Genome-scale Evaluation of the Biotechnological Potential of Red Sea Bacilli Strains

Dissertation by
Ghofran Othoum

In Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy

King Abdullah University of Science and Technology
Thuwal, Kingdom of Saudi Arabia

February, 2018
The dissertation of Ghofran Othoum is approved by the examination committee.

Committee Chairperson: Prof. Vladimir B. Bajic
Committee Members: Prof. Alan Christoffels., Prof. Mani Sarathy, Prof. Takashi Gojobori
ABSTRACT

Genome-scale Evaluation of the Biotechnological Potential of Red Sea Bacilli Strains

Ghofran Othoum

The increasing spectrum of multidrug-resistant bacteria has caused a major global public health concern, necessitating the discovery of novel antimicrobial agents. Additionally, recent advancements in the use of microbial cells for the scalable production of industrial enzymes has encouraged the screening of new environments for efficient microbial cell factories. The unique ecological niche of the Red Sea points to the promising metabolic and biosynthetic potential of its microbial system. Here, ten sequenced Bacilli strains, that are isolated from microbial mat and mangrove mud samples from the Red Sea, were evaluated for their use as platforms for protein production and biosynthesis of bioactive compounds. Two of the species (B. paralicheniformis Bac48 and B. litoralis Bac94) were found to secrete twice as much protein as Bacillus subtilis 168, and B. litoralis Bac94 had complete Tat and Sec protein secretion systems. Additionally, four Red Sea Species (B. paralicheniformis Bac48, Virgibacillus sp. Bac330, B. vallismortis Bac111, B. amyloliquefaciens Bac57) showed capabilities for genetic transformation and possessed competence genes. More specifically, the distinctive biosynthetic potential evident in the genomes of B. paralicheniformis Bac48 and B. paralicheniformis Bac84 was assessed and compared to nine available complete genomes of B. licheniformis and three genomes of B. paralicheniformis. A uniquely-structured trans-acyltransferase (trans-AT) polyketide synthase/nonribosomal peptide synthetase (PKS/NRPS) cluster in strains of this species was identified in the genome of B. paralicheniformis 48. In total,
the two *B. paralicheniformis* Red Sea strains were found to be more enriched in modular clusters compared to *B. licheniformis* strains and *B. paralicheniformis* strains from other environments. These findings provided more insights into the potential of *B. paralicheniformis* 48 as a microbial cell factory and encouraged further focus on the strain’s metabolism at the system level. Accordingly, a draft metabolic model for *B. paralicheniformis* Bac48 (iPARA1056) was reconstructed, refined, and validated using growth rate and growth phenotypes under different substrates, generated using high-throughput Phenotype Microarray technology. The presented studies indicate that several of the isolated strains represent promising chassis for the development of cell factories for enzyme production and also point to the richness of their genomes with specific modules of secondary metabolism that have likely evolved in Red Sea Bacilli due to environmental adaptation.
ACHIEVEMENTS

Journal Publications


Conference Presentations


*Equal contribution*
ACKNOWLEDGEMENTS

My sincerest gratitude goes to my supervisor, Prof. Vladimir Bajic. Throughout the years, he showed me how adopting a pragmatic and professional attitude, in different circumstances, often yield the most satisfying and fruitful outcomes. I am forever grateful for his supervision, guidance, and support.

I would also like to thank everyone at the Computational Biology Research Center (CBRC) for their direct or indirect contribution and assistance. Particularly, I would like to thank Dr. Magbubah Essack, Dr. Salim Bougouffa, and Dr. Rozaimi Razali for brainstorming sessions, tips for good writing, and overall assistance.

Additionally, thanks ought to be expressed to those involved in sequencing and maintaining the isolates, before the onset of the project, mainly Dr. Soha Al-Amoudi, Ameerah Bokhari, and Dr. Feras Lafi.

The project overlaps with a collaborative effort with the Sysbio group in Chalmers University, Sweden. Accordingly, I would like to thank Dr. Ivan Mijakovic, Dr. Sylvain Prigent and Dr. Abderahmane Derouiche for their collaboration, assistance, and insights.

I am also grateful to Dr. Marta Simoes, Dr. Andre Antunes and Mohammed Alarawi for helping me transition from the world of computational systems biology to the intriguing world of experimental microbiology. I would also like to thank Dr. Ramona Marasco and Prof. Daniele Daffonchio for allowing me to use the Biolog machine in their lab as well as loading the samples for analysis.

Finally, I would like to thank my family, especially my mother for her continuous support throughout this journey.
# TABLE OF CONTENTS

EXAMINATION COMMITTEE PAGE .............................................................................. 2
COPYRIGHT PAGE ........................................................................................................ 3
ABSTRACT ...................................................................................................................... 4

ACHIEVEMENTS ........................................................................................................... 6
Journal Publications ......................................................................................................... 6
Conference Presentations ................................................................................................. 6

ACKNOWLEDGEMENTS ............................................................................................... 7

TABLE OF CONTENTS ................................................................................................ 8

LIST OF ABBREVIATIONS ............................................................................................ 10

KEYWORDS ..................................................................................................................... 11

LIST OF ILLUSTRATIONS ........................................................................................... 12

LIST OF TABLES ............................................................................................................. 14

DECLARATION ................................................................................................................ 15

Chapter 1 ....................................................................................................................... 16

1 Introduction and Background .................................................................................. 16

1.1 Microbial Cell Factories: Evaluation and Design .................................................. 17

1.1.1 Constrained-based Modeling for the Evaluation of Metabolic Potential .......... 20

1.1.2 Genome Mining as a Screening Method for Biosynthetically Plausible Microbial Cells .............................................................................................................. 23

1.2 *Bacillus*, A Biotechnologically Rich Genus ......................................................... 25

1.3 Thesis Outline ........................................................................................................... 32

Chapter 2 ....................................................................................................................... 34

2 A Comparative Genomics Study Reveals the Environment-dependent Potential of the Red Sea Bacilli for Production of Secondary Metabolites 34

2.1 Material and Methods ............................................................................................ 35

2.1.1 DNA Extraction and Sequencing ........................................................................ 35

2.1.2 Genome Assembly ............................................................................................... 36

2.1.3 Genome Functional Annotation and Analysis .................................................... 36

2.1.4 Bacterial Strains and Media ................................................................................ 36

2.1.5 Competence Assay .............................................................................................. 37

2.1.6 Sporulation Assay ............................................................................................... 38

2.1.7 Protein Secretion Assay ...................................................................................... 38

2.1.8 Phylogeny Analysis ............................................................................................. 39

2.1.9 Identification of Competence, Protein Secretion and Sporulation Genes ....... 39

2.1.10 Metabolic Reconstruction .................................................................................. 40

2.1.11 Biosynthetic Gene Clusters Prediction ............................................................... 40

2.2 Results and Discussion ......................................................................................... 41

2.2.1 Bacilli Present in the Red Sea Originate from Several Colonization Events .... 41

2.2.2 Several Bacilli Isolated from the Red Sea Present Substantial Potential for Protein Secretion ................................................................................................. 43

2.2.3 Convergent Evolution of Metabolic Networks in Bacilli ................................ 47

2.2.4 Secondary Metabolism of the Red Sea Bacilli Presents Considerable Potential for Production of Novel Active Molecules ......................................................... 50
Chapter 3: Exploring the Biosynthetic Potential of Bacillus paralicheniformis strains from the Red Sea ................................................................. 55

3.1 Materials and Methods ............................................................................................ 57
   3.1.1 Genomic Comparison ...................................................................................... 57
   3.1.2 Strain Identification and Phylogeny ............................................................... 57
   3.1.3 Biosynthetic Gene Cluster Prediction ............................................................ 58
   3.1.4 Urease Activity Testing ................................................................................. 59

3.2 Results and Discussion ............................................................................................ 59
   3.2.1 Genome Sequencing and Assembly ............................................................. 59
   3.2.2 Phylogenetic Positioning of the Red Sea Bacillus paralicheniformis Strains 65
   3.2.3 Genomic Overview of B. licheniformis and B. paralicheniformis Strains ..... 67
   3.2.4 Exploring the Biosynthetic Potential of B. paralicheniformis Bac48 and B. paralicheniformis Bac84 ........................................................................ 71

Chapter 4: A Draft Genome-Scale Metabolic Model of B. paralicheniformis Bac4882 82

4.1 Methods .................................................................................................................. 83
   4.1.1 Bacterial Preparation for Phenotype Microarray Plates Inoculation .............. 83
   4.1.2 Model Reconstruction and Simulation ........................................................... 84

4.2 Results and Discussion ............................................................................................ 84
   4.2.1 Phenotypic Characterization of B. paralicheniformis Bac48 ....................... 84
   4.2.2 Model Reconstruction .................................................................................... 85
   4.2.3 Model Validation .......................................................................................... 89
   4.2.4 Gene Essentiality Analysis ......................................................................... 95
   4.2.5 Case Study: Genetic Targets for the Overproduction of Bacitracin Precursors ........................................................................................... 96

Chapter 5: Concluding Remarks ..................................................................................... 99

5 REFERENCES .............................................................................................................. 102

SUPPLEMENTARY ......................................................................................................... 121

Supplementary 1. Composition of Simulated Growth Media ........................................ 121
Supplementary 2. Biolog Phenotype Data .................................................................... 122
   S2.1. M9 Media Composition ................................................................................. 122
   S2.2. Growth Phenotypes Images ........................................................................ 123
Supplementary 3. Urease Christensen Urea Agar Media ............................................. 125
LIST OF ABBREVIATIONS

BGC  biosynthetic gene cluster
COBRA  constraint based reconstruction and analysis
COG  clustered orthologous group
FBA  flux balance analysis
GCF  gene cluster family
GEM  genome-scale metabolic model
HGT  horizontal gene transfer
HMM  hidden Markov Model
iPARA1056  in silico B. paralicheniformis Bac48, 1056 genes
LB  Luria-Bertani
MCF  microbial cell factory
MeLan  Methylanthionine
MGF  minimum genome factory
MOMA  minimization of metabolic adjustment
NRPS  nonribosomal peptide synthetase
pHMM  profile hidden Markov Model
Lan  Lanthionine
PKS  polyketide synthase
ROOM  regulatory on/off minimization
KEYWORDS

Antimicrobials

*Bacillus* genus

*Bacillus licheniformis*

*Bacillus paralicheniformis*

Bacteriocins

Biosynthetic gene clusters

Flux balance analysis

Genome mining

Genome-scale metabolic models

Industrial enzymes

*In silico* strain design

Lanthipeptides

Metabolic networks

Microbial cell factories

Nonribosomal peptides

Polyketides
Figure 1.1. Project outline describing the pruning approach of the ten Red Sea isolates and showing the evaluation methods at each stage............................... 33

Figure 2.1. Phylogenetic tree of the 10 Red Sea Bacilli and 30 other species. The Red Sea species are displayed in grey while previously sequenced Bacilli are displayed in black................................................................. 42

Figure 2.2. Evaluation of sporulation in the Red Sea strains through gene prediction and in vitro measurements. ................................................................. 44

Figure 2.3. Evaluation of protein secretion in the Red Sea strains through gene prediction for the Tat, Sec and Sip secretion systems and in vitro measurements. ............................................................................. 45

Figure 2.4. Transformation efficiencies of Red Sea Bacilli with pMAD and pJC15 ..... 46

Figure 2.5. Metabolic networks reconstruction. The left part of the figure corresponds to the number of reactions present in the metabolic networks. The right part of the figure corresponds to a clustering performed on the 32 species based on the presence and absence of reactions. The 32 species have been divided into 7 metabolic clades based on this clustering................................. 48

Figure 2.6. Number of predicted secondary metabolic gene clusters. ....................... 50

Figure 2.7. Networks of co-occurrence of presence of secondary metabolism gene clusters. Red sea species are displayed in boxes in different shades of green. Other species are displayed as ellipses. The product of a given secondary-metabolic gene cluster is displayed when identified. Unknown gene clusters are transparent......................................................................................... 51

Figure 2.8. Co-localization of BGCs in studied Bacilli strains................................. 53

Figure 3.1. Circular plots of (a) B. paralicheniformis Bac48 and (b) B. paralicheniformis Bac84 genomes, showing the distribution of genomic islands, prophages and biosynthetic genes in the genomes. The tracks show the following features starting from the outermost track; 1st track (pink): genes on the positive strand; 2nd track (blue): genes on the negative strand; 3rd track (yellow): biosynthetic gene clusters; 4th track (red): horizontally transferred genes; 5th track (green): genes in prophage regions; 6th track: GC-plot where purple and green correspond to below and above average GC content, respectively; 7th track: GC-skew where purple and green correspond to below and above average GC-skew, respectively ......................................................... 63

Figure 3.2. Maximum-likelihood phylogenetic tree of 38 genomes of the class Bacilli constructed using single-copy core genes. Clostridioideae difficile CD196 was used as the outgroup. ......................................................................................... 66

Figure 3.3. Maximum-likelihood phylogenetic tree of 36 genomes of the class Bacilli constructed using orthogroups. Clostridioideae difficile CD196 was used as the outgroup......................................................................................... 67
**Figure 3.4.** Intersection of orthologous gene families between the 14 genomes.
Genomes are represented with filled circles. The core genome is represented in the first bar while unique genes are represented in bars with unconnected filled circles. The figure was generated using UpsetR [296] and the bars were ordered by frequency. 

**Figure 3.5.** Estimation of the Pan and Core genomes sizes ........................................ 69

**Figure 3.6.** COG assignments in the core genome showing the conservation of housekeeping functions. ........................................................................................................ 69

**Figure 3.7.** Distribution of genes in biosynthetic gene clusters in nine B. licheniformis and five B. paralicheniformis genomes. Clusters with modular genes are marked with a star and clusters encoding for ribosomally synthesized peptides are marked with a triangle ......................................................................................... 72

**Figure 3.8.** Similarity network showing 54 groups of similar BGCs. Strains are color coded as per the legend. A product is assigned - shown on top of each group of nodes- if the clusters in the group share more than 60% similarity to the product. .................................................................................................................. 73

**Figure 3.9.** Heat map visualization of the number of genes in BGC groups. There are 54 GCFs with BGCs shared by at least two genomes and 20 BGCs identified to be unique (present in one genome). The number of genes in each GCF is normalized based on the maximum number of genes. Putative clusters are predicted using the ClusterFinder algorithm as implemented in antiSMASH [107] ..................................................................................................................... 74

**Figure 3.10.** Structure of the hybrid PKS/NRPS cluster present in B. paralicheniformis Bac48. Biosynthetic genes are identified with red arrows while non-biosynthetic genes are identified with blue ones. Domains are abbreviated as follows: adenylation (A), ketosynthase (KS), ketoreductase (KR), condensation domain (C), acyl carrier protein (ACP), peptidyl carrier domains (PCP), c-methyltransferase (cMTA), o-methyltransferase (oMT, enoyl-CoA hydratases (ECH), dehydratase (DH), acyltransferase docking site (Trans-AT docking) and acyltransferase (AT) ........................................................................................................ 75

**Figure 3.11.** Homology between the hybrid PKS-NRPS cluster and other clusters with known products.............................................................................................................. 77

**Figure 3.12.** Similarity between the genomes of B. paralicheniformis Bac48 and B. paralicheniformis Bac84. a) Circos figure showing synteny blocks between B. paralicheniformis Bac48 and B. paralicheniformis Bac84. Regions I, II and III are regions in Bac48 that are missing in B. paralicheniformis Bac84. The coordinates of the two largest regions are from 1,884,514 to 1,968,250 with a total length of 83,736 bp and from 3,977,600 to 3,997,816 with a total length of 20,216 bp. b) A dotplot between B. paralicheniformis Bac48 and B. paralicheniformis Bac84 genomes which shows the high concordance between the two genomes. ....... 79

**Figure 4.1.** Maximum height of growth curve for B. paralicheniformis Bac48 under different concentrations of NaCl ........................................................................ 85

**Figure 4.2.** Distribution of reactions, genes and metabolites in iPARA1056 ............. 87

**Figure 4.3.** Experimental growth curve of B. paralicheniformis Bac48. The vertical axis is log₂ of absorbance at OD600 and horizontal axis is time in hours. ....... 90
LIST OF TABLES

