified, were only previewed to be performed in less than half of the species that were expected. For instance, R05581 (represented by only 14 sequences in the database), R05594 (represented by 17 sequences), and R05597 (represented by 17 sequences), path 1 and 2, were only shown to be performed by 2 of the expected 7 species).

Table 16 contains an assembly of the information from tables 13, 14, and 15. The second column indicates the paths that each species is described with the ability to fully perform (according to KEGG annotation) and the third column indicates the paths that were found via OrtAn pipeline. Path 3 was shown to be fully performed by 6 of the 8 species. Path 2 was found to not be completed by any 7 expected species. This leads to the assumption that there are insufficient genes, in any species, to perform some of the required reactions. Path 1 was also not found to be fully performed by any species as expected.

An explanation for the absence of species with the ability to perform the complete path 2 can be tied to low recall values (0.673) after using the BTA P2 database for annotation. When observing the recall values on the annotation evaluation for path 3, they are considerably better. Around 90% of the sequences that should be were effectively annotated, and that translates in better results when inferring the presence of path 3 in this set of species.

Using the species that were not able to perform any of the paths alone, it was tested if it was possible to find complete paths in different combinations of two, three and four species.

Table 14.: **Reactions inference of BTA alternative path 2 in set A.** The calculations were made using the GPR rules relative to the BTA pathway and the annotation results obtained with OrtAn (database – BTA P2, identity cut-off - 40%, score cut-off - 90%). The reference results had into consideration the annotation present in KEGG. x – represents a reaction that was correctly found recurring to OrtAn annotation; o – represents a reaction not found but that should be; * – represents a reaction found that wasn't present in the reference results. The green and red boxes help to visualize the correct and wrong assumptions, respectively.

Species code	Reaction code										
	R00238	R01422	R01976	R02451	R02488	R03026	R05305	R05581	R05586	R05594	R05597
adv	x		o		x	x	x		x		
ath	x		x			x	x		x		
aza	x	o	x	o	x	x	x	o	x	o	o
azd	x	x	x	o	x	x	x	o	x	o	o
azi	x	o	x	o	x	x	x	o	x	o	o
bced	x		x		x	x	x		x		
bvi	x		x		x	x	x		x		
cyq	x					x	x		x		
cza	x					x	x		x		
dor	x		x			x	x		x		
eba	x	o	x	o	x	x	x	o	x	o	o
lcm	x				x	x	x		x		
magx	x	x	x	o	x	x	x	x	x	x	x
parb	x	*	x		x	x	x		x		
rrz	x		x		x	x	x		x		
shd	x	o	x	o	x	x	x	o	x	o	o
sscu	o				x	o	o		x		
tmz	x	o	x	o	x	x	x	x	x	x	x

In combinations of two species, 89 of a total of 253 were found to be able to perform the path 3. For combinations of three, 615 of 816 and for combinations of four, 2646 of 3060. Path 1 and 2 were never reveal.

6.3.2 *Set B*

The analysis performed for genome set B was similar to that of genome set A, however, for set B the only information available was if the species was or not a benzoate degrader. In Tables 17, 18 and 19 it is shown which reactions were identified in each species, for path 1, 2 and 3 respectively. As in genome set A, path 3 is the only one where all the reactions were found in at least one species. For paths 1 and 2, some reactions, similar to the results with set A, were not identified in any of the species. The reactions R03028 and R02451 were not identified in any of the species of set A or set B. Additionally on set B, the reactions R05579, R05581, R05594, and R05597 were not identified either.

Due to the absence of annotation references, it cannot be affirmed that these results are due to the poor representation of the genes involved in reactions in the database, or if, effectively the genomes in analyses don't possess the genes necessary to perform the reactions.

In Table 20 it is possible to compare the information regarding the species that are considered benzoate degraders and the results obtained with the BTA inference from the OrtAn

Table 15.: **Reactions inference of BTA alternative path 3 in set A.** The calculations were made using the GPR rules relative to the BTA pathway and the annotation results obtained with OrtAn (database – BTA P3, identity cut-off - 40%, score cut-off - 90%). The reference results had into consideration the annotation present in KEGG. x – represents a reaction that was correctly found recurring to OrtAn annotation; o – represents a reaction not found but that should be. The green and red boxes help to visualize the correct and wrong assumptions, respectively.

Species code	Reaction code						
	R00228	R00750	R00813	R00816	R02601	R02604	R05621
adv	x	x	x	x	x	x	x
ath							
aza							
azd	x	x		x	x	x	
azi	x	x			x	o	
bced	x	x	x	x	x	x	x
bvi	x	x	x	x	x	x	x
cyq	x	x	x	x	x	x	x
cza	x	x	x	x	x	x	x
dor	x	x			x		
eba							
lcm							
magx							
parb	x	x	x	x	x	x	x
rrz	x	x	x	x	x	o	x
shd							
sscu				o			
tmz	x	x	x	x	x	o	x

Table 16.: **BTA alternative paths inference in set A.** The first column indicates the species code, the second the reference results calculated with KEGG annotation, and the third column indicates the results calculated with OrtAn annotation results (database - the one corresponding to the path in analyses, identity cut-off - 40%, score cut-off - 90%) .The green and red boxes on the third column help to visualize the correct and wrong assumptions, respectively.

Species code	Reference results	Results obtained with OrtAn annotation
adv	P3	P3
ath		
aza	P2	
azd	P2	
azi	P2	
bced	P3	P3
bvi	P3	P3
cyq	P3	P3
cza	P3	P3
dor		
eba	P2	
lcm		
magx	P2	
parb	P3	P3
rrz	P3	
shd	P2	
sscu		
tmz	P2, P3	

Table 17.: **Reactions inference of BTA alternative path 1 in set B.** The calculations were made using the GPR rules relative to the BTA pathway and the annotation results obtained with OrtAn (database – BTA P1, identity cut-off - 40%, score cut-off - 90%). x - represents a reaction that was expected to be performed.

Species code	Reaction code											
	R00238	**R01422**	**R01976**	**R02451**	**R03026**	**R03028**	**R05305**	**R05579**	**R05581**	**R05586**	**R05594**	**R05597**
A	x		x		x		x			x		
B	x		x									
C	x		x		x		x			x		
D	x	x	x		x		x			x		
F	x		x		x					x		
G	x		x		x		x			x		
H	x	x	x		x		x			x		
J	x	x	x		x		x			x		
I	x		x		x		x			x		
E	x	x	x		x		x			x		
K	x		x		x		x			x		
L	x		x		x		x					

Table 18.: **Reactions inference of BTA alternative path 2 in set B.** The calculations were made using the GPR rules relative to the BTA pathway and the annotation results obtained with OrtAn (database – BTA P2, identity cut-off - 40%, score cut-off - 90%). x - represents a reaction that was expected to be performed.

Species code	Reaction code										
	R00238	R01422	R01976	R02451	R02488	R03026	R05305	R05581	R05586	R05594	R05597
A	x		x			x	x		x		
B	x		x								
C	x		x			x	x		x		
D	x	x	x		x	x	x		x		
F	x		x		x	x			x		
G	x		x		x	x	x		x		
H	x	x	x		x	x	x		x		
J	x	x	x		x	x	x		x		
I	x		x		x	x	x		x		
E	x	x	x		x	x	x		x		
K	x		x		x	x	x		x		
L	x		x			x	x				

Table 19.: **Reactions inference of BTA alternative path 3 in set B.** The calculations were made using the GPR rules relative to the BTA pathway and the annotation results obtained with OrtAn (database – BTA P3, identity cut-off - 40%, score cut-off - 90%). x - represents a reaction that was expected to be performed.

Species code	Reaction code						
	R00228	R00750	R00813	R00816	R02601	R02604	R05621
A		x		x			
B	x			x			
C	x	x			x		
D	x	x			x		
F	x	x	x	x	x	x	x
G	x	x	x		x		x
H							
J							
I	x	x			x		
E	x	x	x	x	x	x	x
K	x	x	x	x	x		x
L	x	x	x	x	x	x	x

Table 20.: **BTA alternative paths inference in set B.** The first column indicates the species code, the second the indication if the species is a benzoate degrader or not, and the third column indicates the results calculated with OrtAn annotation results (database - the one corresponding to the path in analyses, identity cut-off - 40%, score cut-off - 90%) . The green and red boxes on the third column help to visualize the correct (species is a benzoate degrader and a complete path was found, or it is not, and no path was found) and wrong (the species is a benzoate degrader but no path was found or species is not a benzoate degrader but a complete path was found) assumptions, respectively.

Species Code	Benzoate Degrader	Results obtained with OrtAn annotation
A	No	
B	No	
C	No	
D	Yes	
E	Yes	P3
H	No	
J	No	
F	Yes	P3
G	Yes	
I	No	
K	Yes	
L	No	P3

annotation pipeline. Path 3 was found in three species, being only two of them considered benzoate degraders.

It was not possible to find combinations of species (excluding the ones where the path 3 was found) that would be able to perform any of the alternative paths.

Currently available data limits the ability to provide in-depth explanations for the obtained results. The fact that the pipeline didn't identify complete paths in the species considered benzoate degraders, could be due to low recall or imprecise annotations. Another possible explanation is that the species identified as benzoate degraders perform degradation using a different set of reactions not included in this study.

6.4 TOOLS PERFORMANCE

6.4.1 *OrtScraper*

The installation is simply done and the usage consists of a simple command. The process of downloading all the sequences from KEGG requires a lot of time, but that was optimized with the use of the library *grequests* that does not wait until a request has an answer to do the next one. But some additional features could make OrtScraper more useful, for example:

- Provide metadata about the KO groups being downloaded and the number of sequences downloaded for each one;
- Provide information about the interactions between reactions, enzymes, and pathways;
- Before creating the database, make a preview of the resources that will be needed to store all the data and attempt to preview the time left to finish all downloads (according to the requests in waiting and the time the responses are taking);
- Use the syntax of the KEGG database to automatically extract the GPR rules and store them together with the database (in the future, this could be used by OrtAn to perform the pathway inference right after the annotation without additional work for the user);
- Add new commands focused on extracting other types of information from KEGG database, for instance, lists of genes from a reaction, lists of species involved in a pathway, etc;
- Allow storing the database in a different format, for instance, the one used by DIAMOND. These options could also be convenient to OrtAn since it would save time creating the database for DIAMOND and space as there wouldn't be needed two databases of different formats containing the same data.

6.4.2 *OrtAn*

The installation is simple, as well as the usage. Despite having more commands, this allows the user to shape the criteria used. The reason to separate the `restrictive_search` and `annotation` is to allow the user to run the annotation more than one time, if need, without having to run the `restrictive_search` (the more time and space demanding task) as well. For instance, if the user chooses over-restrictive parameters for the annotation, the annotation process can be repeated with different values without performing all the DIAMOND searches of the restrictive search again.

The most challenging part of creating the OrtAn tool was the combination of the developed strategy with DIAMOND. The chosen strategy uses an organized database, and, mainly in the restrictive search step, there are various combinations of query files and databases that should be run against each other. Apart from DIAMOND's fast sequence alignment, its power is mostly exploited by using big query and database files. However, because of the OrtAn strategy, the sequences to be compared are divided into different processes of DIAMOND searches and the power of DIAMOND regarding its fast sequence alignment is not fully harnessed. Some adaptations were already made to take more advantage from DIAMOND's and revert this situation. In the restrictive search, when a series of various orthogroups needed to be compared against the same KO group, the sequences of these orthogroups were

put together in only one query file to avoid initiating a lot of different processes unnecessarily. This leads to a storage problem since sequences will be repeated in different FASTA files. A different method to store all the data and avoid having these unnecessary repetitions (same data in different files or formats) should be implemented in future versions of the pipeline. Thus, the most interesting update to OrtAn would be an improved adaptation of the tool strategy to DIAMOND or use other sequence alignment search tool.

7

CONCLUSIONS AND FURTHER WORK

The main objective of this work, create a pipeline for the annotation of genomes through COGs data and the inference of metabolic pathways, was fulfilled. All the pipeline testing, from the clustering to annotation and pathway inference was focused on one metabolic pathway, the benzoate to acetyl-CoA conversion. To evaluate and adjust the parameters used throughout the pipeline, a genome set from whose annotation was known was used. The final test of the pipeline was made with the genome set B where the only information available was about their capability as benzoate degraders.