Table 1.1. Summary of different in silico strain design methods................................. 21
Table 2.1. Average and standard deviation of three measurements of number of heat-resistant spores (10^-4).......................................................................................... 44
Table 2.2. Average and standard deviation of three measurements of secreted protein for each Red Sea species ........................................................................................................... 45
Table 2.3. Measurements for transformation efficiencies for Red Sea strains in number of transformants per microgram of DNA................................................................. 46
Table 3.1. Basic statistics relating to the PacBio SMRT sequencing that was done for B. paralicheniformis B48 and B84. A single SMRT cell was sequenced for each strain. .............................................................................................................................. 61
Table 3.2. Levels of completeness and contamination in B. paralicheniformis Bac48 and B. paralicheniformis Bac84 as determined in CheckM................................................. 61
Table 3.3. Summary of the genomes and annotation of nine B. licheniformis strains and five B. paralicheniformis strains. ................................................................. 62
Table 3.4. List of genomic island regions in the genomes of B. paralicheniformis Bac48 and B. paralicheniformis Bac84, predicted using IslandViewer [213]. .... 64
Table 3.5. Predicted prophage regions in B. paralicheniformis Bac48 and B. paralicheniformis Bac84 and their overlap with GIs. Scores were obtained using PHASTER [293] scoring scheme. Most Common Phage shows the phage ID(s) with the highest number of proteins most similar to proteins in the region. Overlap percentage shows the length of overlap region with respect to the length of prophage......................................................................................................... 64
Table 4.1. COG assignments for genes in the metabolic network .................................. 88
Table 4.2. Computational and experimental growth (G) and no growth (NG) phenotypes in different carbon sources. TP refers to true positive, TN refers true negative, FN refers false negative and FP refers to false positive. ............ 92
Table 4.3. Computational and experimental growth (G) and no growth (NG) phenotypes in different nitrogen substrates. TP refers to true positive, TN refers to true negative, FN refers to false negative and FP refers to false positive. .... 93
DECLARATION

I hereby declare that this dissertation is my original work and it has been written by me in its entirety.
I have duly acknowledged all the sources of information, which have been used in the dissertation.
Ghofran Othoum
Chapter 1

1 Introduction and Background

There is now an unprecedented enthusiasm towards the use of microbial cells for bio-based production of industrial enzymes, pharmaceutical antimicrobials, and energy fuels [1-3]. This interest is partly motivated by the depletion of some of the current resources used in industry [4] and by the reduction of environmental costs associated with the use of microbial cell factories (MCFs) [5-7]. Another growing issue is the emergence of multi-drug resistant pathogens, an inevitable consequence of the unsupervised use of antibiotics in animal husbandry and over-prescription of medicinal antibiotics [8-11]. This problem is further complicated by the limited number of new classes of antibiotics and the slow rate at which novel antibiotics are being approved [10, 12].

The factors mentioned above motivate bioprospecting new environments for new microbial systems that have the metabolic infrastructure to produce useful bioactive compounds. Accordingly, in this study, Bacilli strains isolated from the Red Sea are screened for their potential as platform MCFs capable of producing antimicrobials and enzymes of biotechnological relevance. For the evaluation, the strains’ production capabilities, as shown by their genomes’ repertoire, their ability to take exogenous DNA (competence), their protein secretion and their sporulation were considered. Additionally, a draft metabolic model was built for one of the 10 Red Sea strains (*Bacillus paralicheniformis* Bac48) and offered as a computational guide for the future generation of metabolic engineering strategies.
1.1 Microbial Cell Factories: Evaluation and Design

Using microbial systems for the efficient and optimal production of selected industrial compounds provides an economical and environmentally-friendly alternative to other industrial routes [13]. There are already several such products produced through industrial microbiology that include pharmaceuticals [3, 14], amino acids [15, 16], biofuels [17-21], bioplastics [22] and platform chemicals [23-25]. Conventionally, the range of products produced is limited by obstacles that hinder the utilization of microbial systems for industrial-scale production, including low product yield and low strain fitness in industrial settings [26]. However, the advancement in genome sequencing technology and the development of genome manipulation tools allowed the ‘smart’ and controlled manipulation of microbial cells by mapping differences in metabolism to single nucleotide mutations. Compared to traditional methods (e.g., random mutagenesis), rationally-designed and metabolically-engineered MCFs [27, 28] are more efficient and have a higher probability of reaching the desired phenotype without compromising microbial growth and well-being [29, 30].

One of the most critical steps for a successful metabolic engineering experiment is the selection of an optimal microbial cell. The criteria for this ranking can be summarized in four main points: high production yield, genomic stability, minimal genomic interventions necessary, and low-cost requirements [31]. Intuitively, a good MCF should have the potential to produce the target product at high yield [32, 33], with high expression of genes that increase production, low or no expression of genes that negatively affect the production, and high metabolic flux for ATP regeneration [34]. Additionally, the bacterium is preferably able to grow anaerobically
or in microaerobic conditions, as this is expected to lead to a lower production rate of byproducts and a reduction in unwanted consequences associated with heat from aerobic respiration [35-37]. Moreover, the bacterium should be able to utilize a wide array of substrates for fermentation, including those in abundant feedstock (lignocellulosic substrates) [38-40], especially those that are not in competition with human food sources [41]. Furthermore, MCFs ought to tolerate toxic and inhibitory byproducts produced in the pretreatment of lignocellulose sources [38, 42] and/or live in a reduced-cost medium [43]. Some additional metabolic capabilities that might increase the efficiency and reduce the costs associated with MCFs include photosynthesis, CO₂ fixation, methanogenesis, and nitrogen fixation [44].

Although the factors mentioned above collectively increase the efficiency of MCFs, it is almost impossible to find a microbial cell that naturally has all the desired features. The logic applied is that the more favorable metabolic capabilities the strain has, the less genetic intervention it requires. Microbial cells are usually artificially modified by inducing desirable phenotypes, preferably with the least number of interferences required [45]. It is often the case that some non-native enzymatic components [44, 46], or their resulting intermediates [47], might be toxic to the bacterial host into which they are introduced, requiring further intervention [35] or the selection of another host. Moreover, the presence of a complete native biosynthetic pathway in a strain is not necessarily indicative of its ability to produce the target efficiently and/or at sufficient yields [48]. The main setback of using natural microbial systems is not the presence of incomplete native biosynthetic pathways but rather the low quantities at which compounds and proteins of interest are produced.
In some cases, to produce the compound at a rate feasible for commercialization, enzymatic and metabolic components are synthetically incorporated in the MCF [49-51]. Additionally, some instances require limiting carbon flux to nonessential competing pathways and thus re-directing carbon flux toward the target biosynthetic pathway [52, 53]. Equally, as the introduction of novel genetic components may be needed, the removal of unnecessary elements may also be required, including the removal of prophages and biosynthetic gene clusters.

Other instances for the use of MCFs entail the introduction of a complete heterologous pathway in a desired chassis, often through the utilization of promiscuous enzymes [54-56]. There are cases where *de novo* heterologous pathways are used. It could be the case that the compound of interest is produced by an uncultivatable microorganism or produced by a non-microbial organism (mostly plants), or it is not a natural product but rather a product of a synthetic metabolic pathway. In some cases, the compound of interest is produced naturally in a cultivatable microorganism, albeit in low quantities, urging the introduction of the biosynthetic pathway into other favorable hosts.

The availability of genomic tools for model organisms, mainly *Escherichia coli* in Gram-negative bacteria and *Bacillus subtilis* in Gram-positive bacteria, is what often directs researchers and industrial experimentalists to use these well-studied microorganisms as platforms. However, the availability of methods to reconstruct entire genomes using the approach described in 2009 by Gibson et al. [57], and the advancements in tools for the development of optimized and minimized genomes [58], increased the spectrum of cells as putative microbial hosts.
1.1.1 Constrained-based Modeling for the Evaluation of Metabolic Potential

Significant developments in Systems Biology allow the exploration of the cellular metabolism of a strain using reconstructed metabolic networks at the genome scale [59]. These reconstructions are based on annotated genomes obtained from current knowledge bases that link information of genomic, biochemical and metabolic nature into the resulting so-called genome-scale models [59, 60]. Kinetic modeling of complicated metabolic networks is a difficult task, as it requires the quantification of a large number of parameters (e.g., enzymatic kinetic rates) that are frequently not known [61]. A more practical alternative is constraint-based modeling, in which constraints are used to reduce the solution space, allowing the exploration of the theoretical flux of each reaction in a given metabolic network [62, 63].

The most common types of constraints in metabolic models are: physiochemical ones, including mass conservation and thermodynamics [64], topological ones as metabolites and associated enzymes are located in different cellular compartments, and environmental ones that reflect the availability of suitable nutrition and necessary exchange elements.

The ability to mathematically represent these networks, in stoichiometric matrices, has allowed the development of a number of in silico modeling methods to predict different phenotypes related to the growth rate of the cell and/or the production rate of any metabolite in the network. One of these methods is flux balance analysis (FBA), which computes metabolites flow through the network, assuming the model is at a pseudo-steady state with no change in the concentration of metabolites in the model [65]. FBA also assumes that the objective function, such
as growth rate or ATP production maximization [66], is optimized through evolution. These assumptions facilitate the study of the effect of perturbations on the cell, both external perturbations and network-specific ones. More constraints can be considered to assure that predicted fluxes of the engineered strain are reflective of the natural state of the cell, immediately after the introduction of these manipulations. Two of the most significant approaches are minimization of metabolic adjustment (MOMA) [67] and regulatory on/off minimization (ROOM) [68]. MOMA aims at minimizing the Euclidean distance between the metabolite fluxes in the wild-type strain and the optimized strain, while optimizing the target product. On the other hand, ROOM minimizes the number of regulatory changes that cause a significant range of flux changes.

One of the challenging tasks in metabolic engineering is the process of modifying cellular metabolism to direct carbon flux toward a specific pathway of interest. With the availability of numerous metabolically reconstructed models, along with different reaction databases (e.g., ModelSEED [69], Path2Models [70] and MetaNetX [71]), it is now possible to develop biologically rational strategies for the design of strains capable of overproducing biochemical compounds in a metabolic engineering framework. Here, I provide a summary of the most common methods for in silico strain design (primarily but not exclusively using FBA) (Table 1.1), followed by a suggested approach for selecting the appropriate method for a given task.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>OptKnock</td>
<td>The bilevel optimization structure of OptKnock includes an inner component that optimizes a function expected to be selected in natural evolution. Moreover, the outer component aims to find the gene knock-out targets for the production of the chemical of interest.</td>
<td>[72]</td>
</tr>
<tr>
<td>Method</td>
<td>Description</td>
<td>Reference</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td><strong>OptStrain</strong></td>
<td>Identifies plausible reaction additions using a comprehensive database of biotransformations that is compiled from different bio-pathway databases. After the necessary non-native functionalities are identified and added from the database, knock-out strategies are designed in order to remove competing reactions.</td>
<td>[73]</td>
</tr>
<tr>
<td><strong>OptReg</strong></td>
<td>Considers a reaction to be repressed if the reaction flux in the engineered strain is higher than the steady-state flux. The opposite applies when a reaction is considered to be activated. One reaction can have one type of three states: knocked-out, up-regulated, and down-regulated.</td>
<td>[74]</td>
</tr>
<tr>
<td><strong>OptGene</strong></td>
<td>Predicts knock-out strategies using probability flux-balance analysis and MOMA. The identification of the optimal set of knock-outs is made through the use of evolutionary algorithms; primarily genetic algorithm and simulated annealing.</td>
<td>[75]</td>
</tr>
<tr>
<td><strong>OptFlux</strong></td>
<td>Optimizes the objective yield while maintaining a minimum level of biomass using evolutionary algorithms with the options to use FBA, MOMA and ROOM.</td>
<td>[76]</td>
</tr>
<tr>
<td><strong>RobustKnock</strong></td>
<td>Aims at eliminating competing pathways using a max-min approach to solve the dual optimization problem rather than the max-max approach in OptKnock. By doing so, the number of FBA solutions is reduced to only those that guarantee a minimum production rate of the chemical of interest.</td>
<td>[77]</td>
</tr>
<tr>
<td><strong>OptForce</strong></td>
<td>Uses differences in flux values between the wild-type strain and the engineered strain to classify all reactions in the network to three main categories: reaction fluxes should be increased, decreased and eliminated. Thus, OptForce differs from other methods that rely on the bi-optimization of both the biomass and the target chemical as it reports the most conservative set of targets given the stoichiometric boundaries.</td>
<td>[78]</td>
</tr>
<tr>
<td><strong>EMILIO</strong></td>
<td>Enhancing Metabolism with Iterative Linear Optimization attempts to solve the scalability problem of the other tools especially with large genome-scale models primarily using Successive Linear Programming (SLP). It optimizes the growth-coupled problem by first identifying reactions that serve as possible targets and then predicting the flux range for each reaction in the proposed optimal engineered strain.</td>
<td>[79]</td>
</tr>
<tr>
<td><strong>GDBB</strong></td>
<td>Genetic Design through Branch and Bound is a fast method that uses 'truncated branch and bound algorithm' to find modification targets. The heuristic nature of the method makes it faster than other in silico strain design ones.</td>
<td>[80]</td>
</tr>
<tr>
<td><strong>DySScO</strong></td>
<td>Dynamic Strain Scanning Optimization uses dynamic flux-balance analysis (dFBA) as its modeling approach by incorporating process optimization in addition to yield optimization. Thus, It offers a more industrial design by incorporating factors such as product yield, titer, and productivity volume in the optimization function. By using dFBA, it incorporates a more detailed simulation of the microbial cell factory by considering the substrate uptake rate in both batch and fed-batch bioreactors.</td>
<td>[81]</td>
</tr>
</tbody>
</table>

Based on the summary provided in the table, it is evident that the selection of the most suitable approach depends on the nature of the genome-scale model and the aim of the engineering rationale. For instance, OptStrain is most fitting for designing strains that do not contain genes necessary for the synthesis of the product of interest. Additionally, if the analyzed network is large in size and leads to computationally expensive simulations, utilizing a heuristic-based approach, such as GDBB or EMILIO, this would solve the scalability problem. Moreover, if the application is industrially motivated, where maximizing the titer is the main focus, process optimization methods (such as DySScO) will be the most suitable. Finally, for investigations that require a holistic evaluation, where each reaction is inspected as an engineering target, OptForce is recommended as it provides a thorough evaluation.
of the network.

1.1.2 Genome Mining as a Screening Method for Biosynthetically Plausible Microbial Cells

The first antimicrobial secondary metabolite (penicillin) was discovered in the lab of Alexander Fleming (produced by *Penicillium notatum*) in a Petri dish in 1929. It was an accidental discovery when Fleming noted how a fungal colony on an agar plate caused the cell walls of the colonies near the mold to lyse as they were evidently transparent [82, 83]. From these fungal metabolites, the idea of using toxic microorganisms for the production of antibacterial compounds was ignited. These compounds assisted in critical medicinal applications, aimed at fighting infections and controlling the spread of pathogens in mammalian systems. Further experiments were performed to detect microorganisms that show an inhibition zone against both Gram-positive and Gram-negative pathogens. In 1943, there was an early success in the discovery of antibiotics from bacteria with the discovery of streptomycin produced from *Streptomyces griseus* [84]. Since then, *Actinomycetes* have been the biggest contributors in the industrial production of antibiotics used in clinical applications, as they are usually not toxic to human cells [85, 86].

The extensive focus on *Actinomycetes* overshadowed other taxonomic groups that might be enriched with novel antibiotics. Fortunately, genome sequencing and functional genome annotation showed that antibiotics and other secondary metabolites are usually encoded by clusters of co-localized genes referred to as Biosynthetic Gene Clusters (BGCs) [87, 88]. Accordingly, the detection of BGCs in
sequenced genomes facilitates the computational prediction of biosynthetic genes and subsequent evaluation of the richness of different classes of BGCs in different taxa [89].

Nonribosomal Peptide Synthetases (NRPSs) and Polyketide Synthases (PKSs) are two classes of BGCs that are often associated with the synthesis of antitumor, antimicrobial, antifungal, and immunosuppressive products [90, 91]. NRPS and PKS BGCs are known to be modular in their structure, consisting of megasynthases (with a number of enzymatic domains) that catalyze the biosynthesis of nonribosomal peptides and polyketides respectively [92]. Core domains in NRPS synthetases (adenylation, thiolation and condensation domains) catalyze the ‘sequential condensation’ of amino acids as building blocks [93], while core domains in polyketides synthases (ketosynthase, acyltransferase and acyl carrier protein) catalyze decarboxylative condensation of short carboxylic acids (acetate) as building blocks, resulting in the repetitive addition of two-carbon ketide units to the growing chain [92, 94]. There are also additional non-core domains, in tailoring enzymes, that catalyze the modifications of the peptide and ketide backbones in NRPS and PKS clusters respectively [95]. Usually, the structure of the end product can be inferred by the number, specificity, and order of the enzymatic domains (collinearity rule) [96].