The clustering, despite not resulting in perfect orthogroups (i.e., all the sequences in a orthogroup sharing the same function) helps in reducing the work performed in the subsequent steps of the pipeline. The annotation with OrtAn exhibited good precision results. However, one of the biggest problems of the pipeline was the low recall values of the annotation. This problem arises because of two main reasons:

- Some of the functions in the database had a small number of sequences representing them. This makes it more difficult to find these functions in a genome set with variability in their genomes and with genomes that vary from the ones present in the database.
- DIAMOND lower sensitivity when compared to traditionally-used sequence alignment search tools such as BLAST.

The homology search is key for the annotation. Due to the increase in the genomic data, the homology search tools were constructed to work on big sets of data, and their strategy was optimized for that. That is the purpose of DIAMOND as well, where some sensibility was compromised for faster results. The pipeline developed in this thesis is constantly reducing the number of sequences to be compared with these search tools. It starts with a database smaller than usual since the analyses are focused on only one metabolic pathway. In the relaxed search, only one per 10 sequences in each orthogroups is selected to be part of the query file. In the restrictive search, the orthogroups are only compared with groups of sequences that are already expected to be similar. Thus, the power of DIAMOND is potential not being completely explored, especially on the restrictive search when we only allow being compared

sequences that have a good chance of being a good match. Thus, the question is raised: is DIAMOND a good option for this pipeline? Or should it be used another homology search tool, that did not comprise so much sensibility, to get better annotation results?

In summary, improvements can and should be made to the developed pipeline to make it more attractive for users:

- Add to OrtScraper the ability to extract GPR rules automatically from the KEGG database and solve the cases where sequences with the same function are assigned to different KO groups;
- Add to OrtAn a feature for metabolic network inference;
- Add to OrtAn the option to use a more sensitive homology search tool;

BIBLIOGRAPHY

Parser for command-line options, arguments and sub-commands — Python 3.8.0 documentation. URL https://docs.python.org/3/library/argparse.html.

Beautiful Soup 4.4.0 documentation. URL https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

grequests · PyPI. URL https://pypi.org/project/grequests/.

Asynchronous HTTP Request Processing. URL https://docs.jboss.org/resteasy/docs/1.0.1.GA/userguide/html/Asynchronous_HTTP_Request_Processing.html.

GitHub - MartaLopesGomes/OrtAn, a. URL https://github.com/MartaLopesGomes/OrtAn.

GitHub - MartaLopesGomes/OrtScraper, b. URL https://github.com/MartaLopesGomes/OrtScraper.

Setuptools 41.4.0 documentation. URL https://setuptools.readthedocs.io/en/latest/.

Virtualenv 16.7.7 documentation. URL https://virtualenv.pypa.io/en/latest/.

Steven D Allison and Jennifer B H Martiny. Colloquium paper: resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 105 Suppl(Supplement_1):11512–9, 2008. ISSN 1091-6490. doi: 10.1073/pnas.0801925105. URL http://www.pnas.org/content/105/Supplement_1/11512.full.

Tomer Altman, Michael Travers, Anamika Kothari, Ron Caspi, and Peter D Karp. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC bioinformatics*, 14:112, 3 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-112. URL http://www.ncbi.nlm.nih.gov/pubmed/23530693http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3665663.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 10 1990. ISSN 00222836. doi: 10.1016/S0022-2836(05)80360-2. URL https://www.sciencedirect.com/science/article/pii/S0022283605803602?via%3Dihubhttp://www.sciencedirect.com/science/article/pii/S0022283605803602http://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

Stephen F Altschul, Thomas L Madden, A A Schäffer, J Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402, 9 1997a. ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/pdf/253389.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/9254694http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC146917.

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, 1997b. ISSN 03051048.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: Tool for the unification of biology, 5 2000. ISSN 10614036. URL http://www.ncbi.nlm.nih.gov/pubmed/10802651http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3037419http://www.nature.com/doifinder/10.1038/75556.

Ayansina Ayangbenro and Olubukola Babalola. A New Strategy for Heavy Metal Polluted Environments: A Review of Microbial Biosorbents. *International Journal of Environmental Research and Public Health*, 14(1):94, 1 2017. ISSN 1660-4601. doi: 10.3390/ijerph14010094. URL http://www.ncbi.nlm.nih.gov/pubmed/28106848http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5295344http://www.mdpi.com/1660-4601/14/1/94.

Ramakrishnan B. Microbial Diversity and Degradation of Pollutants. *Journal of Bioremediation and Biodegradation*, 03(12):12, 2012. ISSN 21556199. doi: 10.4172/2155-6199.1000e128. URL https://www.omicsonline.org/microbial-diversity-and-degradation-of-pollutants-2155-6199.1000e128.pdfhttps://www.omicsonline.org/2155-6199/2155-6199-3-e128.digital/2155-6199-3-e128.htmlhttps://www.omicsonline.org/microbial-diversity-and-degradation-of-po.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995. ISBN 978-0-19-853864-6. URL https://global.oup.com/academic/product/neural-networks-for-pattern-recognition-9780198538646?cc=pt&lang=en&.

Eva Boon, Conor J. Meehan, Chris Whidden, Dennis H.J. Wong, Morgan G.I. Langille, and Robert G. Beiko. Interactions in the microbiome: Communities of organisms and communities of genes, 2014. ISSN 01686445.

Elhanan Borenstein and Marcus W Feldman. Topological Signatures of Species Interactions in Metabolic Networks. *Journal of Computational Biology*, 16(2):191–200, 2 2009. ISSN 1066-5277. doi: 10.1089/cmb.2008.06TT. URL http://www.ncbi.nlm.nih.gov/pubmed/19178139http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3035845http://www.liebertonline.com/doi/abs/10.1089/cmb.2008.06TT.

Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 1 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3176. URL http://www.nature.com/articles/nmeth.3176.