Another type of antimicrobial BGCs is ribosomally synthesized and posttranslationally modified peptides (RiPPs) [97] including modified and unmodified bacteriocins. RiPPs biosynthesis starts with a structural gene that encodes a ribosomal-based peptide [97]. This leader peptide is hypothesized to be a recognition signal for post-translational modification enzymes, a secretion signal, and a chaperone
for the folding and stabilization of the core peptide [98]. In some other cases, there are C-terminal recognition sequences used for the cyclization and excision of the main peptide. In modified bacteriocins, the core peptide undergoes post-translational modifications, allowing diverse structures to form, as well as more chemical functionalities, and increased stability [97]. For instance, lanthipeptides (one type of post-translationally modified bacteriocins) are identified by the having lanthionine (Lan) and methyl-lanthionine (MeLan) as signature amino acids [99]. Lanthipeptides have been extensively reported to show antimicrobial activities, and thus the term lantibiotics was given to such lanthipeptides [99].

Fast advancements in next-generation sequencing technologies, along with the development of computational genome-annotation tools have facilitated the mining of sequenced genomes for genes with specific biosynthetic function and gene clusters that account for biosynthetic capabilities [100-105] [106]. Thus, genome mining can be considered a suitable approach for the evaluation of microbial cells as ‘natural product factories’. Some tools for the prediction of BGCs in sequenced genomes include Antibiotics & Secondary Metabolite Analysis Shell (antiSMASH) which utilizes profile Hidden Markov Models (pHMMs) of signature genes [107]; BAGEL, which mines for modified and unmodified bacteriocins [108]; and ClusterFinder [109], which uses prediction models to predict clusters of unknown types.

1.2 *Bacillus*, A Biotechnologically Rich Genus

Historically, species from the *Bacillus* genus were one of the first described Bacteria in 1835, initially referred to as *vibrio subtilis* [110]. Later, in 1872, it was
renamed as *B. subtilis* and microscopically described as a rod-shaped bacteria [111]. The genus is most notably known for its distinctive endospore-forming ability; allowing its species to withstand unfavorable environmental conditions and physiochemical stress [112, 113].

Strains in this Gram-positive genus are ubiquitously present in nature as they can survive in a variety of environments with different types of organic and inorganic nutritional resources [114]. Accordingly, they show different phenotypes with versatile production and biosynthesis capabilities for antimicrobials, toxins and other industrially relevant compounds. Some of the previously reported environments from which *Bacillus* strains were isolated include, but are not limited to, marine environments (seawater [115], tidal flat [115-117], and sediments [118-123]); soil environments (rhizospheres) [124-128]; human gut samples [129-131]; and food samples (dairy products [132, 133] and fermented soybean [132]).

Interestingly, a study by Alcaraz L.D. *et al.* [134], that aimed at identifying evolutionary and functional relationships between twenty complete and draft Bacilli genomes isolated from different environments, found that most of the metabolic variation stemmed from genes with functions necessary for environmental adaptation. These findings highlight the ecologically-specific genetic richness of Bacilli genomes and warrant detailed investigation of the metabolic capabilities of different Bacilli dwelling in different niches. The potential of Bacilli for the production of antibiotics (mostly peptide antibiotics) has been established for more than 50 years [135]. There are several studies that have provided excellent reviews on structures, biochemistry, and biosynthesis of antimicrobials in Bacilli species [135-138]. A recent work, by Zhao and Kuipers (2016), has additionally used computational genome
mining methods to mine a total of 328 strains in 57 different Bacilli genomes for antimicrobial secondary metabolites [139]. The study found that the genomes of some Bacilli species - previously not known as antimicrobial producers - have BGCs for the synthesis of antimicrobial compounds. This finding stresses the advantage of using computational genome-wide screening over just relying on phenotypic observations to evaluate the species as a potential producer. The fact that some genes encoding bioactive compounds are only expressed under specific (unknown) conditions, results in an inaccurate evaluation of the actual biosynthetic potential of the strain.

A number of known Bacilli strains have shown strong biosynthetic capabilities, both in silico, using computational methods, [139, 140] and in vitro, where the secretion of different proteins was detected [141-148]. Accordingly, distinguished Bacilli strains are used for industrial production of an array of pharmacologically and industrially relevant compounds including biosurfactants [149, 150], antimicrobials [151-153], hydrolysis/deproteinizing enzymes [154-156], and livestock probiotics [157]. The extensive study of the metabolism of B. subtilis 168 facilitated the use of metabolic and genetic modification methods to transform these strains into overproducing MCFs for products of interest [158]. Some Bacilli strains that have natural competence (allowing easy genetic manipulation), and efficient protein secretion (allowing easy recovery of products without having to lyse the cells), are ranked as better MCF candidates [159-163].

Thermophilic and halophilic Bacillus strains [34, 154, 164-171] naturally have more ability to ferment in different thermal and pH ranges [162, 172] which add to their industrial value. B. subtilis, B. licheniformis and B. amyloliquefaciens have been optimized for their use as MCFs with industrial relevance. Some of the favorable
characteristics of these species are survival in fermentation media, high product yield, and low number of toxic by-products [173]. Moreover, the detailed understanding of *B. subtilis* metabolism and sporulation, its high transformation rate (i.e. ability to take exogenous naked DNA) and its strong secretion system (through the *Sec* and *Tat* pathways), add to its value as an industrial producer [173]. Additionally, the availability of genomic modification tools make strains of this species good candidates as chassis to overproduce an array of biotechnologically important compounds [174].

Moreover, metabolic engineering strategies have been generated and utilized to overproduce several targets of interest in *Bacillus* hosts. For instance, although the model strain for amino acids production (*Corynebacterium glutamicum*) is not within the *Bacillus* genus, there have been some attempts for the overproduction of amino acids as precursors for secondary metabolites (peptides) in *B. subtilis* 168. In one specific work [175], an *in silico* approach was used to generate a metabolic engineering strategy for the targeted overproduction of the amino acid precursors for surfactants. Specifically, six gene knock-outs in *B. subtilis* 168 resulted in a decrease of leucine degradation and an increase in a 1.6 to 20.9-fold in the production of surfactants.

Additionally, the optimization of d-lactic acid, a basic component of a degradable and environmentally friendly bioplastic material, was completed in *B. subtilis* [176]. Since the strain does not have the necessary genes for d-lactic acid synthesis, five genes, related to d-lactate dehydrogenase from *B. coagulans* and *Lactobacillus delbrueckii*, were introduced in a *B. subtilis* 168 strain using cloned genes in *E. coli* plasmids (pDA9, pDA34, pDA38, pDA42). The introduction and expression of
d-lactate dehydrogenase genes resulted in a total titer of 0.6 M at a yield of 0.99 g/g of glucose.

Other strategies required deleting genes that negatively affect the production of the end product. For instance, to overproduce meso-2,3-butanediol, a promising biofuel [177], competing pathways were eliminated by knocking out genes which encode d-2,3-butanediol dehydrogenase as it consumes the precursor of meso-2,3-butanediol (acetoin). The strategy also inactivated phosphotransacetylase (pta) and lactate dehydrogenase (ldh) to decrease the concentrations of the byproducts of the introduced reaction (acetate and l-lactate). These knock-outs were followed by the introduction of meso-2,3-butanediol dehydrogenase genes. One final intervention was the expression of the gene which encodes a soluble transhydrogenase (udhA) to increase the available NADH pool. These three steps led to 0.487 grams of meso-2,3-butanediol per gram of glucose and decreased the buildup of acetoin to 1.1 g/L.

Another strategy that used NADH levels to increase an NADH-dependent product is that for 2,3-butanediol [178]. Inactivation of enzymes which catalyze reactions for the consumption of NADH led to higher 2,3-butanediol titer as more NADH becomes available. In the first attempt, genes which encode NADH oxidases (yodC) were knocked out to reduce NADH oxidization, but this did not result in sufficiently higher yields. Accordingly, the strategy was modified by the introduction of formate dehydrogenase, which increased the NADH pool and resulted in overproduction of 2,3-butanediol. The approach also decreased the accumulation levels of acetoin, but did not prevent the accumulation of lactate, which had a negative regulatory effect on 2,3-butanediol. Thus, the final intervention was knocking
out lactate dehydrogenase resulting in a 19.9 % increase in 2,3-butanediol with minimal lactate and acetoin accumulation.

In a study by Meng, W. et al., 2015 [179], 2,3-butanediol was considered as a competing pathway to the production of another chemical, tetramethylpyrazine and its precursor (acetoin). In this strategy, tetramethylpyrazine production was increased by blocking 2,3-butanediol synthesis which increased the yield of tetramethylpyrazine from 21.8 to 27.8 g/l, and also increased the yield of its precursor from 11.3 g/l to 16.4 g/l. In one additional strategy [180], a mutant *B. subtilis* 168 strain, genes which encode acetoin dehydrogenase (*acoA*) and 2,3-butanediol dehydrogenase (*bdhA*) were knocked out, and the operon for acetoin formation (*alsSD*) was overexpressed, resulting in an increase in acetoin synthesis of 52%.

Some other examples in which *B. subtilis* strains were used as optimized MCFs include the overproduction of α-amylase up to 2500-folds [181], the overproduction of the prohormone precursor to insulin (proinsulin) up to 1 mg ml⁻¹ [182], and the overproduction of streptavidin up to 35 to 50 mg/liter [183].

Ever since the first *B. subtilis* genome in 1997 [184], detailed analysis of the essentiality of genes encouraged the use of synthetic biology to develop strains with minimized genomes, referred to as **Minimal Genome Factories (MGFs)**. In metabolic engineering, genome minimization efforts aim to simplify the regulatory and metabolic network of a strain in order for it to be more amenable to genetic modification and to result in a predictable behavior where desired phenotypes are easily reached.
One of the first attempts to minimize the genome of *B. subtilis* 168 was completed by Westers *et al.* in 2003 [163]. 7.7% of the genome, represented in two prophage regions, three prophage-like regions and one PKS cluster, was deleted. The resulting mutant strain grew but did not have overproduction capabilities, compared to the wildtype parent strain.

It was not until 2008 that genome minimization of *B. subtilis* 168 showed how a minimized genome has more production capabilities compared to the wildtype strain. Thus, this work can be considered the first successful attempt at using synthetic biology for efficient MGFs [161]. However, reaching this overproducing minimized strain was not a straightforward task. First, all regions that are known not to participate in essential functions, related to primary metabolites, DNA, and RNA functions, were identified to be excluded from the list of regions that can be deleted. The minimized strains had around 2 Mb of the genome eliminated, but did not yield any growing functional strains in all tested media. Correspondingly, the authors decided to change the deletion strategy and opted for a more rational stepwise approach. In this approach, each region that could be deleted was tested and evaluated for growth in a stepwise fashion, reducing the size of DNA regions that can be deleted from 2 Mb to 874 kb, and resulted in strains with no morphological or chromosomal abnormalities, albeit with a relatively lower growth rate. Finally, when plasmids with genes necessary for the production of alkaline protease and cellulose were introduced, it was noted that the production capability of the tested strain positively correlated with the length of the deleted region, proving the efficiency of MGFs.
One of the most recent works on genome minimization of *B. subtilis* 168 aimed to use the well-validated metabolic model for *B. subtilis* 168 (iBsu1103), to test the accuracy of predicted essential DNA regions (157 in total) in different media compositions (one complex media and one defined media) [185]. Of the 157 attempted deletions, 146 deletions were not critical, in at least one of the tested media types, while the remaining deletions were. In cases where predicted phenotype upon deletion is different from experimental results, iBsu1103 was iteratively refined and corrected, resulting in a new version of the model iBsu1103V2. The availability of the set of validated essential and coessential genes, along with phenotypical information resulting from the exclusion of each gene in the refined model, provide an excellent recourse for *B. subtilis* genome minimization.

### 1.3 Thesis Outline

In Chapter 1, all of the background and introductory information related to the study are introduced, primarily focusing on evaluation methods of potential MCFs, as well as the richness of the *Bacillus* genus with efficient microbial strains. In Chapter 2, comparative genomics and evolutionary analysis are used to highlight the metabolic distinction of the Red Sea strains. In Chapter 3, species that are identified as optimally performing from the study reported in Chapter 2, are further evaluated as biotechnological platforms and compared to strains in close species. In Chapter 4, the genome of the strain with the most unique metabolic and bioactive profile, identified from the study in Chapter 3, is analyzed and evaluated using constrained-based modeling. The pruning process used to select the most optimal strain is depicted in Figure 1.1. Finally, in Chapter 5, a general summary of the significant conclusions made
from different parts of the study are provided, accompanied by recommendations worthy of future investigation.

**Figure 1.1.** Project outline describing the pruning approach of the ten Red Sea isolates and showing the evaluation methods at each stage.
Chapter 2

2 A Comparative Genomics Study Reveals the Environment-dependent Potential of the Red Sea Bacilli for Production of Secondary Metabolites

One way of identifying different functional patterns between taxonomically close strains is through the utilization of reconstructed metabolic networks. This has been facilitated by the currently available large databases of biochemical pathways, which provide a comprehensive view of the metabolic functions that are variant between organisms of interest [186]. For instance, functional genome-scale metabolic models and FBA were successfully utilized to identify metabolic functions related to the virulent phenotype of Staphylococcus aureus strains in around 300 different environments [187].

A recent survey of biosynthetic gene clusters in Bacilli genomes found that some types of BGCs (lipopeptides and polyketides) are found to be constrained in certain Bacillus species [140], a confirmation that different species from different environments have unique biosynthetic capabilities. For example, mining the genomes of the rhizosphere-dwelling B. amyloliquefaciens, has helped in identifying antimicrobial and antitumor agents produced by this bacterium [187-194]. Additionally, analyzing the phenotypic and genomic properties of B. thuringiensis, which is believed to be omnipresent in marine environments, resulted in the identification of cry genes and zwittermycin [141, 145, 195-201] that encode for proteins with toxic phenotypes.
The Red Sea is one of the recently explored environments, that was found to be populated with Bacilli strains [202]. The ecological uniqueness of this environment, regarding its salinity and temperature, marks its microbial genomic repertoire as putatively attractive for the discovery of unique metabolic and biosynthetic capabilities. Here, overrepresented and underrepresented metabolic reactions in a dataset containing 22 reference Bacilli genomes and the ten Red Sea strains were compared. Additionally, secondary metabolic gene clusters were predicted and cataloged to analyze their homology and co-localization patterns. Finally, to further assess the biotechnological potential of these strains, their capacity for protein secretion, competence, and sporulation were ranked against levels detected in the reference strain *B. subtilis* 168. The results suggest that specific modules of secondary metabolism have evolved in the Red Sea Bacilli due to environmental adaptation, and several of the isolated strain represent promising platforms for the development of cell factories.

### 2.1 Material and Methods

#### 2.1.1 DNA Extraction and Sequencing

Biomass of the ten strains was obtained after growth under optimal conditions (as described in Al-Amoudi et al., 2016) [203]). Genomic DNA was extracted using the Sigma Gen Elute Bacterial Genomic DNA Kit (Sigma, USA) as per the manufacturer’s protocol followed by a second purification step utilizing MO BIO PowerClean Pro Clean-Up Kit (USA). As a quality control, DNA purification was assessed by overnight gel electrophoresis and NanoDrop (Thermo Fisher Scientific, USA), while Qubit 2.0 (Life Technologies, Germany) was used to quantify the DNA. A Single-Molecule Real-
Time (SMRT) cell was run on the PacBio RSII platform (Pacific Biosciences, USA) in the King Abdullah University of Science and Technology (KAUST) BioCore Lab (Jeddah, Saudi Arabia) with P6-C4 chemistry.

2.1.2 Genome Assembly

Raw data from PacBio’s RS II were assembled using PacBio’s SMRT Analysis pipeline v2.3.0. using default parameters and the genomeSize parameter set to 6,000,000 bp, which produced a single contig per library. Overlapping ends were checked using Gepard v1.40 [204] which would indicate circular genomes. To circularize both genomes, one end of each contig was trimmed to reduce the amount of overlap, then each contig was split into two halves which were then rejoined using minimus2 [205]. After circularization, multiple rounds of assembly polishing were performed using the SMRT Analysis Resequecing protocol until convergence. To assess the quality of the genomes and estimate their completeness and contamination, checkM v1.0.6 [206] taxonomic workflow was used, utilizing single-copy genes in the genus Bacillus.