Stuart H M Butchart, M. Walpole, B. Collen, A. van Strien, J. P. W. Scharlemann, R. E. A. Almond, J. E. M. Baillie, B. Bomhard, C. Brown, J. Bruno, K. E. Carpenter, G. M. Carr, J. Chanson, A. M. Chenery, J. Csirke, N. C. Davidson, F. Dentener, M. Foster, A. Galli, J. N. Galloway, P. Genovesi, R. D. Gregory, M. Hockings, V. Kapos, J.-F. Lamarque, F. Leverington, J. Loh, M. A. McGeoch, L. McRae, A. Minasyan, M. H. Morcillo, T. E. E. Oldfield, D. Pauly, S. Quader, C. Revenga, J. R. Sauer, B. Skolnik, D. Spear, D. Stanwell-Smith, S. N. Stuart, A. Symes, M. Tierney, T. D. Tyrrell, J.-C. Vie, and R. Watson. Global Biodiversity: Indicators of Recent Declines. *Science*, 328(5982):1164–1168, 5 2010. ISSN 0036-8075. doi: 10.1126/science.1187512. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1187512.

Bradley J. Cardinale. Biodiversity improves water quality through niche partitioning. *Nature*, 472(7341):86–91, 2011. ISSN 00280836. doi: 10.1038/nature09904. URL http://dx.doi.org/10.1038/nature09904.

Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 1 2016. ISSN 13624962. doi: 10.1093/nar/gkv1164. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1164.

Ron Caspi, Richard Billington, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E Midford, Quang Ong, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes. *Nucleic acids research*, 46(D1):D633–D639, 1 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx935. URL http://www.ncbi.nlm.nih.gov/pubmed/29059334http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5753197http://academic.oup.com/nar/article/46/D1/D633/4559117.

Feng Chen, Aaron J. Mackey, Jeroen K. Vermunt, and David S. Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4), 2007. ISSN 19326203. doi: 10.1371/journal.pone.0000383.

M O Dayhoff, P. J. McLaughlin, W. C. Barker, and L. T. Hunt. Evolution of sequences within protein superfamilies. *Die Naturwissenschaften*, 62(4):154–161, 4 1975. ISSN 0028-1042. doi: 10.1007/BF00608697. URL http://www.ncbi.nlm.nih.gov/pubmed/181273http://link.springer.com/10.1007/BF00608697.

Prasenjit Debbarma, M.G.H. Zaidi, Saurabh Kumar, Shikha Raghuwanshi, Amit Yadav, Yogesh Shouche, and Reeta Goel. Selection of potential bacterial strains to develop bacterial consortia for the remediation of e-waste and its in situ implications. *Waste Management*, 79:526–536, 9 2018. ISSN 0956053X. doi: 10.1016/j.wasman.2018.08.026. URL https://www.sciencedirect.com/science/article/pii/S0956053X18305130?via%3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S0956053X18305130.

Emms D.M. and Kelly S. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv*, page 466201, 2018. doi: 10.1101/466201. URL https://www.biorxiv.org/content/10.1101/466201v1.

J.W. Doran, M. Sarrantonio, and M.A. Liebig. Soil Health and Sustainability. In *Tropical Ecology*, volume 37, pages 1–54. Academic Press, 1 1996. ISBN 0564-3295. doi: 10.1016/S0065-2113(08)60178-9. URL https://www.sciencedirect.com/science/article/pii/S0065211308601789http://linkinghub.elsevier.com/retrieve/pii/S0065211308601789.

Jéssica Janzen dos Santos and Leila Teresinha Maranho. Rhizospheric microorganisms as a solution for the recovery of soils contaminated by petroleum: A review. *Journal of Environmental Management*, 210:104–113, 3 2018. ISSN 03014797. doi: 10.1016/j.jenvman.2018.01.015. URL https://www.sciencedirect.com/science/article/pii/S030147971830015X?via%3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S030147971830015X.

Sean R Eddy. A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics*, 23(1):205–11, 10 2009. ISSN 0919-9454. doi: 10.1142/9781848165632{_}0019. URL www.worldscientific.comhttp://www.worldscientific.com/doi/abs/10.1142/9781848165632_0019http://www.ncbi.nlm.nih.gov/pubmed/20180275.

SEAN R. EDDY, GRAEME MITCHISON, and RICHARD DURBIN. Maximum Discrimination Hidden Markov Models of Sequence Consensus. *Journal of Computational Biology*, 2(1):9–23, 1 1995. ISSN 1066-5277. doi: 10.1089/cmb.1995.2.9. URL

http://www.ncbi.nlm.nih.gov/pubmed/7497123http://www.liebertpub.com/doi/10.1089/cmb.1995.2.9http://www.liebertonline.com/doi/abs/10.1089/cmb.1995.2.9.

Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 10 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq461. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq461.

Jane Elith and John R. Leathwick. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 2009. ISSN 1543-592X. doi: 10.1146/annurev.ecolsys.110308.120159.

David M Emms and Steven Kelly. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(1):157, 8 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0721-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531804/pdf/13059_2015_Article_721.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/26243257http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4531804.

Jan Fassler and Peter Cooper. BLAST Glossary. *BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US)*, 2011.

Karoline Faust and Jeroen Raes. Microbial interactions: From networks to models, 2012. ISSN 17401526.

Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, 7(2):129–43, 2 2009. ISSN 1740-1534. doi: 10.1038/nrmicro1949. URL http://www.ncbi.nlm.nih.gov/pubmed/19116616http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3119670.

F. Fernández-Luqueño, C. Valenzuela-Encinas, R. Marsch, C. Martínez-Suárez, E. Vázquez-Núñez, and L. Dendooven. Microbial communities to mitigate contamination of PAHs in soil—possibilities and challenges: a review. *Environmental Science and Pollution Research*, 18(1):12–30, 1 2011. ISSN 0944-1344. doi: 10.1007/s11356-010-0371-6. URL http://link.springer.com/10.1007/s11356-010-0371-6.

Ingo Fetzer, Karin Johst, Robert Schäwe, Thomas Banitz, Hauke Harms, and Antonis Chatzinotas. The extent of functional redundancy changes as species' roles shift in different environments. *Proceedings of the National Academy of Sciences*, 112(48):14888–14893, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1505587112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1505587112.

Gabriel M. Filippelli, Martin Risch, Mark A.S. Laidlaw, Deborah E. Nichols, and Julie Crewe. Geochemical legacies and the future health of cities: A tale of two neurotoxins in urban soils. *Elementa: Science of the Anthropocene*, 3(0):000059, 7 2015. ISSN 2325-1026. doi: 10.12952/journal.elementa.000059. URL http://www.elementascience.org/articles/10.12952/journal.elementa.000059.

Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic acids research*, 42(Database issue):222–30, 1 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1223. URL http://www.ncbi.nlm.nih.gov/pubmed/24288371http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965110.

Walter M. Fitch. Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, 19(2):99, 1970. ISSN 00397989. doi: 10.2307/2412448. URL https://academic.oup.com/sysbio/article-lookup/doi/10.2307/2412448.

Shiri Freilich, Anat Kreimer, Isacc Meilijson, Uri Gophna, Roded Sharan, and Eytan Ruppin. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Research*, 38(12):3857–3868, 7 2010. ISSN 03051048. doi: 10.1093/nar/gkq118. URL http://www.ncbi.nlm.nih.gov/pubmed/20194113http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2896517.

Sharon Greenblum, Peter J Turnbaugh, and Elhanan Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences of the United States of America*, 109(2):594–9, 1 2012. ISSN 1091-6490. doi: 10.1073/pnas.1116053109. URL http://www.ncbi.nlm.nih.gov/pubmed/22184244http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3258644http://www.pnas.org/cgi/doi/10.1073/pnas.1116053109.

M Gribskov, A D McLachlan, and D Eisenberg. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 84(13):4355–8, 7 1987. ISSN 0027-8424. doi: 10.1111/j.1540-4781.1953.tb04118.x. URL http://www.ncbi.nlm.nih.gov/pubmed/3474607http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC305087http://doi.wiley.com/10.1111/j.1540-4781.1953.tb04118.x.

W.D. Hamilton. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):17–52, 7 1964. ISSN 00225193. doi: 10.1016/0022-5193(64)90039-6. URL https://www.sciencedirect.com/science/article/pii/0022519364900384http://linkinghub.elsevier.com/retrieve/pii/0022519364900396.

P.J. Harvey, M. Rouillon, C. Dong, V. Ettler, H.K. Handley, M.P. Taylor, E. Tyson, P. Tennant, V. Telfer, and R. Trinh. Geochemical sources, forms and phases of soil contamination in an industrial city. *Science of The Total Environment*, 584-585:505–514, 4 2017. ISSN 00489697. doi: 10.1016/j.scitotenv.2017.01.053. URL https://www.sciencedirect.com/science/article/pii/S0048969717300530?via%3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S0048969717300530.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 1992. ISSN 00278424. doi: 10.1073/pnas.89.22.10915.

Benjamin Horemans, Karolien Bers, Erick Ruiz Romero, Eva Pose Juan, Vincent Dunon, René De Mot, and Dirk Springael. Functional Redundancy of Linuron Degradation in Microbial Communities in Agricultural Soil and Biopurification Systems. *Applied and Environmental Microbiology*, 82(9):2843–2853, 5 2016. ISSN 0099-2240. doi: 10.1128/AEM.04018-15. URL http://www.ncbi.nlm.nih.gov/pubmed/26944844http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4836412http://aem.asm.org/lookup/doi/10.1128/AEM.04018-15.

Tim Hulsen, Martijn A. Huynen, Jacob de Vlieg, and Peter M.A. Groenen. Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, 7(4), 2006. ISSN 1474760X. doi: 10.1186/gb-2006-7-4-r31.

Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 12 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-119.

Lucian Ilie, Silvana Ilie, Shima Khoshraftar, and Anahita Mansouri Bigvand. Seeds for effective oligonucleotide design. *BMC Genomics*, 12(1):280, 12 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-280. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-280.

Chen Jiang, Yi Chen Lu, Jiang Yan Xu, Yang Song, Yue Song, Shu Hao Zhang, Li Ya Ma, Feng Fan Lu, Ya Kun Wang, and Hong Yang. Activity, biomass and composition of microbial communities and their degradation pathways in exposed propazine soil. *Ecotoxicology and Environmental Safety*, 145:398–407, 11 2017. ISSN 01476513. doi: 10.1016/j.ecoenv.2017.07.058. URL https://www.sciencedirect.com/science/article/pii/S0147651317304700?via%3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S0147651317304700.

Dazhi Jiao, Yuzhen Ye, and Haixu Tang. Probabilistic inference of biochemical reactions in microbial communities from metagenomic sequences. *PLoS computational biology*, 9(3):e1002981, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002981. URL http://omics.informatics.indiana.edu/mg/MetaNetSam/.http://www.ncbi.nlm.nih.gov/pubmed/23555216http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3605055.

Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, 11 2019. ISSN 0961-8368. doi: 10.1002/pro.3715. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3715.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 1 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw1092. URL http://www.ncbi.nlm.nih.gov/pubmed/27899662http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5210567https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1092.

Cheryl A. Kerfeld and Kathleen M. Scott. Using BLAST to teach "E-value-tionary" concepts. *PLoS Biology*, 2011. ISSN 15449173. doi: 10.1371/journal.pbio.1001014.

Ruchir A. Khandelwal, Brett G. Olivier, Wilfred F.M. Röling, Bas Teusink, and Frank J. Bruggeman. Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS ONE*, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0064567.

Ian Korf, Mark Yandell, and Joseph Bedell. *BLAST*. O'Reilly, 2003. ISBN 9780596002992 0596002998.

Martin C. Krueger, Hauke Harms, and Dietmar Schlosser. Prospects for microbiological solutions to environmental pollution with plastics. *Applied Microbiology and Biotechnology*, 99(21):8857–8874, 11 2015. ISSN 0175-7598. doi: 10.1007/s00253-015-6879-4. URL http://link.springer.com/10.1007/s00253-015-6879-4.

Stefan Kurtz, Adam Phillippy, A. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and S. Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004. ISSN 1465-6914. doi: 10.1186/gb-2004-5-2-r12. URL http://www.ncbi.nlm.nih.gov/pubmed/14759262http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC395750.

D. E. LAROWE, A. W. DALE, and P. REGNIER. A thermodynamic analysis of the anaerobic oxidation of methane in marine sediments. *Geobiology*, 6(5):436–449, 12 2008. ISSN 14724677. doi: 10.1111/j.1472-4669.2008.00170.x. URL http://www.ncbi.nlm.nih.gov/pubmed/18699783http://doi.wiley.com/10.1111/j.1472-4669.2008.00170.x.

Roie Levy and Elhanan Borenstein. Reverse Ecology: From Systems to Environments and Back. In *Advances in Experimental Medicine and Biology*, volume 751, pages 329–345. 2012. ISBN 9781461435662. doi: 10.1007/978-1-4614-3567-9{_}15. URL http://link.springer.com/10.1007/978-1-4614-3567-9_15.

Roie Levy and Elhanan Borenstein. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences*, 110(31):12804–12809, 7 2013. ISSN 0027-8424. doi: 10.1073/pnas.1300926110. URL http://www.ncbi.nlm.nih.gov/pubmed/23858463http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3732988http://www.pnas.org/cgi/doi/10.1073/pnas.1300926110.

Roie Levy, Rogan Carr, Anat Kreimer, Shiri Freilich, and Elhanan Borenstein. NetCooperate: A network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics*, 16(1):164, 5 2015. ISSN 14712105. doi: 10.1186/s12859-015-0588-y. URL http://www.ncbi.nlm.nih.gov/pubmed/25980407http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4434858.

Li Li, Christian J Jr Stoeckert, and David S Roos. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes – Li et al. 13 (9): 2178 – Genome Research. *Genome Research*, 13(9):2178–2189, 2003. ISSN 1088-9051. doi: 10.1101/gr.1224503.candidates. URL http://genome.cshlp.org/cgi/content/full/13/9/2178.

Qi Li, Feibi Lin, Chen Yang, Juanping Wang, Yan Lin, Mengyuan Shen, Min S Park, Tao Li, and Jindong Zhao. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming Microcystis-epibiont communities. *Frontiers in Microbiology*, 9(APR):746, 2018. ISSN 1664302X. doi: 10.3389/fmicb.2018.00746. URL http://www.ncbi.nlm.nih.gov/pubmed/29731741http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5919953.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 7 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158.

Weizhong Li, Limin Fu, Beifang Niu, Sitao Wu, and John Wooley. Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, 13(6):656–668, 2012. ISSN 14675463. doi: 10.1093/bib/bbs035.

Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics, 6 2015. ISSN 14710064. URL http://www.ncbi.nlm.nih.gov/

pubmed/25948244http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5204302.

Yaci Liu, Zhaoji Zhang, Yasong Li, Yi Wen, and Yuhong Fei. Response of soil microbial communities to roxarsone pollution along a concentration gradient. *Journal of Environmental Science and Health, Part A*, 52(9):819–827, 7 2017. ISSN 1093-4529. doi: 10.1080/10934529.2017.1281687. URL http://www.ncbi.nlm.nih.gov/pubmed/28276888https://www.tandfonline.com/doi/full/10.1080/10934529.2017.1281687.

Bin Ma, John Tromp, and Ming Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 3 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.3.440. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/18.3.440.

Fernando T. Maestre, Andrea P. Castillo-Monroy, Matthew A. Bowker, and Raúl Ochoa-Hueso. Species richness effects on ecosystem multifunctionality depend on evenness, composition and spatial pattern. *Journal of Ecology*, 100(2):317–330, 2012. ISSN 00220477. doi: 10.1111/j.1365-2745.2011.01918.x.

Sharmila S. Mande, Monzoorul Haque Mohammed, and Tarini Shankar Ghosh. Classification of metagenomic sequences: Methods and challenges. *Briefings in Bioinformatics*, 2012. ISSN 14675463. doi: 10.1093/bib/bbs054.

Raja Mazumder, Sona Vasudevan, and Anastasia N. Nikolskaya. Protein functional annotation by homology. *Methods in molecular biology (Clifton, N.J.)*, 484:465–90, 2008. ISSN 1064-3745. doi: 10.1007/978-1-59745-398-1{_}28. URL http://link.springer.com/10.1007/978-1-59745-398-1_28http://www.ncbi.nlm.nih.gov/pubmed/18592196.

Sara Mitri and Kevin Richard Foster. The Genotypic View of Social Interactions in Microbial Communities. *Annual Review of Genetics*, 47(1):247–273, 11 2013. ISSN 0066-4197. doi: 10.1146/annurev-genet-111212-133307. URL www.annualreviews.orghttp://www.annualreviews.org/doi/10.1146/annurev-genet-111212-133307.

Yuki Moriya, Masumi Itoh, Shujiro Okuda, Akiyasu C Yoshizawa, and Minoru Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*, 35(Web Server issue):182–5, 7 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm321. URL http://www.ncbi.nlm.nih.gov/pubmed/17526522http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1933193.

C P Mulder, D D Uliassi, and D F Doak. Physical stress and diversity-productivity relationships: the role of positive interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 98:6704–6708, 2001. ISSN 0027-8424. doi: 10.1073/pnas.111055298.

Lynne Reed Murphy, Anders Wallqvist, and Ronald M. Levy. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering, Design and Selection*, 13(3):149–152, 3 2000. ISSN 1741-0134. doi: 10.1093/protein/13.3.149. URL https://academic.oup.com/peds/article-lookup/doi/10.1093/protein/13.3.149.

Shahid Naeem. Biodiversity equals instability? *Nature*, 416(6876):23–24, 2002. ISSN 00280836. doi: 10.1038/416023a.

Bruno T. L. Nichio, Jeroniza Nunes Marchaukoski, and Roberto Tadeu Raittz. New Tools in Orthology Analysis: A Brief Review of Promising Perspectives. *Frontiers in Genetics*, 8(OCT):165, 10 2017. ISSN 1664-8021. doi: 10.3389/fgene.2017.00165. URL www.frontiersin.orghttp://journal.frontiersin.org/article/10.3389/fgene.2017.00165/full.

Irene Nobeli, Angelo D Favia, and Janet M Thornton. Protein promiscuity and its implications for biotechnology, 2 2009. ISSN 15461696. URL http://www.nature.com/articles/nbt1519.

Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 3 2010. ISSN 1087-0156. doi: 10.1038/nbt.1614. URL http://www.ncbi.nlm.nih.gov/pubmed/20212490http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3108565http://www.nature.com/articles/nbt.1614.

Andrea B. Pfisterer and Bernhard Schmid. Diversity-dependent production can decrease the stability of ecosystem functioning. *Nature*, 416(6876):84–86, 2002. ISSN 00280836. doi: 10.1038/416084a.

Daniel C. Reed, Christopher K. Algar, Julie A. Huber, and Gregory J. Dick. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proceedings of the National Academy of Sciences*, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1313713111.