2.1.3 Genome Functional Annotation and Analysis

The complete genome sequences of the ten strains were annotated using the Automatic Annotation of Microbial Genomes pipeline (AAMG) [207] with default parameters (BLAST bit score of 30) and Prodigal [208] as the chosen gene predictor. For details about the annotation pipeline, tools and databases used, refer to [207].

2.1.4 Bacterial Strains and Media

All strains were cultured in Luria-bertani (LB) broth at 37°C in shaking flask. The minimal media M9 [209] was used for secretion essays and the Difco Sporulation
Medium (DSM) [210] for sporulation essays. When relevant, kanamycin (25 µg/ml) and erythromycin (1 µg/ml) were added for the Bacilli strains genetic transformation.

2.1.5 Competence Assay

The competence of Bacilli strains was tested by the transformation of the electrocompetent cell prepared from the tested strains, using two replicative plasmids: pMAD for erythromycin resistance [211] and pJC15 for kanamycin [212]. The electrocompetent cells from the Bacilli strains were prepared according to the method described by Xue et al. [213]. The strains were cultured overnight in LB culture. The culture was diluted 16-fold into 50 ml of LB growth medium with 0.5 M sorbitol, and grown at 37°C to an optical density (OD 600) of 0.85-0.95. The cells were cooled on ice at 4°C for ten minutes and harvested by centrifugation at 4°C and 5000X g for ten minutes. This was followed by four successive washing steps in cold electroporation medium with 0.5 M sorbitol, 0.5 M mannitol, and 10 % v/v glycerol. Then, cells were suspended in electroporation medium (1/40 (v/v) ) and divided into 50µl aliquots. 50 µl of the competent cells was mixed with 100 ng pMAD or pJC15 vectors and then transferred to a cold electroporation cuvette with 1-mm electrode gap. Following five minute incubation on ice, the mixture of cells-DNA was shocked with a MicroPulser Electroporator by a single pulse of 21 Kv, with a time constant (τ) of 5 ms. The cells were immediately added into 1 ml of outgrowth LB medium with 0.5 M sorbitol and 0.38 M mannitol, and incubated in shaking media at 37 °C for four hours to allow expression of the antibiotic-resistant genes. The cultures were then spread onto agar plate with the kanamycin antibiotics (25 µg/ml) and erythromycin (1 µg\ml) for pJC15 transformation and pMAD transformation respectively. Counted colonies on the
plates after 24h of incubation at 37°C (transformants /μg DNA using 100ng of vector) was considered as the measuring parameter of transformation efficiency.

2.1.6 Sporulation Assay

The overnight LB cultures of the Bacilli strains at 37°C were diluted to OD600 0.1 in preheated liquid DSM medium [210] and cultured under shaking conditions (200 rpm) at 37°C to OD600 of two (around 24 hours of culture). 500μl of the culture was heated at 80°C for 10 minutes. Serial dilutions from the heated samples were made in 0.9% NaCl (from $10^{-1}$ to $10^{-7}$) and 100μl of the material from dilution with $10^{-3}$ to$10^{-7}$ plated in LB and incubated for 36 h at 37°C. After incubation, the colonies were counted on plates and considered as the number corresponding to heat-resistant spores. Three biological replicates were performed, providing similar results. One representative biological replicate for each Bacilli strain is reported with standard deviation from three technical replicates.

2.1.7 Protein Secretion Assay

The tested Bacilli strains were grown for overnight culture in LB at 37°C and 200 rpm shaking speed. 20 mls of minimal media (M9) that was supplemented with 0.4 % glucose [209] were inoculated to OD600 of 0.1. At the late exponential phase of growth, 3 Mm of the protease inhibitor Phenylmethylsulfonyl fluoride (PMSF) was added to the cultures to prevent proteolytic digestion and then the cells were grown for 24 hours to OD600 of 2. The cultures were centrifuged at (8000xg) for 10 minutes and the supernatants were filtered with 0.2 nm nitrocellulose to eliminate the rest of cells. To precipitate the secreted proteins, 50% (w/v) of ammonium sulphate was added to the supernatants and kept on ice for 1 hour. The mixtures were then
centrifuged at 25000xg for 20 minutes. The precipitated proteins were resuspended in PBS buffers at pH 6.8. The proteins concentration was measured using NanoDrop, and compared with samples of M9 media (Table S2.1) without bacteria treated with the same procedure for culture and protein precipitation. One representative biological replicate is reported from three technical replicates.

2.1.8 Phylogeny Analysis

The phylogeny analysis was performed by identifying the single-copy genes shared amongst the 40 strains included in the tree, yielding a total of 188 single-copy genes. Those genes were aligned using MUSCLE v3.8.31 [214], and the alignments were concatenated using FASconCAT-G v1.02 [215]. ProtTest3 v.3.4.2 [216] was then used to predict the amino-acid replacement model. Finally a maximum likelihood tree was built using PhyML v3.1 [217] with 100 as bootstrap value using the LG + I + G (LG) model as recommended by ProtTest3.

2.1.9 Identification of Competence, Protein Secretion and Sporulation Genes

To identify competence and protein genes and sporulation genes, homologous genes to those present in B. subtilis 168 were identified using bidirectional blastp. Genes related to sporulation, protein section and competence functions were retrieved from the SubtiList database available at [http://genolist.pasteur.fr/SubtiList/] [218, 219]. A gene was considered to be present in the genome if it has a bidirectional hit with its homologue in B. subtilis. The minimum e-value was set to be 1e-5, all top hits had e-value of 1e-20 or less. If a gene is absent in a genome, it was given a gene coverage value of 0. Ranking the genomes was performed using the average coverage of all the genes in each genome as well as considering the average gene copy number. To search
for intergenic regions that are conserved in \textit{B. subtilis} and the 10 local strains, GET_HOMOLOGUES [220] was used to calculate orthologous ORF clusters and for conserved intergenic regions, with minimum intergenic region length of 200 nts, maximum size of 700 nts and 180 nts were allowed to be borrowed from neighboring ORFs.

\subsection*{2.1.10 Metabolic Reconstruction}

The metabolic networks were reconstructed using MetaCyc database [221]. For each reaction of this database, the sequence of the enzymes catalyzing the reaction has been retrieved. When only one enzyme sequence was available for a given reaction, a blast was run between this sequence and the predicted proteome of a given species. The chosen threshold is a bit-score of 100. When several enzyme sequences were available, an HMM model characterizing those sequences was reconstructed using HMMer [222]. This model was searched against the predicted proteome. The chosen threshold is again a bit-score of 100.

To find over and under-represented reactions per metabolic clade in the heat map, a binomial test was run for each reaction, comparing the number of positive occurrence between the considered clade and the other species. A threshold of 1e-5 was chosen to consider a reaction as overrepresented or underrepresented in a given clade.

\subsection*{2.1.11 Biosynthetic Gene Clusters Prediction}

BGCs were predicted on the public web version of antiSMASH 4.0 (http://antismash.secondarymetabolites.org) [223]. The ClusterFinder algorithm was used to predict the borders of BGCs. Manual curation was performed to define more
precisely the boundaries of the gene clusters based on similarities with known gene clusters present in the MIBiG database. Some of the predicted gene clusters have been split into two different gene clusters when sufficient information was available to make this decision. The secondary metabolite product of a given gene cluster was identified if this gene cluster showed at least 60% similarity with one characterized gene cluster present in the MIBiG database. MultiGeneBlast [224] was used to infer similarities between all predicted gene clusters in the \textit{Bacillus} species. Two BGCs were considered to be similar if they show at least 60% similarity with each other. To identify the co-localization of gene clusters, a normalized position was computed for each BGC based on the length of the chromosome and the position of the middle of the BGC. This was only completed for the species possessing a complete chromosomal sequence.

\section*{2.2 Results and Discussion}

\subsection*{2.2.1 Bacilli Present in the Red Sea Originate from Several Colonization Events}

Sequencing the genomes of the Red Sea strains using the SMRT (single molecule real-time) sequencing platform produced on average 128,522 subreads with a mean length of 9,527 bp and an average coverage of 272 (lowest 112 and highest 344). The assembly produced a single circular chromosome without plasmids for all of the genomes, except for \textit{B. foraminis} Bac44, \textit{Virgibacillus} species Bac324, Bac332 and \textit{B. amyloliquefaciens} Bac5, with an average genome size of 4.536 Kb, 4,483 predicted open reading frames (ORFs), and 29.6 rRNAs.

To assess the phylogenetic positions of the 10 newly sequenced Red Sea species, whole-genome phylogenetic analysis was completed using 10 Red Sea species,
together with the 29 reference species of Bacilli and one outgroup. The obtained tree is available as Figure 2.1.

**Figure 2.1.** Phylogenetic tree of the 10 Red Sea Bacilli and 30 other species. The Red Sea species are displayed in grey while previously sequenced Bacilli are displayed in black.

This phylogenetic tree confirmed that the Red Sea species are located among Bacilli in this phylogenetic tree. The analysis showed that these 10 Bacilli species, present in the Red Sea, come from several colonization of this environment and are not the result of a single colonization followed by speciation. Therefore, similarities in the overall metabolic setup among the Red Sea species discussed here are due to convergent evolution.
Phylogenetic analysis also showed that some of the species are closely related to other known Bacilli species, such as *B. vallismortis* Bac111 that is closely related to *B. subtilis* 168, or *B. paralicheniformis* Bac84 and *B. paralicheniformis* Bac48 that are closely related to *B. licheniformis* ATCC 14580. Interestingly, *B. subtilis* 168 and *B. licheniformis* ATCC 14580 both have a soil habitat while the newly sequenced species live in the sea environment. Given the close relationship between their genomes, it is probable that some small adjustment of the metabolism of the newly sequenced species had occurred, enabling them to colonize this specific environment.

### 2.2.2 Several Bacilli Isolated from the Red Sea Present Substantial Potential for Protein Secretion

Bacilli are extensively used for production of industrial enzymes. In order to qualify for such applications, a candidate strain would ideally be non-sporulating and have an efficient protein secretion machinery. The Red Sea strains were tested along these lines, comparing them to the standard laboratory strain *B. subtilis* 168. Three of the species did not grow on standard laboratory media, so the results shown in Fig. 2.2, 2.3, 2.4 concern the seven Red Sea species that were able to grow in laboratory conditions. Sporulation is a complex developmental process, involving several hundred genes [134, 225]. The sporulation capabilities of the different species were consistent with their genetic capabilities, i.e. the presence of sporulation genes. There is a good correlation between the number of genes involved in sporulation mechanisms and the number of spores produced by the different species (Figure 2.2). Interestingly, strains *Virgibacillus* sp. Bac332 and *Virgibacillus* sp. Bac330 were completely incapable of sporulating. This can be related to the fact that in both
species, despite the presence of many sporulation genes, the gene spo0B is missing. This gene is a key element in the phosphorelay regulating sporulation initiation [226]. Its absence has already been reported to lead to completely arrested sporulation in phase 0 [226]. No clear explanation has been found concerning the quasi-absence of sporulation for *B. paralicheniformis* Bac84. Further studies will be needed to explain this phenomenon. Globally, it is interesting to note that none of the Red Sea species sporulate more than *B. subtilis 168* under the studied conditions.

Table 2.1. Average and standard deviation of three measurements of number of heat-resistant spores (10⁴)

<table>
<thead>
<tr>
<th>Species</th>
<th>Average</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. paralicheniformis</em> Bac48</td>
<td>421.67</td>
<td>27.43</td>
</tr>
<tr>
<td><em>B. amyloliquefaciens</em> Bac57</td>
<td>467.33</td>
<td>50.85</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em> Bac-84</td>
<td>3.33</td>
<td>2.08</td>
</tr>
<tr>
<td><em>B. litoralis</em> Bac94</td>
<td>153</td>
<td>29.21</td>
</tr>
<tr>
<td><em>B. vallismortis</em> Bac111</td>
<td>492.66</td>
<td>23.69</td>
</tr>
<tr>
<td>Virgibacillus sp. Bac330</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Virgibacillus sp. Bac332</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>B. subtilis</em> 168</td>
<td>493.33</td>
<td>24.13</td>
</tr>
</tbody>
</table>
Protein secretion in *Bacillus* species is mainly controlled by two systems: the *Tat* and *Sec* systems, that transport respectively folded and unfolded proteins across the membrane. In *B. subtilis*, both systems involve only a small number of genes: *tatAd*, *tatAy*, *tatAc*, *tatCd* and *tatCy*; and *secA*, *secY*, *secE*, *secG* and *secDF*. While comparing the genes present in the Red Sea species and the actual protein secretion (Figure 2.3) show no clear correlation between the presence of genes of the *Tat* and *Sec* systems, and the actual protein secretion capabilities. It is nevertheless worth noting that two of the species from the Red Sea secrete twice as much protein as *B. subtilis* 168 when grown in LB medium. One of these species, *B. litoralis* Bac94 possesses complete *Tat* and *Sec* systems. This indicates that *B. litoralis* Bac94 and Bac48 could be interesting candidates for the development of cell factories for enzyme production and for the study of the protein secretion pathways involved in such secretion.

![Figure 2.3. Evaluation of protein secretion in the Red Sea strains through gene prediction for the Tat, Sec and Sip secretion systems and in vitro measurements.](image)

**Table 2.2.** Average and standard deviation of three measurements of secreted protein for each Red Sea species

<table>
<thead>
<tr>
<th>Species</th>
<th>Measurement 1</th>
<th>Measurement 2</th>
<th>Measurement 3</th>
<th>Average</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bac-94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac-230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac-232</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac-57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-118</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac-111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac-48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac-94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Additionally, four Red Sea Species (Bac48, *Virgibacillus* sp. Bac330, *B. vallismortis* Bac111, *B. amyloliquefaciens* Bac57) showed capabilities for genetic transformation, however not as much as *B. subtilis* 168 (Fig 2.4).

<table>
<thead>
<tr>
<th>Species</th>
<th>Transformants/µg DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pMAD</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em></td>
<td></td>
</tr>
<tr>
<td>Bac48</td>
<td>3.36</td>
</tr>
<tr>
<td></td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td><em>B. amyloliquefaciens</em></td>
<td>0.98</td>
</tr>
<tr>
<td>Bac57</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>0.18</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em></td>
<td>0.19</td>
</tr>
<tr>
<td>Bac84</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
</tr>
<tr>
<td><em>B. litoralis</em> Bac94</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td><em>B. vallismortis</em> Bac111</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td><em>Virgibacillus sp.</em> Bac330</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
</tr>
<tr>
<td><em>Virgibacillus sp.</em> Bac332</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td><em>B. subtilis</em> 168</td>
<td>1.91</td>
</tr>
<tr>
<td></td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 2.3. Measurements for transformation efficiencies for Red Sea strains in number of transformants per microgram of DNA

Figure 2.4. transformation efficiencies of Red Sea Bacilli with pMAD and pJC15
B. amyloliquefaciens Bac57  
5.71E+00  1.43E+01

B. paralicheniformis Bac84  
0.00E+00  0.00E+00

B. litoralis Bac94  
0.00E+00  0.00E+00

B. vallismortis Bac111  
1.43E+01  7.14E+00

Virgibacillus sp. Bac330  
8.57E+01  8.57E+00

Virgibacillus sp. Bac332  
0.00E+00  0.00E+00

B. subtilis 168  
1.29E+02  1.10E+02

2.2.3 Convergent Evolution of Metabolic Networks in Bacilli

To have better input into the metabolism of the 32 studied species, and to discover eventual metabolic features specific to the ten Red Sea species, their metabolic networks were reconstructed. Those metabolic networks ranged from 671 to 1,398 reactions, with an average of 1,238 reactions (median: 1,291 reactions) and from 1,050 to 1,897 different metabolites involved in those reactions, with an average of 1,667 metabolites (median: 1,715 metabolites). It is interesting to note that the three networks possessing the most metabolic reactions correspond to Red Sea species. Globally, the Red Sea species possess on average 57 more reactions than the other species (1,277 vs. 1,220 reactions), and on average 75 more metabolites (1,719 vs. 1,644 metabolites). The reactions are inferred based on metabolic genes. Between 477 and 1,042 genes per species, 903 on average (median: 941) were used.