Maido Remm, Christian E.V. Storm, and Erik L.L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052, 2001. ISSN 00222836. doi: 10.1006/jmbi.2000.5197.

C. Sansom. Database searching with DNA and protein sequences: an introduction. *Briefings in bioinformatics*, 1(1):22–32, 2 2000. ISSN 1467-5463. doi: 10.1093/bib/1.1.22. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/1.1.22http://www.ncbi.nlm.nih.gov/pubmed/11466971.

Michael C Schatz. Computational thinking in the era of big data biology. *Genome biology*, 13(11):177, 11 2012. ISSN 1474-760X. doi: 10.1186/

gb-2012-13-11-177. URL http://www.ncbi.nlm.nih.gov/pubmed/23194371http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3580488.

Florence Servant, Catherine Bru, Sébastien Carrère, Emmanuel Courcelle, Jérĵme Gouzy, David Peyruc, and Daniel Kahn. ProDom: automated clustering of homologous domains. *Briefings in bioinformatics*, 3(3):246–51, 9 2002. ISSN 1467-5463. doi: 10.1093/bib/3.3.246. URL http://www.ncbi.nlm.nih.gov/pubmed/12230033.

SHADI SHOKRALLA, JENNIFER L. SPALL, JOEL F. GIBSON, and MEHRDAD HAJIBABAEI. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21(8):1794–1805, 4 2012. ISSN 09621083. doi: 10.1111/j.1365-294X.2012.05538.x. URL http://doi.wiley.com/10.1111/j.1365-294X.2012.05538.x.

S. Sieuwerts, F. A. M. de Bok, J. Hugenholtz, and J. E. T. van Hylckama Vlieg. Unraveling Microbial Interactions in Food Fermentations: from Classical to Genomics Approaches. *Applied and Environmental Microbiology*, 74(16):4997–5007, 8 2008. ISSN 0099-2240. doi: 10.1128/AEM.00113-08. URL http://www.ncbi.nlm.nih.gov/pubmed/18567682http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2519258http://aem.asm.org/cgi/doi/10.1128/AEM.00113-08.

Christian J A Sigrist, Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at PROSITE. *Nucleic acids research*, 41(Database issue):344–7, 1 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1067. URL http://www.ncbi.nlm.nih.gov/pubmed/23161676http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3531220http://academic.oup.com/nar/article/41/D1/D344/1055798/New-and-continuing-developments-at-PROSITE.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 1981. ISSN 00222836. doi: 10.1016/0022-2836(81)90087-5.

Hyun-Seob Song, William Cannon, Alexander Beliaev, and Allan Konopka. Mathematical Modeling of Microbial Community Dynamics: A Methodological Review. *Processes*, 2(4):711–752, 10 2014. ISSN 2227-9717. doi: 10.3390/pr2040711. URL http://www.mdpi.com/2227-9717/2/4/711http://www.mdpi.com/2227-9717/2/4/711/.

Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big Data: Astronomical or Genomical? *PLOS Biology*, 13(7):e1002195, 7 2015. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002195. URL http://www.ncbi.nlm.nih.gov/pubmed/26151137http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4494865https://dx.plos.org/10.1371/journal.pbio.1002195.

Shuji Suzuki, Masanori Kakuta, Takashi Ishida, and Yutaka Akiyama. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PloS one*, 9(8):e103833, 8 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0103833. URL https://dx.plos.org/10.1371/journal.pone.0103833http://www.ncbi.nlm.nih.gov/pubmed/25099887http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4123905.

Susannah Green Tringe. Comparative Metagenomics of Microbial Communities. *Science*, 308(5721):554–557, 4 2005. ISSN 0036-8075. doi: 10.1126/science.1107851. URL http://www.ncbi.nlm.nih.gov/pubmed/15845853http://www.sciencemag.org/cgi/doi/10.1126/science.1107851.

UN-HABITAT and Earthscan Publications Ltd. State of the World's Cities 2010/2011, Bridging The Urban Divide. Technical report, 2011. URL https://unhabitat.org/books/state-of-the-worlds-cities-20102011-cities-for-all-bridging-the-urban-divide/.

Stjin van Dongen. *Graph clustering by flow simulation*. PhD thesis, 2000.

Diana H Wall and Johan Six. Give soils their due. *Science*, 347(6223):695–695, 2 2015. ISSN 0036-8075. doi: 10.1126/science.aaa8493. URL http://www.ncbi.nlm.nih.gov/pubmed/25678633http://www.sciencemag.org/cgi/doi/10.1126/science.aaa8493.

Diana H. Wall, Uffe N. Nielsen, and Johan Six. Soil biodiversity and human health. *Nature*, 528(7580):69–76, 11 2015. ISSN 0028-0836. doi: 10.1038/nature15744. URL http://www.nature.com/doifinder/10.1038/nature15744.

Stuart A. West, Ashleigh S. Griffin, Andy Gardner, and Stephen P. Diggle. Social evolution theory for microorganisms. *Nature Reviews Microbiology*, 4(8):597–607, 8 2006. ISSN 1740-1526. doi: 10.1038/nrmicro1461. URL http://www.nature.com/articles/nrmicro1461.

David Wheeler and Medha Bhagwat. BLAST QuickStart. 2007. doi: 10.1007/978-1-59745-514-5{_}9.

Yuzhen Ye and Thomas G. Doak. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Metagenomes. In Christos A. Ouzounis, editor, *Handbook of Molecular Microbial Ecology I*, volume 5, pages 453–460. John Wiley & Sons, Inc., Hoboken, NJ, USA, 5 2011. ISBN 9780470644799. doi: 10.1002/9781118010518.ch52. URL https://dx.plos.org/10.1371/journal.pcbi.1000465http://doi.wiley.com/10.1002/9781118010518.ch52.

Yuzhen Ye, Jeong-Hyeon Choi, and Haixu Tang. RAPSearch: a fast protein similarity search tool for short reads. *BMC bioinformatics*, 12:159, 5 2011. ISSN 1471-2105. doi: 10.

1186/1471-2105-12-159. URL http://www.ncbi.nlm.nih.gov/pubmed/21575167http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3113943.