Those numbers could be compared to the most complete existing metabolic model of B. subtilis 168, iBsu1103 [227]. This metabolic model contains 1,437 reactions accounting for 1,103 metabolic genes. In comparison our reconstruction of the metabolic network of B. subtilis 168 contains 1,366 reactions associated with 1,006 genes. The numbers are similar and the small difference could be attributed to the absence of gap-filling and the fact that less manual curation was carried on in our case.
To have better understanding of the metabolism of Bacilli and to search for signatures in their metabolism, the reactions were compared for their presence in different metabolic networks. A clustering, represented in Figure 2.5, has been made based on the presence/absence of reactions. Based on this clustering, several ‘metabolic clades’ appeared clearly. The 32 species were divided into seven different metabolic clades based on the species dendrogram obtained. Based solely on the dendrogram, clades 5 and 6 may have been merged but they were split to understand the meaning of the group of reactions predominantly present in the clade number 6. The genome for *B. coahuilensis* m4-4 was not included in the analysis due to the low quality of its genome, leading automatically to a low quality of its metabolic network. *B. pumilus* SAFR-032 was also not included in this study since it appeared to be too distant to all the other clades for the subsequent statistical analysis to be relevant.

![Figure 2.5](image)

**Figure 2.5.** Metabolic networks reconstruction. The left part of the figure corresponds to the number of reactions present in the metabolic networks. The right part of the figure corresponds to a clustering performed on the 32 species based on the presence and absence of reactions. The 32 species have been divided into 7 metabolic clades based on this clustering.

The first insight that appeared from this analysis concerns a comparison between the obtained dendrogram and the previous phylogenetic analysis. Some of the metabolic
clades are conserved compared to the phylogenetic clustering (clades 1, 4, 5 and 6). On the other hand, metabolic clades 2, 3 and 7 do not correspond to phylogenetic clustering. This could be a sign of the convergent evolution of the metabolism of these species, which could have evolved to adapt to special environmental conditions. To facilitate the compositional analysis of the metabolic networks, a binomial test was performed on each reaction to find if the given reaction is statistically over-represented or under-represented in a given clade. The obtained list of over under represented-reactions were grouped in pathways as defined in the MetaCyc database. From this analysis, several hypotheses can be drawn about the metabolism of the analyzed species. For example, metabolic clade number 4 possesses statistically more reactions involved in ectoine biosynthesis. Since ectoine is already known for its osmoprotectant effect [228, 229], one could hypothesize that *Virgibacillus sp. Bac332, Virgibacillus* sp. Bac330, *Virgibacillus halodenitrificans Bac324, Oceanobacillus iheyensis HTE831* and *Virgibacillus sp. SK37* are highly tolerant against osmotic stress. The same species have statistically more reactions involved in glucuronoarabinoxylan degradation compared to others, which can be linked with utilization of plant cell wall polysaccharides. Species in metabolic clades number 5 and 6 are potentially capable of degrading plant cell walls as the clades are over-represented with reactions involved in rhamnogalacturonan type I degradation and D-galacturonate degradation. This assumption is validated by the presence of *B. licheniformis, B. subtilis* and *B. amyloliquefaciens* in these clades, which are all known to be associated with plant and plant material in nature [188, 230]. These results are also consistent with a study by Alcaraz L.D. *et al.* [134] with most of the variations in metabolic genes accounting for changes in the environment.
2.2.4 Secondary Metabolism of the Red Sea Bacilli Presents Considerable Potential for Production of Novel Active Molecules

There is a big diversity in the number of BGCs predicted in the different *Bacillus* species, ranging from 6 to 23 gene clusters per species (Figure 2.6). On average, there are 13 gene clusters per species (the median is 15). Mostly, there is no clear correlation between the number of BGC and phylogenetic clades. For example, *Virgibacillus* sp. Bac330 and *Virgibacillus* sp. Bac332 are very close, from a phylogenetic and metabolic point of view, but the first possess almost twice the number of gene clusters compared to the second. The species containing the most BGCs is a *Bacillus* strain coming from the Red Sea, *B. amyloliquefaciens* Bac57. It possess ten more gene clusters than the average.

![Figure 2.6. Number of predicted secondary metabolic gene clusters.](image)

To compare more precisely the different species from a secondary metabolism point of view, gene clusters that are shared between several species and gene clusters that
are already known were identified (Figure 2.7). In total, 417 gene clusters are present in the 32 studied species. Since the same gene cluster can be present in different species, in total, 200 different gene clusters can be distinguished: 55 are shared between at least two species while 145 are unique for a given species. Of those 200 gene clusters, 21 are known, while 179 are still unknown, highlighting a significant potential for new discoveries in the following years. In particular, 54 unknown gene clusters are only present in the genomes of the Red Sea species.

![Networks of co-occurrence of presence of secondary metabolism gene clusters.](image)

**Figure 2.7.** Networks of co-occurrence of presence of secondary metabolism gene clusters. Red sea species are displayed in boxes in different shades of green. Other species are displayed as ellipses. The product of a given secondary-metabolic gene cluster is displayed when identified. Unknown gene clusters are transparent.

Because most of the studied species (29 out of 32) possess complete genomes, co-localization of the gene clusters in different species was analyzed. Studying the localization of the BGC in the genomes could serve as a proxy to study the evolution
of apparition of those gene clusters in the *Bacillus* species. From this study, it appears that most of the gene clusters, shared by at least two species, co-localize in the genomes, pointing out that those gene clusters appeared in one of the common ancestors of those species. Among the 54 clusters of gene clusters shared by at least two species and present in strains with complete genomes, 38 are strictly co-localized, four possess genes that co-localize in at least two species but are also localized somewhere else in other genomes, and 12 are randomly localized in all the genomes (Figure 2.8). Interestingly, while clusters No. 49 and 51 perfectly co-localize in the genomes of *B. amyloliquefaciens DSM7* and *B. amyloliquefaciens* Bac57, clusters No. 48 and 51 that are also shared by both species co-localize poorly. Cluster No. 37, assigned to the production of the antibiotic Bacillaene is shared and co-localize in the genomes of *B. amyloliquefaciens* Bac57, *B. amyloliquefaciens DSM7* and *B. subtilis 168* although it is not present in the genome of *B. vallismortis* Bac111, suggesting a loss of this cluster in that species during evolution. On the other hand, cluster No. 21 is present in *B. marisflavi* Bac144 and in *B. pumilus SAFR 32* but not co-localized, suggesting an independent apparition of this gene cluster in those species. In the same manner, cluster No. 3 is present at the same position in all genomes from clade number 1 suggesting an apparition in the common ancestor of those species, while this cluster is also present in *B. clausii KSM K16* in another position suggesting an independent apparition in this species. Finally, the gene clusters present in cluster No. one seems to have appeared two times in the evolution, once in the *B. cereus* group and then in the *B. subtilis* group, while also appearing in *Virgibacillus sp.* Bac332.
Concerning the Red Sea strains, they possess, in total, 95 different BGCs. Amongst them, 38 are shared by at least one other species while 57 are unique. A known product was assigned to ten of the shared gene cluster (29% of them) and five of the unique ones (9%). It is worth noting that only three BGCs present in more than one species are shared solely within the Red Sea species. There is a big over-representation of BGCs involved in the production of antibiotics and lantibiotics, as it represents 66% of the identified BGCs (15 identified BGCs in the Red Sea species). The three identified BGCs shared by the biggest number of species, including the Red Sea Species, are involved in the protection against environmental changes. For example, genes involved in the production of Bacillibactin (a siderophore) [231], teichuronic acid (a part of the cell wall produced in low-phosphate conditions) [232], or ectoine (a protectant against osmotic stress), were identified.

Among the identified BGCs, those that are less spread within the studied species mostly correspond to antibiotics, lantibiotics, and antifungals, such as the identified
gene clusters involved in the production of the bacitracin antibiotic, the class II lantibiotic pseudomycoicidin, or the antifungal fengycin shared by four of the ten Red Sea species.

This highlights once again a potential for new secondary metabolites discovery, especially among molecules that would be produced by gene clusters uniquely present in those species. For example, it is worth noting that the species *B. foraminis* Bac44, *B. litoralis* Bac94 and *B. marisflavi* Bac144, while possessing between 12 and 15 BGCs, only share two of these gene clusters with other species, which is an indication of the specialization of the secondary metabolism to specific environmental conditions.
Chapter 3

3 Exploring the Biosynthetic Potential of Bacillus paralicheniformis
strains from the Red Sea

*B. licheniformis* is a Gram-positive facultative anaerobe, dubbed an industrial workhorse due to its use in several fields of biotechnology and its ability to secrete large amounts of commercially-used biomolecules and enzymes [233, 234]. These include specialty chemicals (e.g., citric acid and poly-γ-glutamic acids) and enzymes (e.g., proteases and α-amylases used in the food, detergent, textile, and paper industries) [235-238]. Most importantly, the antimicrobial capabilities of *B. licheniformis* have been widely reported [239-243] and several *B. licheniformis* strains have been used as biocontrol agents [230, 244-246] (e.g., EcoGuard). Moreover, *B. licheniformis* strains are used in the petroleum industry for microbial enhanced oil recovery [239, 247] due to their ability to produce lipopeptide biosurfactants.

*B. paralicheniformis* is a recently described new species within the *Bacillus* genus which was classified as *B. licheniformis* [248]. Despite the phylogenetic proximity and the expected biotechnological relevance, this species remains largely unexplored. The first description of *B. paralicheniformis* showed that it displayed a wider range of antimicrobial capabilities, despite being unable to produce licheinicidin or bacteriocins as does *B. licheniformis* [249].

A genomic-scale comparison of strains in both species can provide insights into their potential metabolic processes, their biosynthetic capabilities, and their stress adaptations. The evaluation of these properties helps to identify potential industrially relevant strains with novel and/or improved production capabilities of desired
compounds [250-253]. One way of assessing the production capabilities of these strains is through the identification of gene clusters that are co-localized in the genome and are responsible for the synthesis of secondary metabolites that, for example, exhibit antifungal, antimicrobial, herbicidal, or anticancer activities including NRPSs, PKSs, and RiPPs.

Ecologically, strains of *B. licheniformis* and *B. paralicheniformis* inhabit diverse environments including marine, freshwater, and food-related niches. This diversification in ecological and phenotypic properties has led *B. licheniformis* to become one of the most studied *Bacillus* species. In industrial settings, strains are often challenged with increased external osmolarity due to the high-level excretion of metabolites into the growth medium, threatening their productivity and/or viability [254-256]. Therefore, closely related *Bacillus* strains that are adapted to survive in high osmolarity environments, and have metabolic capacities similar to industrial strains are highly desirable.

One such environment is the Red Sea with relatively high salinity (36–41 p.s.u) and temperature (24 °C in spring, and up to 35 °C in summer) [257]. It is expected that strains from the Red Sea are able to produce a number of thermo-tolerant enzymes, and provide robust MCFs that are able to survive frequent exposure to high salinity and high temperature, and produce sturdier enzymes that might be better suited for industrial applications [258].

The biosynthetic potential of the two Red Sea strains, Bac48 and Bac84, isolated from the Rabigh harbor lagoon, along with nine *B. licheniformis* and three *B. paralicheniformis* strains were estimated. By grouping identified BGCs into families of gene clusters, the overall unutilized biosynthetic capabilities of strains from both
groups were highlighted. The unique presence of putative antimicrobial clusters in the Red Sea strains, focusing on one uniquely structured hybrid PKS/NRPS cluster that was identified in the genome of the Bac48, was highlighted.

3.1 Materials and Methods

3.1.1 Genomic Comparison

The overall genome similarities between *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84 were inspected using a dot plot that was generated with Gepard v1.40 [204]. Genome variation and synteny were inspected between the two strains using Sibelia v3.0.6 [259]. Prediction of genomic islands was done using IslandViewer v4 [260] and the identification of phage inserts was performed using PHASTER [261]. Finally, circular visualization of the genomes and annotated features were plotted using DNAPlotter [262].

3.1.2 Strain Identification and Phylogeny

To build the phylogenetic tree, orthologous protein groups were obtained using GET_HOMOLOGUES v1.3 [220] with default settings. Single-copy genes were then retrieved from the core genome output of GET_HOMOLOGUES. Each set of single-copy genes was aligned using MUSCLE v3.8.31 [214]. FASconCAT-G v1.02 [215] was used to concatenate the alignments and ProtTest v.3.4.2 [216] was used to predict the best-fit evolutionary model, which was used to generate the Maximum Likelihood (ML) tree in RAxML v8.2.3 [263] with the rapid bootstrapping algorithm and bootstrap value of 100. Additionally, another tree was generated using the complete core genome of the analyzed species. To build this phylogeny tree, orthologous protein groups (orthogroups) were obtained using OrthoFinder v2.2.1 [264] with default
settings. Briefly, an all-vs-all BLASTp analysis [265] was initially performed for the preliminary assignment of gene pairs. Gene pairs were then filtered based on the length-normalized BLAST bitscores to generate a gene pair graph for all-vs-all species. Next, orthogroups were inferred from the graph using the MCL tool v14.137 [266]. After establishing orthology, gene trees were constructed for all orthogroups in the core genomes (all species present) using the alignment-free tool DendroBlast [267] and FastMe v2.1.10 [268]. The Species tree was then reconstructed with support values from the consensus of all gene trees using STAG v1.0.0 (https://github.com/davidemms/STAG) and rooted based on duplication events using STRIDE v1.0.0 [https://doi.org/10.1093/molbev/msx259]. We visualized the consensus trees using iTOL [269].

3.1.3 Biosynthetic Gene Cluster Prediction

Only published strains with complete genomes were included in the analysis to ensure that the identified variations were indeed due to functional differences and not due to the quality of assembly. At the time of this study (May 2017), 12 strains satisfied these requirements, nine *B. licheniformis* and three *B. paralicheniformis*. To avoid potential bias resulting from using different annotation pipelines, all strains were re-annotated using the same set of tools and databases.

Biosynthetic and secondary metabolic gene clusters were predicted using antiSMASH v3.0 [107] with the *ClusterFinder* option [270]. Additionally, the *KnownClusterBlast* option was used to identify potential products for the clusters from the MIBiG database. Similar gene clusters from different genomes were then classified
into groups based on homology using BiG-SCAPE [271]. The resultant network from BiG-SCAPE was visualized in Cytoscape [272].

3.1.4 Urease Activity Testing

Christensen Urea Agar media [89] was used to test for urease activity. 5.82 grams of media was suspended in 200 ml of distilled water. After the medium has dissolved completely, it was autoclaved for 120 minutes. It was then cooled to 50°C and 30 ml of filter-sterilized 20% urea solution was aseptically added with a final concentration of 2%. The medium was dispensed in sterile (autoclaved) tubes set in a slanting position. The medium was inoculated in triplicates by spreading the pure culture of \textit{B. paralicheniformis} Bac48 and \textit{B. paralicheniformis} Bac84 with a control set for each strain. This was followed by overnight incubation at 37 °C. A change in color, compared to the control, was set as a positive indication.

3.2 Results and Discussion

3.2.1 Genome Sequencing and Assembly

Each SMRT cell produced 138,867 and 108,978 filtered subreads and 1,331,266,339 and 1,194,885,446 bases for \textit{B. paralicheniformis} Bac48 and \textit{B. paralicheniformis} Bac84, respectively (Table 3.1). \textit{De novo} assembly using the SMRT Analysis pipeline v2.3.0 using the smrtpipe.py commandline script with default parameters and the genomeSize parameter set to 6 Mb. Briefly, the assembly pipeline begins by filtering the SMRT reads (\texttt{minLength = 50, minSubReadLength = 50, readScore = 0.75}), followed by error correction. Minimum Seed Read Length is calculated automatically to produce a minimum coverage of 30x; 22,202 bp minimum read cut-off for \textit{B. paralicheniformis} Bac48 and 22,146 bp minimum read cut-off for \textit{B. paralicheniformis}
Bac84. Assembly of the corrected reads is carried out using the Celera assembler which generates an initial draft assembly followed by a polishing step using the Quiver program. The average read coverage is 234x and 215x for Bac48 and Bac84, respectively. The assembly produced two contigs for Bac48 (4,490,805 and 22,038 bp) and one contig for *B. paralicheniformis* Bac84 (4,400,372 bp). However, the second contig in *B. paralicheniformis* Bac48 had coverage from 2x to 15x and therefore was discarded.

The contigs from the initial assembly were checked for circularization using Gepard [204]. An overlap was observed at peripherals of each contig indicating that both genomes are circular. To circularize, each contig was split at a random location in the middle, then contigs were rejoined using minimus2 (part of AMOS) [205] producing circular contigs of 4,464,397 bp and 4,376,845 bp for Bac48 and Bac84, respectively. Finally, each contig was re-polished using the SMRT Analysis resequencing protocol taking the circularized contigs as the starting reference. The re-polishing step, which uses Quiver, is repeated multiple times taking the polished contig from the previous round, as input for the next, until convergence. Sequencing the genomes of the Red Sea strains using the SMRT sequencing platform produced 138,867 subreads with a mean length of 9,586 bp (234x genome coverage) for *B. paralicheniformis* Bac48 and 108,978 subreads with a mean length of 10,964 bp (215x genome coverage) for *B. paralicheniformis* Bac84 (Table 3.1). The assembly produced a single circular chromosome without plasmids for both strains. *B. paralicheniformis* Bac48’s circular chromosome is 4,464,381 bp in length containing 4,366 predicted open reading frames (ORFs); 51.5% of the genes are on the positive strand, and 48.5% are on the negative one. *B. paralicheniformis* Bac84’s circular chromosome is
4,376,831 bp in length containing 4,306 predicted ORFs; 47.8% of genes are on the positive strand, and 52.2% are on the negative one. Both genomes have 24 rRNAs and 81 tRNAs genes (Table 3.3).

**Table 3.1.** Basic statistics relating to the PacBio SMRT sequencing that was done for *B. paralicheniformis* B48 and B84. A single SMRT cell was sequenced for each strain.

<table>
<thead>
<tr>
<th></th>
<th>Bac48</th>
<th>Bac84</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polymerase Reads</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reads</td>
<td>86,252</td>
<td>75,932</td>
</tr>
<tr>
<td>Bases (bp)</td>
<td>1,334,497,375</td>
<td>1,197,038,773</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>22,785</td>
<td>21,104</td>
</tr>
<tr>
<td>Average Length (bp)</td>
<td>15,472</td>
<td>15,764</td>
</tr>
<tr>
<td><strong>Subreads</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subreads</td>
<td>138,867</td>
<td>108,978</td>
</tr>
<tr>
<td>Bases</td>
<td>1,331,266,339</td>
<td>1,194,885,446</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>12,578</td>
<td>15,215</td>
</tr>
<tr>
<td>Average Length (bp)</td>
<td>9,586</td>
<td>10,964</td>
</tr>
</tbody>
</table>

The completeness and contamination of *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84 were evaluated using CheckM (Version1.0.5) [206] (Table 3.2). Specifically, the taxonomic workflow was utilized, which looks for gene markers within the *Bacillus* genus.

**Table 3.2.** Levels of completeness and contamination in *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84 as determined in CheckM

<table>
<thead>
<tr>
<th></th>
<th>Completeness (%)</th>
<th>Contamination</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. paralicheniformis</em> Bac48</td>
<td>98.03</td>
<td>0.05</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em> Bac84</td>
<td>98.62</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Table 3.3. Summary of the genomes and annotation of nine *B. licheniformis* strains and five *B. paralicheniformis* strains.

<table>
<thead>
<tr>
<th>Strain</th>
<th>GenBank accession number</th>
<th>Genome size (Mb)</th>
<th>N contigs</th>
<th>N ORFs</th>
<th>N rRNA genes</th>
<th>GC content%</th>
<th>Environment</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bac48</td>
<td>CP023666</td>
<td>4.46</td>
<td>1</td>
<td>6,366</td>
<td>24</td>
<td>45.87</td>
<td>Mangrove mud</td>
<td>-</td>
</tr>
<tr>
<td>Bac84</td>
<td>CP023665</td>
<td>4.38</td>
<td>1</td>
<td>6,306</td>
<td>24</td>
<td>45.84</td>
<td>Microbial mat</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> DSM13</td>
<td>AE017333.1</td>
<td>4.22</td>
<td>1</td>
<td>6,216</td>
<td>21</td>
<td>46.19</td>
<td>N/A</td>
<td>[273]</td>
</tr>
<tr>
<td><em>B. licheniformis</em> HRBL-1STD</td>
<td>CP014781.1</td>
<td>4.25</td>
<td>1</td>
<td>6,293</td>
<td>24</td>
<td>45.92</td>
<td>Fermented food</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> WH-62</td>
<td>CP012110.1</td>
<td>4.29</td>
<td>1</td>
<td>6,343</td>
<td>26</td>
<td>46.10</td>
<td>Soil from salt mine</td>
<td>[274]</td>
</tr>
<tr>
<td><em>B. licheniformis</em> BI1302</td>
<td>CP017347.1</td>
<td>4.42</td>
<td>1</td>
<td>6,533</td>
<td>24</td>
<td>46.0</td>
<td>Soybean paste</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> SCK BI1</td>
<td>CP014795.1</td>
<td>4.30</td>
<td>1</td>
<td>6,372</td>
<td>24</td>
<td>45.93</td>
<td>Korean soybean paste</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> SCD8 14</td>
<td>CP016482.1 and CP016483.1</td>
<td>6.34</td>
<td>2</td>
<td>6,360</td>
<td>24</td>
<td>46.27</td>
<td>Korean soybean paste</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> SCD8 34</td>
<td>CP0164793.1</td>
<td>4.48</td>
<td>1</td>
<td>6,612</td>
<td>24</td>
<td>45.69</td>
<td>Korean soybean paste</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> SCD8 57</td>
<td>CP0164794.1</td>
<td>4.40</td>
<td>1</td>
<td>6,496</td>
<td>24</td>
<td>45.96</td>
<td>Korean soybean paste</td>
<td>-</td>
</tr>
<tr>
<td><em>B. licheniformis</em> YNP1-TSU</td>
<td>CM0037615.1</td>
<td>4.28</td>
<td>10</td>
<td>6,376</td>
<td>6</td>
<td>45.87</td>
<td>Hot spring</td>
<td>-</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em> ATCC 9054</td>
<td>CP000580.1</td>
<td>4.38</td>
<td>1</td>
<td>6,392</td>
<td>21</td>
<td>45.90</td>
<td>Salt</td>
<td>[275]</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em> BL-09</td>
<td>CP0010524.1</td>
<td>4.39</td>
<td>1</td>
<td>6,421</td>
<td>21</td>
<td>45.90</td>
<td>Natural fermented sour congee</td>
<td>-</td>
</tr>
<tr>
<td><em>B. paralicheniformis</em> MDK30</td>
<td>CP0010522.1</td>
<td>4.35</td>
<td>1</td>
<td>6,363</td>
<td>24</td>
<td>45.90</td>
<td>Rhizosphere</td>
<td>-</td>
</tr>
</tbody>
</table>

Genomic island (GI) prediction identified five GIs in *B. paralicheniformis* Bac48 that include three unique regions (totaling 64.3 Kb and representing 1.4% of the genome) and 14 GIs in *B. paralicheniformis* Bac84 (totaling 142.8 Kb and representing 3.3% of the genome) (Figure 3.1, Table 3.4). Analysis of prophage sequences in the
genome revealed three prophage regions in *B. paralicheniformis* Bac48 (124 genes), with one of them partially overlapping with a GI. Similar analysis in *B. paralicheniformis* Bac84 also identified three prophage regions (121 genes), with two of them partially overlapping with GIs (Figure 3.1, Table 3.5). When compared with the complete genome, the relative percentage of these prophages is relatively low (2.4% for *B. paralicheniformis* Bac48 and 2.6% for *B. paralicheniformis* Bac84).

These values suggest a reduced number of horizontally transferred elements compared to the genome of the industrially important strain *B. licheniformis* DSM 13 where GIs represent 4.8% and prophages represent 6.2% of the genome. This paucity is an advantage of Bac48 and Bac84 strains, as removing GIs and prophages is a necessary step for stabilizing minimized genomes and for streamlining metabolism in biotechnological hosts [276].

**Figure 3.1.** Circular plots of (a) *B. paralicheniformis* Bac48 and (b) *B. paralicheniformis* Bac84 genomes, showing the distribution of genomic islands, prophages and biosynthetic genes in the genomes. The tracks show the following features starting from the outermost track; 1st track (pink): genes on the positive strand; 2nd track (blue): genes on the negative strand; 3rd track (yellow): biosynthetic gene clusters; 4th track (red): horizontally transferred genes; 5th track (green): genes in prophage regions; 6th track: GC-plot where purple and green correspond to below and above average GC content, respectively; 7th track: GC-skew where purple and green correspond to below and above average GC-skew, respectively.
Table 3.4. List of genomic island regions in the genomes of *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84, predicted using IslandViewer [213].

<table>
<thead>
<tr>
<th>Genome</th>
<th>Name</th>
<th>Start position</th>
<th>End position</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. paralicheniformis</em> Bac84</td>
<td>GI₁</td>
<td>249,229</td>
<td>268,017</td>
<td>18,788</td>
</tr>
<tr>
<td></td>
<td>GI₂</td>
<td>792,253</td>
<td>798,734</td>
<td>6,481</td>
</tr>
<tr>
<td></td>
<td>GI₃</td>
<td>1,003,363</td>
<td>1,020,318</td>
<td>16,955</td>
</tr>
<tr>
<td></td>
<td>GI₄</td>
<td>1,012,588</td>
<td>1,019,457</td>
<td>6,869</td>
</tr>
<tr>
<td></td>
<td>GI₅</td>
<td>1,223,810</td>
<td>1,227,971</td>
<td>4,161</td>
</tr>
<tr>
<td></td>
<td>GI₆</td>
<td>1,917,364</td>
<td>1,929,075</td>
<td>11,711</td>
</tr>
<tr>
<td></td>
<td>GI₇</td>
<td>1,973,909</td>
<td>1,994,365</td>
<td>20,456</td>
</tr>
<tr>
<td></td>
<td>GI₈</td>
<td>1,986,695</td>
<td>1,992,844</td>
<td>6,149</td>
</tr>
<tr>
<td></td>
<td>GI₉</td>
<td>2,300,230</td>
<td>2,307,438</td>
<td>7,208</td>
</tr>
<tr>
<td></td>
<td>GI₁₀</td>
<td>2,727,262</td>
<td>2,734,510</td>
<td>7,248</td>
</tr>
<tr>
<td></td>
<td>GI₁₁</td>
<td>2,823,477</td>
<td>2,841,438</td>
<td>17,961</td>
</tr>
<tr>
<td></td>
<td>GI₁₂</td>
<td>3,363,336</td>
<td>3,370,202</td>
<td>6,866</td>
</tr>
<tr>
<td></td>
<td>GI₁₃</td>
<td>3,395,245</td>
<td>3,403,627</td>
<td>8,382</td>
</tr>
<tr>
<td></td>
<td>GI₁₄</td>
<td>3,399,935</td>
<td>3,403,534</td>
<td>3,599</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome</th>
<th>Name</th>
<th>Start position</th>
<th>End position</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. paralicheniformis</em> Bac48</td>
<td>GI₆₉</td>
<td>69,655</td>
<td>80,660</td>
<td>11,005</td>
</tr>
<tr>
<td></td>
<td>GI₇₀</td>
<td>776,075</td>
<td>799,315</td>
<td>23,240</td>
</tr>
<tr>
<td></td>
<td>GI₇₁</td>
<td>1,884,577</td>
<td>1,898,632</td>
<td>14,055</td>
</tr>
<tr>
<td></td>
<td>GI₇₂</td>
<td>2,446,070</td>
<td>2,454,030</td>
<td>7,960</td>
</tr>
<tr>
<td></td>
<td>GI₇₃</td>
<td>3,970,771</td>
<td>3,978,833</td>
<td>8,062</td>
</tr>
</tbody>
</table>

Table 3.5. Predicted prophage regions in *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84 and their overlap with GIs. Scores were obtained using PHASTER [277] scoring scheme. Most Common Phage shows the phage ID(s) with the highest number of proteins most similar
to proteins in the region. Overlap percentage shows the length of overlap region with respect to the length of prophage.

<table>
<thead>
<tr>
<th></th>
<th>Size (Kb)</th>
<th>Completeness</th>
<th>PHASTER Score</th>
<th>No. of proteins</th>
<th>Region</th>
<th>Most common phage</th>
<th>Overlapping GI region</th>
<th>Size of overlap (Kb)</th>
<th>GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B. paralicheniformis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac48</td>
<td>28.1</td>
<td>Incomplete</td>
<td>50</td>
<td>17</td>
<td>55436-83598</td>
<td>PHAGE_Bacillus_L_G_NC_023 719(2)</td>
<td>55436-80,660</td>
<td>25.2</td>
<td>40.63</td>
</tr>
<tr>
<td></td>
<td>36.2</td>
<td>Intact</td>
<td>110</td>
<td>46</td>
<td>916501-952741</td>
<td>PHAGE_Brevib_limmen1_Jim1_NC_029104(7)</td>
<td>-</td>
<td>-</td>
<td>47.35</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>Intact</td>
<td>97</td>
<td>61</td>
<td>3407364-3451392</td>
<td>PHAGE_Bacillus_l_phi105_NC_004167(33)</td>
<td>-</td>
<td>-</td>
<td>42.68</td>
</tr>
<tr>
<td><strong>B. paralicheniformis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bac84</td>
<td>36.2</td>
<td>Intact</td>
<td>110</td>
<td>46</td>
<td>1415893-1452151</td>
<td>PHAGE_Bacillus_l_BalMu_1_NC_030945(8)</td>
<td>-</td>
<td>-</td>
<td>47.23</td>
</tr>
<tr>
<td></td>
<td>25.8</td>
<td>Incomplete</td>
<td>50</td>
<td>17</td>
<td>2279889-2305734</td>
<td>PHAGE_Bacillus_L_G_NC_023 719(2)</td>
<td>2,300,230-2305734</td>
<td>5.5</td>
<td>40.97</td>
</tr>
<tr>
<td></td>
<td>57.6</td>
<td>Intact</td>
<td>97</td>
<td>58</td>
<td>3349191-3406828</td>
<td>PHAGE_Bacillus_l_phi105_NC_004167(33)</td>
<td>3,363,336-3,370,202</td>
<td>3.6</td>
<td>43.20</td>
</tr>
</tbody>
</table>

### 3.2.2 Phylogenetic Positioning of the Red Sea *Bacillus paralicheniformis* Strains

For a comprehensive comparative analysis of the genomes and to ascertain the phylogenetic position of Bac48 and Bac84 within the *Bacillus* genus, a phylogenetic tree was generated using the concatenated alignment of 250 single-copy genes.
Another tree was generated using 494 gene trees accounting for the core genome of the species. According to Wang and Ash [278], phylogenetic trees of *Bacillus* that use this approach are more in line with results from the whole genome feature frequency profiling and are more accurate than phylogenetic trees based on single marker genes such 16S rRNA, *gryB* (gyrase subunit B) or *aroE* (shikimate-5-dehydrogenase) genes.

**Figure 3.2.** Maximum-likelihood phylogenetic tree of 38 genomes of the class Bacilli constructed using single-copy core genes. *Clostridoides difficile* CD196 was used as the outgroup.
Other than the two Red Sea strains, the phylogenetic analysis included nine *B. licheniformis* strains, three *B. paralicheniformis* strains and 24 genomes from other representative species within the genus [279]. However, it was noted that *Listeria* species are misplaced in the phylogenetic analysis and accordingly were not included in the phylogeny tree.

All resulting trees (Figure 3.2 and 3.3) show the phylogenetic proximity of Bac48 and Bac84 to *B. paralicheniformis* strains and reveals them to be more distantly related to *B. licheniformis* than previously reported [203].

### 3.2.3 Genomic Overview of *B. licheniformis* and *B. paralicheniformis* Strains

All of the strains included in this study (nine *B. licheniformis* and five *B. paralicheniformis*) have comparable genome sizes (4.22 - 4.48 Mb) while the number
of predicted ORFs range from 4,256 in *B. licheniformis* DSM13 to 4,612 in *B. licheniformis* SCDB 34 (Table 3.3). Across these 14 genomes, the pangenome size was found to be 6,299 gene families (Figure 3.5a) with 3,324 core ones (Figure 3.5b) and 2,975 accessory ones. *B. paralicheniformis* BL-09 (505) has the highest number of unique genes while *B. paralicheniformis* MDJK30 has the lowest (45) (Fig. 3.4).

**Figure 3.4.** Intersection of orthologous gene families between the 14 genomes. Genomes are represented with filled circles. The core genome is represented in the first bar while unique genes are represented in bars with unconnected filled circles. The figure was generated using UpsetR [280] and the bars were ordered by frequency.