Yongan Zhao, Haixu Tang, and Yuzhen Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(1):125–6, 1 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr595. URL http://omics.informatics.indiana.edu/mg/RAPSearch2/.http://www.ncbi.nlm.nih.gov/pubmed/22039206http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3244761.

Ali R. Zomorrodi and Costas D. Maranas. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Computational Biology*, 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002363.

Ali R. Zomorrodi, Mohammad Mazharul Islam, and Costas D. Maranas. D-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology*, 2014. ISSN 21615063. doi: 10.1021/sb4001307.

Xianchun Zou, Guijun Wang, and Guoxian Yu. Protein Function Prediction Using Deep Restricted Boltzmann Machines. *BioMed Research International*, 2017:1729301, 2017. ISSN 23146141. doi: 10.1155/2017/1729301. URL http://www.ncbi.nlm.nih.gov/pubmed/28744460http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5506480.

A

SUPPORT MATERIAL

Table 21.: **BTA database information.** All the selected reactions, the enzymes, the related KO groups and the number of gene sequences existent to each KO group are shown. Additionally, the total number of reactions, enzymes, KO groups, and gene sequences is also indicated.

Reaction ID	EC number	KO	Number of sequences
R01422	6.2.1.25	K04105	19
		K04110	68
R02451	1.3.7.8	K04113	28
		K04112	24
		K04114	18
		K04115	18
R05597	4.2.1.100	K07537	17
R05581	1.1.1.368	K07538	14
R05594	3.7.1.21	K07539	17
R05305	1.1.1.35	K01825	835
		K00022	258
		K07514	130
		K07547	5
		K07516	2044
		K01782	1815
		K08683	227
		K07548	6
		K15016	176
		K10527	369
R05586	2.3.1.16	K07508	272
		K00632	4110
		K07509	224
		K07513	699
R05579	1.3.99.32	K16173	26
R03028	7.2.4.5	K01615	45
R03026	4.2.1.17	K07515	246
		K01825	835
		K07514	130
		K07511	348
		K01782	1815
		K15016	176
		K01692	6554
		K13767	232
		K01715	1057
		K10527	369
R01976	1.1.1.157	K00074	4699
R00238	2.3.1.9	K00626	13376
R02488	1.3.8.6	K00252	2526
R05621	1.14.12.10	K05550	475
		K05549	499
R00813	1.3.1.25	K05783	424
R00816	1.13.11.2	K00446	530
		K07104	1188
R02604	3.7.1.9	K10216	143
R02601	4.2.1.80	K02554	923
		K18364	185
R00750	4.1.3.39	K01666	1112
		K18365	157
R00228	1.2.1.10	K00132	63
		K18366	156
		K04072	1571
		K04073	827
Total Reactions	20		
Total Enzymes	20		
Total KO groups	47		
Total Sequences	48755		

Table 22.: **BTA P1 database information.** All the selected reactions, the enzymes, the related KO groups and the number of gene sequences existent to each KO group are shown. Additionally, the total number of reactions, enzymes, KO groups, and gene sequences is also indicated.

Reaction ID	**EC number**	**KO**	**Number of sequences**
R01422	6.2.1.25	K04105	19
		K04110	68
R02451	1.3.7.8	K04114	18
		K04113	28
		K04112	24
		K04115	18
R05597	4.2.1.100	K07537	17
R05581	1.1.1.368	K07538	14
R05594	3.7.1.21	K07539	17
R05305	1.1.1.35	K10527	369
		K07516	2044
		K08683	227
		K00022	258
		K01825	835
		K15016	176
		K01782	1815
		K07547	5
		K07514	130
		K07548	6
R05586	2.3.1.16	K07509	224
		K07513	699
		K07508	272
		K00632	4110
R05579	1.3.99.32	K16173	26
R03028	7.2.4.5	K01615	45
R03026	4.2.1.17	K10527	369
		K01715	1057
		K01825	835
		K15016	176
		K01692	6554
		K07514	130
		K13767	232
		K01782	1815
		K07515	246
		K07511	348
R01976	1.1.1.157	K00074	4699
R00238	2.3.1.9	K00626	13376
Total Reactions	12		
Total Enzymes	12		
Total KO groups	32		
Total Sequences	37976		

Table 23.: **BTA P2 database information.** All the selected reactions, the enzymes, the related KO groups and the number of gene sequences existent to each KO group are shown. Additionally, the total number of reactions, enzymes, KO groups, and gene sequences is also indicated.

Reaction ID	**EC number**	**KO**	**Number of sequences**
R01422	6.2.1.25	K04105	19
		K04110	68
R02451	1.3.7.8	K04114	18
		K04113	28
		K04112	24
		K04115	18
R05597	4.2.1.100	K07537	17
R05581	1.1.1.368	K07538	14
R05594	3.7.1.21	K07539	17
R05305	1.1.1.35	K10527	369
		K07516	2044
		K08683	227
		K00022	258
		K01825	835
		K15016	176
		K01782	1815
		K07547	5
		K07514	130
		K07548	6
R05586	2.3.1.16	K07509	224
		K07513	699
		K07508	272
		K00632	4110
R02488	1.3.8.6	K00252	2526
R03026	4.2.1.17	K10527	369
		K01715	1057
		K01825	835
		K15016	176
		K01692	6554
		K07514	130
		K13767	232
		K01782	1815
		K07515	246
		K07511	348
R01976	1.1.1.157	K00074	4699
R00238	2.3.1.9	K00626	13376
Total Reactions	11		
Total Enzymes	11		
Total KO groups	31		
Total Sequences	40431		

Table 24.: **BTA P3 database information.** All the selected reactions, the enzymes, the related KO groups and the number of gene sequences existent to each KO group are shown. Additionally, the total number of reactions, enzymes, KO groups, and gene sequences is also indicated.

Reaction ID	**EC number**	**KO**	**Number of sequences**
R05621	1.14.12.10	K05549	499
		K05550	475
R00813	1.3.1.25	K05783	424
R00816	1.13.11.2	K07104	1188
		K00446	530
R02604	3.7.1.9	K10216	143
R02601	4.2.1.80	K18364	185
		K02554	923
R00750	4.1.3.39	K18365	157
		K01666	1112
R00228	1.2.1.10	K18366	156
		K04073	827
		K00132	63
		K04072	1571
Total Reactions	7		
Total Enzymes	7		
Total KO groups	14		
Total Sequences	8253		