The core genome represents 74.3% of the average number of ORFs. Assigned Clusters of Orthologous Groups (COGs) to the coding sequences of the core genome are associated with key functions: amino acid transport and metabolism (11%),
carbohydrate transport and metabolism (9%), inorganic ion transport and metabolism (7%), transcription (8%), and translation, ribosomal structure and biogenesis (5%) (Figure 3.6).

Figure 3.6. COG assignments in the core genome showing the conservation of housekeeping functions.

The unique features of the core genome of the *B. paralicheniformis* strains, including the Red Sea strains, was compared against the core genome of *B. licheniformis* strains. Of these genes, 24 are biosynthetic genes present in two NRPS clusters that synthesize the lipopeptides fengycin and bacitracin.
Further analysis revealed that some industrially relevant genes distinguish strains from both species. The first is the presence of a complete urease operon, consisting of nine genes in *B. paralicheniformis*, which is in line with previous reports [248, 249]. The presence of the urease operon in *B. paralicheniformis* creates a number of possible industrial applications for it. Urease plays an important role in the utilization of environmental nitrogenous compounds through the hydrolysis of urea into ammonia and carbon dioxide. The presence of the urease operon in bacteria has been linked with the capability to induce the precipitation of calcium carbonate, a process with a wide range of applications which include the production of self-healing concrete, bioremediation of sites contaminated with heavy metals, and sequestration of CO$_2$ [281]. In fact, precipitation of carbonates by ureolytic bacteria is reported as the most commonly found process, most straightforward and most easily controlled mechanism. The latter has particularly large implications as it is seen as a promising method to reduce CO$_2$ emissions and curb climate change [281-284]. Species of the *Bacillus* genus have been reported as strong players in the industrial application of biomineralization of calcium carbonate [285]. This capability has not been previously reported for *B. paralicheniformis* and warrants further studies, as it could be relevant for industry. Urease activity was verified in Christensen Urea Agar media for *B. paralicheniformis* Bac48 (Figure S3.1).

The second unique feature in *B. paralicheniformis* is a group of enzymes that collectively play critical roles for the depolymerization of xylan, a hemicellulose constituent of plant cell wall and the second most common feedstock source [286]. These enzymes are namely xylan 1,4-β-xylosidase (EC 3.2.1.37), glucuronoarabinoxylan endo-1,4-beta-xylanase (GAX) (EC 3.2.1.36) and arabinoxylan
arabinofuranohydrolase (AXH) (EC 3.2.1.55). 1,4-β-xylosidase and endo-1,4- β-xylanase are required to break-down the xylan backbone. β-xylosidase releases xylose monomers from xylo-oligosaccharides [287] and GAX cleaves the β-1,4 glycosidic linkage between xylose residues. Additionally, AXH cleaves arabinose substituents in the xylan backbone [287, 288]. Bacteria’s ability to break-down abundant feedstock sources (hemicellulose), is often desired to increase the economic and environmental efficiencies of cell factories that are able to use plant biomass (e.g., agricultural waste) as a nutritional source. The interest in xylosidase, is also motivated by its industrial potential application in the paper industry, as complete bio-degradation of xylans improves the bio-bleaching process and hence reduces chlorine usage [288]. Concerning the synthesis of bioactive compounds, Nieto-Domínguez et al. [288] hinted at the involvement of xylosidase in the synthesis of biosurfactant compounds with antimicrobial properties due to its regioselective transxylosylation.

3.2.4 Exploring the Biosynthetic Potential of *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84

To evaluate the biosynthetic potential of the two species (*B. licheniformis* and *B. paralicheniformis*), nine complete *B. licheniformis* and five complete *B. paralicheniformis* genomes, including the two Red Sea strains, were used (Table 3.3). On average, the analyzed genomes are each comprised of 34 putative BGCs, predicted by antiSMASH [107]. These clusters encode peptides/proteins associated with the biosynthesis of one of the following types of secondary metabolites: bacteriocins, lanthipeptides, NRPS, type III PKSs, hybrid NRPS/PKS clusters and unclassified clusters (Figure 3.7). It is clear that, overall, *B. paralicheniformis* strains have more biosynthetic
genes (8.5% of the average number of predicted ORFs) compared to \textit{B. licheniformis} (6.3% of the average number of predicted ORFs). In this study, I focus on two types of compounds that are often associated with high antimicrobial activity: 1/ modular clusters (NRPS and modular PKS), and 2/ ribosomally synthesized peptides, namely modified and unmodified bacteriocins.

\textbf{Figure 3.7.} Distribution of genes in biosynthetic gene clusters in nine \textit{B. licheniformis} and five \textit{B. paralicheniformis} genomes. Clusters with modular genes are marked with a star and clusters encoding for ribosomally synthesized peptides are marked with a triangle.

A total of 480 BGCs were classified into 54 groups (also referred to as gene cluster families GCFs) using scoring similarity networks as implemented in BiG-SCAPE (Figure 3.8) [271]. Interestingly, only 6 GCFs (ca. 11% of the total) were assigned to known products or a similar pathway using threshold similarity of 60%. This highlights the limited knowledge available for the analyzed strains. Furthermore, these unexplored secondary metabolites can potentially provide new antimicrobial agents and compounds of industrial importance, thus warranting future studies of these BGCs to identify their functions.
3.2.4.1 Nonribosomal Peptides and Modular Polyketides

Modular genes in NRPS and PKS clusters are of critical importance when assessing the biotechnological value of strains. Firstly, known polyketides and nonribosomal peptides are often found to be responsible for producing important antibiotic, antifungal, and antitumor pharmaceutical products [289]. Additionally, understanding the organization of domains in modules could help advance efforts for the synthesis of products with amended physiochemical properties and enhanced bioactivity [290].

The identified NRPS clusters were grouped into four GCFs with known products (Figure 3.9). The first group, found to be conserved across all *B. licheniformis* and *B. paralicheniformis* strains, has on average 46 genes per genome and shared 46% of its genes with the bacillibactin cluster, a siderophore commonly produced in the *Bacillus*
genus [291]. The second GCF of NRPS clusters has 43 genes that include the lichenysin operon \((\text{licABC})\), an efficient biosurfactant from the surfactin family [150, 292, 293]. The third and fourth NRPS clusters were only detected in the \(B. \text{paralicheniformis}\) strains, including \(B. \text{paralicheniformis} \text{Bac48}\) and \(B. \text{paralicheniformis} \text{Bac84}\), with 50 and 45 genes and with 86% and 100% similarity to the BGC of the antifungal fengycin [294-296] and the narrow-spectrum antibiotic bacitracin [297-300], respectively.

Figure 3.9. Heat map visualization of the number of genes in BGC groups. There are 54 GCFs with BGCs shared by at least two genomes and 20 BGCs identified to be unique (present in one genome). The number of genes in each GCF is normalized based on the maximum number of genes. Putative clusters are predicted using the ClusterFinder algorithm as implemented in antiSMASH [107]

A hybrid PKS/NRPS cluster was identified in the genome of \(B. \text{paralicheniformis}\) Bac48. To the best of our knowledge, this is the first trans-acyltransferase (trans-AT) PKS/NRPS cluster reported in strains of this species. Trans-AT PKS biosynthetic clusters are an
Figure 3.10. Structure of the hybrid PKS/NRPS cluster present in *B. paralicheniformis* Bac48. Biosynthetic genes are identified with red arrows while non-biosynthetic genes are identified with blue ones. Domains are abbreviated as follows: adenylation (A), ketosynthase (KS), ketoreductase (KR), condensation domain (C), acyl carrier protein (ACP), peptidyl carrier domains (PCP), c-methyltransferase (cMTA), o-methyltransferase (oMT), enoyl-CoA hydratases (ECH), dehydratase (DH), acyltransferase docking site (Trans-AT docking) and acyltransferase (AT).

emerging class of modular PKSs that are becoming more commonly found in microbial genomes. Structurally, a trans-AT PKS cluster is different from a typical cis-AT PKS in that the AT domain, which loads the substrate onto acyl carrier protein domains (ACPs), is encoded in a separate ORF as independent polypeptide and not integrated into the assembly line [301]. Other trans-AT PKS/NRPS clusters reported within the genus *Bacillus* are the antibiotic bacillaene *pksX* found in *B. subtilis* and its homologous the *baeX* operon in *B. amyloliquefaciens*.

The hybrid trans-AT PKS/NRPS cluster is located 14.6 Kb downstream of a lichenysin synthase operon (*licABC*). The cluster was predicted as a single BGC along with the lichenysin operon; however, due to the large non-biosynthetic gap between the two clusters, the predicted cluster was split into two. The resultant BGC is composed of 29 genes, eight of which were horizontally transferred. The cluster extends over 82.8 Kb, which is close in size to the bacillaene and pksX cluster (~80 Kb) (Figure 3.10) [190].

One of the architectural differences between this cluster and the other trans-AT PKS clusters in *Bacillus* is that there is one NRPS module with its domains (adenylation, condensation and peptidyl carrier domains) extended over two ORFs, while on the other hand, the other *Bacillus* trans-AT clusters have two NRPS modules in two ORFs.
The cluster encodes nine multi-domain ORFs, consisting of one adenylation domain (A), 16 ketosynthase domains (KS), ten ketoreductase domains (KR), two peptidyl carrier domains (PCP), 18 acyl carrier protein domains (ACP), nine dehydratase domains (DH), two enoyl-CoA hydratases domains (ECH), two c-methyltransferase domains (cMT), two o-methyltransferase domains (oMT), and one condensation (C) domain. I also identified truncated AT domains that could be used as binding sites for trans-acting AT. The order of the PKS domains and the absence of integrated AT domains in all of the nine PKS/NRPS ORFs in this gene cluster suggest that this is indeed a trans-AT PKS cluster, with two trans-acting AT domains encoded by ORFs that are independent from the polypeptide assembly line.

Moreover, the cluster showed similarity to known trans-AT PKSs (71% to elansolid and 57% to thiomarinol) (Figure 3.11). Although the cluster had the highest number of genes similar to genes in the elansoid BGC, its modular organization was more similar to that of the thiomarinol cluster. Comparing this cluster to known PKS clusters in *Bacillus* revealed a 57% similarity to the bacillaene cluster in *B. amyloliquefaciens* FZB 42.
Figure 3.11. Homology between the hybrid PKS-NRPS cluster and other clusters with known products

The two strains, *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84, have been previously reported to have different antibiotic production capabilities [203]. In spite of this variation in antimicrobial production, genomic comparison showed that the two genomes share 3,953 orthologous genes, of which 3,909 are single copies. The high similarity between the two genomes is also clearly demonstrated by whole-genome alignment, as the genomes are highly syntenic, except for three large regions present in *B. paralicheniformis* Bac48 and missing in *B. paralicheniformis* Bac84 (Figure 3.12 a, b). The largest non-syntenic block is a ~83 Kb region in which the previously described NRPS/trans-AT PKS cluster resides, an indication that the antimicrobial
variation could be a result of this uncharacterized cluster.
Figure 3.12. Similarity between the genomes of *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84. a) Circos figure showing synteny blocks between *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84. Regions I, II and III are regions in Bac48 that are missing in *B. paralicheniformis* Bac84. The coordinates of the two largest regions are from 1,884,514 to 1,968,250 with a total length of 83,736 bp and from 3,977,600 to 3,997,816 with a total length of 20,216 bp. b) A dotplot between *B. paralicheniformis* Bac48 and *B. paralicheniformis* Bac84 genomes which shows the high concordance between the two genomes.

3.2.4.2 Ribosomally Synthesized Peptides and Post-Translationally Modified Peptides (RiPPs): Bacteriocins and Lanthipeptide.

There is at least one bacteriocin cluster family in each of the analyzed genomes. One of the families was conserved across all the *B. licheniformis* and *B. paralicheniformis* strains, with an average of nine genes. The clusters in this group had three biosynthetic genes (ribosomal mythelotransferase accessory protein, carbohydrate esterase, and an uncharacterized protein) and showed no similarity to any known bacteriocin. Another bacteriocin family was found to be encoding for a circularized head-to-tail bacteriocin. The cluster was detected in the genomes of *B. paralicheniformis* Bac84 and *B. paralicheniformis* strains ATCC 9945a and BL-09. Clusters in this family had mostly uncharacterized genes and showed no evident similarity to any known bacteriocin.

Lanthipeptides are a type of bacteriocins that often contain lanthionine and undergo posttranslational modification. The fact that these post-transitional modification genes are highly conserved assists in the *in silico* prediction of lanthipeptide clusters [302]. Immunity genes and ABC transporters are other common functional elements of lanthipeptide clusters [303].

It is noted that two-component class II lanthipeptides, in which two peptides processed by a modifying enzyme (*lanM*) [304], are the most common lanthipeptides in the analyzed genomes. *B. licheniformis* strains have three genes mapping to *LchA1*,...
LchA2 and LchM1 in the class II lanthipeptide lichenicidin VK21 cluster. The absence of lichenicidin posttranslational modification genes in *B. paralicheniformis* is a distinguishing feature between the two species. A lanthipeptide cluster was detected in the *B. paralicheniformis* genomes (MDJK30, BL-09 and ATCC 9945a), and in *B. licheniformis* SCDB 34 with a mersacidin-like structural gene. The cluster is predicted to be of class II lanthipeptides as it has the *lanM* posttranslational modification enzyme. However, other mersacidin genes (*mrsK2, mrsR2, mrsF, mrsG* and *mrsE*) were not detected, indicating that the cluster might be involved in the synthesis of a new product with partial genomic similarity to the genes encoding for the antibiotic mersacidin. No lanthipeptide clusters were predicted in the Red Sea strains; however, the genomes of *B. paralicheniformis* Bac84 harbored a lantibiotic-like cluster, with the subtilin biosynthesis posttranslational modification gene *spaB* that encodes the dehydratase of the lanthionine in the subtilin gene cluster (PFAM: PF04738) and subtilin ABC transporter permease (*spaG*). The cluster was not predicted as a lanthipeptide as it lacked other genomic features including the posttranslational modification enzyme necessary for the cyclization of lanthionine (*spaC* in subtilin) and other immunity genes. Additionally, seven genes in the cluster were similar to genes in the rhizoctin biosynthetic cluster, which is an unusual peptide with antimicrobial activity.

When evaluating the industrial value of known proteins synthesized by *B. paralicheniformis* strains, the analysis showed that known products encoded by NRPS clusters (e.g., lipopeptides) significantly outnumber known products encoded by RiPP clusters as only lichenicidinVK21 was identified in these clusters. This difference is
expected as most of the previous focus in Firmicutes has been on lipopeptides, especially as these are highlighted to be attractive pharmaceutical or/and industrial products. Investigating the functions of genes in RiPPs showed that, although some of their genes are similar to the ones in known clusters, they are incomplete, preventing us from using databases, such as MIBiG, to determine their final products. This lack of information leads us to conclude that the *B. licheniformis* and *B. paralicheniformis* RiPPs are understudied. However, genes in RiPPs from partial genomes, encode other known products, for instance the recently discovered novel lanthipeptide produced by *B. paralicheniformis* APC 1576 (formicin) [305], the bacteriocin produced by *B. licheniformis* 490/5 (bacillocin 490) [306] and the bacteriocin-like lichen produced by *B. licheniformis* 26L-10/3RA [307].
Chapter 4

4 A Draft Genome-Scale Metabolic Model of *B. paralicheniformis*

**Bac48**

Genome-scale metabolic models are excellent methods to be utilized for the investigation of an organism metabolism. Specifically, a metabolic model assists in identifying critical dependencies between different elements of the genomes, which cannot be easily obtained by analyzing pathways in the organism independently. For instance, the accurate metabolic reconstruction of a specific genome can be used to design an efficient MCF by recognizing genetic interventions that optimize the production of a specific metabolite or its precursor. It can also be used to evaluate the strain potential as a minimized genome through the prediction of essential genes for growth. Finally, metabolic models can predict essential genes in the genomes of pathogens as putative drug targets.

The most notable metabolic model in the *Bacillus* genus is that of *B. subtilis* 168. It has been subjected to different rounds of model generation and simulation [227, 308]. The well-studied nature of the organism and the availability of numerous studies provide a wealth of information for model refinement and validation. The availability of other ‘omics’ data for the strain (transcriptomics, metabolomics and proteomics) facilitate the reconstruction of omics-integrated models for *B. subtilis* 168.

Given the aforementioned advantages of metabolic models, and the scarcity of *Bacillus* metabolic model, I will describe the first draft model of *B. paralicheniformis* Bac48. The model, referred to as *iPARA1056* (*in silico* B. *paralicheniformis* with 1056}
genes), accurately predicts in vitro growth rate for the strain, as well as growth phenotypes under different substrates.

4.1 Methods

4.1.1 Bacterial Preparation for Phenotype Microarray Plates Inoculation

Biolog Phenotype microarray (PM) assay was used to characterize the cellular phenotypes of B. paralicheniformis Bac48 strain. The frozen-stock bacterial cells stored at -80°C were grown on rich LB medium for 24h at 37°C, followed by two consecutive growth-steps on diluted LB medium (1:50). Isolated colonies were removed from the agar plate using a sterile swab and added to a tube containing 16 ml of physiological solution (0.9% NaCl) until a cell suspension of 80% T (transmittance) was measured using the Biolog turbidimeter. Biolog microplates for carbon sources (PM1 and PM2), nitrogen sources (PM3, PM6, PM7 and PM 8), osmolytes (PM9) and pH (PM10) have been tested. Different IF (Inoculating Fluids) solutions were prepared in order to provide the minimal substrates necessary to guarantee bacterial growth once in presence of the carbon/nitrogen sources that the studied bacterium was able to use. In the case of osmolyte and pH plates, a rich medium (LB 1:2) has been used. Inoculation of the PM plates was performed according to the Biolog PM protocol for Gram positive bacteria. Plates were incubated at 37°C in the OmniLog plate reader. The reduction of the redox potential indicator (Dye F, Biolog) due to the metabolic activity of bacterial cells caused the formation of a blue-purple color which is recorded by a CCD camera every 30 minutes [309]. The color changes was monitored and measured for 120 hours. The kinetic colorimetric data were stored in computer files and analyzed using the Biolog software.
4.1.2 Model Reconstruction and Simulation

To generate the first draft of the metabolic model, ModelSeed automatic reconstruction pipeline [69] was used. For gap filling, ModelSeed and Meneco [310] were used to add reactions to simulate growth on LB media. Since there are no transport reactions for *B. paralicheniformis* in transportDB [311], those for *B. licheniformis* were retrieved and used, only if supported by gene(s) in the model. The biomass equation generated by ModelSeed for Gram-positive bacteria was used as a template and modified according to known biomass components of *Bacillus* species. Biomass precursors were individually checked to confirm their suitability for *B. paralicheniformis*. Dead-end metabolites were detected and removed from the model using the *removeDeadEnds* function in COBRA toolbox. To find essential genes, all reactions that are exclusively associated with the knocked-out gene are removed from the model. The knock-out experiments also account for genes that partially encode a protein. All reactions and metabolites were maintained in Systems Biology Markup Language (SBML) format, level 2. All flux-balance and flux-variability analyses were completed using COBRA toolbox [312] utilizing glpk solver version 4.63.

4.2 Results and Discussion

4.2.1 Phenotypic Characterization of *B. paralicheniformis* Bac48

Phenotype MicroArray technology (Biolog) has a total of 25 plates that test cellular growth under different conditions [309]. Each plate has 96 wells, where each well has a specific substrate (metabolite) to be tested. For this study, Biolog plates PM1 and PM2 were used for carbon sources (a total of 192 wells, two of which are negative controls) (Figure S2.1), PM3 was used for nitrogen sources (96 wells with one
negative control) (Figure S2.2), PM6-8 were used for peptide nitrogen sources (Figure S2.3) and PM 9-10 for osmolytes and pH ranges (192 wells) (Figure S2.4). The strain could utilize a total of 153 carbon substrates with the highest dye reduction observed for d-xylene and d-arabinose. For nitrogen sources, the strain grew in 67 wells, with the highest dye reduction observed for the dipeptide ala-asp (a maximum growth curve height of 268). For PM9 and PM10, the strain grew under all osmolyte stresses, including high NaCl concentrations (10%) (Fig. 4.1), except that the log growth phase started late in wells with sodium lactate as the osmolyte. Finally, the strain could not grow in wells with pH lower than 4.5 except with the addition of l-norvaline as an osmoprotectant. It also had limited growth with the addition of l-lysine, l-homoserine and urea.

![Figure 4.1](image1.png)

**Figure 4.1.** Maximum height of growth curve for *B. paralicheniformis* Bac48 under different concentrations of NaCl.

### 4.2.2 Model Reconstruction

To reconstruct a metabolic model, one must start by utilizing all available genome annotation data, and metabolic information available in biochemical databases including: KEGG [313], MetaCyc [314] and SEED [315] among others. In the reconstruction of *B. paralicheniformis* Bac48, I first utilized ModelSeed [69] to reconstruct a draft model for *B. paralicheniformis* Bac48. The reconstruction resulted
in 1056 genes (24.1% of the genome). The percentage of gene-associated reactions in the B. subtilis 168 model (iBSU1103) and the B. licheniformis WX-02 model (iWX1009) are 26.42% and 23.8% respectively, which are comparable to that of iPARA1056. Exchange reactions are expectedly not associated with genes, as they are incorporated in the model to simulate experimental conditions, such as exchange of oxygen, carbon sources, nitrogen sources and ions. ModelSeed added 39 reactions as part of gap-filling, for the model to grow in LB media, and 29 reactions were added by Meneco.

To refine the automatically reconstructed model, each reaction was checked for the following, when possible: 1/ mass and charge balance 2/ directionality 3/ agreement with data in literature. First, the reaction was checked in other closely related high-quality models, primarily that of B. subtilis 168 (iBSU1103) [227] and of B. licheniformis WX-02 (iWX1009) [316]. If the reaction was not in the model, the Rhea database [317], which has manually curated reactions, was used. All reactions with quinones as electron receptors were individually checked and changed in directionality, if the necessary. All of the above-mentioned refinements are reflected in the number of reactions from the draft model to the number of the reactions in the resulting model, after manual refinement and gap-filling as shown in Figure 4.2.
H+ ions were the most consumed and produced metabolites in the model in 48.6% of the reactions (25 of which are in reactions with proton symports). Water (H₂O) is present in 29.8% of the reactions, followed by adenosine triphosphate (ATP), adenosine diphosphate (ADP), phosphate, reduced nicotinamide adenine dinucleotide (NADH), nicotinamide adenine dinucleotide phosphate (NADP), reduced nicotinamide adenine dinucleotide phosphate (NADPH), ammonia (NH₃) and adenosine monophosphate (AMP).

Genes in the reconstruction were classified into functional categories using Clusters of Orthologous Groups (COGs) (Table 4.1), they fell mainly into the functional categories for ‘energy production and metabolism (C)’ (11.6%), ‘amino acid transport and metabolism (E)’ (18.7%), ‘carbohydrate transport and metabolism (G)’ (15.8%), ‘coenzyme transport and metabolism (H)’ (8.1%) and ‘inorganic ion transport and metabolism (P)’ (8.5%). The COG analysis was used to assure that the model had all of the components necessary for growth. The most enriched categories in the model were expectedly ‘amino acid transport (E) and metabolism’ and ‘carbohydrate
transport and metabolism (G’), while the least enriched was ‘signal transduction mechanisms (T)’ (0.29%) and ‘defense mechanisms (V)’ (0.38%).

Table 4.1. COG assignments for genes in the metabolic network

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy production and conversion (C)</td>
<td>122</td>
</tr>
<tr>
<td>Amino acid transport and metabolism (E)</td>
<td>197</td>
</tr>
<tr>
<td>Nucleotide transport and metabolism (F)</td>
<td>60</td>
</tr>
<tr>
<td>Carbohydrate transport and metabolism (G)</td>
<td>166</td>
</tr>
<tr>
<td>Coenzyme transport and metabolism (H)</td>
<td>85</td>
</tr>
<tr>
<td>Lipid transport and metabolism (I)</td>
<td>72</td>
</tr>
<tr>
<td>Secondary metabolites biosynthesis (Q)</td>
<td>36</td>
</tr>
<tr>
<td>Translation, ribosomal structure and biogenesis (J)</td>
<td>15</td>
</tr>
<tr>
<td>Transcription (K)</td>
<td>10</td>
</tr>
<tr>
<td>Cell wall/membrane/envelope biogenesis (M)</td>
<td>55</td>
</tr>
<tr>
<td>Posttranslational modification, protein turnover, chaperones (O)</td>
<td>17</td>
</tr>
<tr>
<td>Inorganic ion transport and metabolism (P)</td>
<td>89</td>
</tr>
<tr>
<td>Secondary metabolites biosynthesis, transport and catabolism (Q)</td>
<td>36</td>
</tr>
<tr>
<td>General function prediction only (R)</td>
<td>66</td>
</tr>
<tr>
<td>Function unknown (S)</td>
<td>23</td>
</tr>
<tr>
<td>Signal transduction mechanisms (T)</td>
<td>3</td>
</tr>
<tr>
<td>Defense mechanisms (V)</td>
<td>4</td>
</tr>
</tbody>
</table>

The biomass equation can be, in a way, considered as a pseudo-reaction that is supplementary to the reconstruction, and serving as a computational representation to simulate cellular growth of an organism. It should, as sufficiently as possible, be inclusive of all the necessary building blocks of the biomass. Accordingly, there are certain indispensable elements that are almost ubiquitous to all prokaryotic and eukaryotic cells, mainly: nucleotides, amino acids, lipids, and cofactors. Given the laborious and demanding nature of experimental elucidation of all coefficients of the components in the biomass, it has now become common to utilize biomass
compositions from representative species with experimentally identified biomass components, mostly *E. coli* in Gram-negative bacteria and *B. subtilis* in Gram-positive bacteria. In this reconstruction, the general biomass composition of Gram–positive bacteria generated by ModelSeed was used.

Each component was checked for suitability in the biomass, and accordingly some elements were eliminated either because there is no genomic evidence for their biosynthesis in the genome or if these elements precursors were already accounted for in the biomass equation. Due to the absence of chemostat experiments for Bac48, the ATP growth associated maintenance requirement in *B. subtilis* 168 (105 mmol gDW⁻¹) was used in iPara1056.

4.2.3 Model Validation

To validate iPARA1056, experimental growth rate and growth phenotypes were compared to simulations of cellular growth in LB media. When the model was simulated on complete media, FBA showed that the first draft reconstruction showed a growth rate of 156.5207 h⁻¹, while the final model, after refinement, showed a growth rate of 9.4338 h⁻¹. The experimental cellular growth rate in rich LB medium (with glucose as the carbon source) was estimated to be 0.936 h⁻¹ (Figure 4.3) using log₂ of the curve of measured absorbance at OD600.
Figure 4.3. Experimental growth curve of *B. paralicheniformis* Bac48. The vertical axis is log₂ of absorbance at OD600 and horizontal axis is time in hours.

To simulate growth in rich LB media, the estimated constituents of commercial LB media in Oh Y.K. *et al.*, 2007 [318] (Table S1.1) were used. The experimental growth rate was obtained when the lower bound of each exchange reaction of these metabolites was constrained to -200 mmol/gDW/h while keeping the upper bound at 1000 mmol/gDW/h. This setting assures that metabolites can freely exit the system while limiting their uptake rate. The lower bound for all other exchange reactions, not estimated to be a component in commercial LB media, was constrained to zero. The uptake rate of oxygen was unlimited (lower bound constrained to -1000 mmol/gDW/h). Glucose was used as the main carbon source and its uptake rate was constrained to -18.5 mmol/gDW/h. Running FBA analysis with biomass as the objective function, yielded a growth rate of 0.9434 h⁻¹, a value that is comparable to the experimentally determined growth rate.

Phenotype MicroArrays (PMs) were used to identity carbon and nitrogen sources on which *B. paralicheniformis* Bac48 could grow on. To determine experimental growth/no-growth phenotypes, maximum curve heights above the
average of curve heights of all readings for a plate was considered as the threshold. Additionally, growth curves that don’t follow the typical bacterial growth curve (lag, log and stationary phases) were considered as no-growth phenotypes. For in silico simulations, cases where there is a positive increase in the optimized growth rate (when the tested substrate uptake rate is constrained to -1000 mm/gDW/h) were considered as positive in silico growth phenotypes. On the other hand, cases where there is no difference in the optimized cellular growth with and without the tested substrate were considered as no-growth in silico growth phenotypes. Accuracy was calculated based on the agreement between experimental and in silico growth phenotypes using true and false positive predictions (TPs and FPs), and true and false negative predictions (TNs and FNs).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]

Out of 190 substrates tested for *B. paralicheniformis* Bac48, 59 substrates have transport reactions in the model and accordingly they were used for testing (47 with positive growth phenotypes and 12 with negative growth phenotypes) (Table 4.2). Of the 47 metabolites that were experimentally utilized, 40 were correctly predicted (TP) and seven were incorrectly predicted (FN). Of the 12 substrates which the model could not utilize, 11 were correctly predicted (TN) and only one was incorrectly predicted (FP). The model was predicted to have an overall accuracy of 86.44%.

For 95 nitrogen sources, 44 had transport reactions and thus were tested (32 with positive growth phenotypes and 12 with negative growth phenotypes) (Table 4.3). Of the 32 positive growth phenotypes, 25 were correctly predicted as positive.
Of the 12 negative growth phenotypes, 11 were correctly predicted. This resulted in 25 TPs, seven FNs, 11 TNs and one FP and an overall accuracy of 81.8%.

As to incorrect false negative predictions, one of the factors could be that some of the substrates used in the Biolog microplates are not completely well-studied and thus adding corresponding reactions for their transportation was not possible. It is also possible that other elements in the media are affecting the strains ability to grow, resulting in no-growth phenotypes.

Table 4.2. Computational and experimental growth (G) and no growth (NG) phenotypes in different carbon sources. TP refers to true positive, TN refers true negative, FN refers false negative and FP refers to false positive.

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Experimental growth</th>
<th>In silico growth</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Glucose-6- Phosphate</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>m-Inositol</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Inosine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Threonine</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>D-Serine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>D-Galacturonic Acid</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Thymidine</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Maltotriose</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Adenosine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Arbutin</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Salicin</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>L-Histidine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Isoleucine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Leucine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Lysine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Methionine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Valine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>DL-Carnitine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>2L-Carnitine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>D-Ribose</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>D-Fructose</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Arabinose</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>N-Acetyl-D- Glucosamine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>D-Galactose</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Aspartic Acid</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Metabolite</td>
<td>Experimental growth</td>
<td>In silico growth</td>
<td>Classification</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------</td>
<td>------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>L-Histidine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Leucine</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Compound</td>
<td>G</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>------------------</td>
<td>----</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>L-Methionine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-Proline</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>L-Tryptophan</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>L-Tyrosine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>D-Serine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>D-Valine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Adenosine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Thymidine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Uracil</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Inosine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Xanthine</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Allantoin</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>L-Threonine</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Ala-Leu</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Ala-Thr</td>
<td>G</td>
<td>NG</td>
<td>FN</td>
</tr>
<tr>
<td>Ammonia</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Nitrite</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Nitrate</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Urea</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Alanine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Arginine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Aspartic Acid</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Glutamic Acid</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Glutamine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Glycine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Isoleucine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Serine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Valine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>D-Alanine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>L-Ornithine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Glucuronamide</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Cytidine</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Cytosine</td>
<td>NG</td>
<td>G</td>
<td>FP</td>
</tr>
<tr>
<td>Uridine</td>
<td>NG</td>
<td>NG</td>
<td>TN</td>
</tr>
<tr>
<td>Ala-Asp</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Ala-Gln</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Ala-Glu</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Gly-Asn</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Gly-Gln</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Gly-Glu</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Gly-Met</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>Met-Ala</td>
<td>G</td>
<td>G</td>
<td>TP</td>
</tr>
</tbody>
</table>
4.2.4 Gene Essentaility Analysis

One of the most notable applications of metabolic models is the prediction of essential genes. The essential set of genes for maintaining the biomass of \textit{B. paralicheniformis} Bac48 was predicted using the \textit{iPARA1056} model. This set will serve as a starting point for future genome minimization attempts made for engineering \textit{B. paralicheniformis} Bac48 and for the identification of gene knock-out targets that should be avoided in metabolic engineering experiments. \textit{In silico} simulations resulted in the identification of 145 essential genes in both minimal media and in LB media. This list was compared to the experimentally verified list in \textit{B. subtilis} 168. Genes in the model agreeing with experimentally verified essential genes in \textit{B. subtilis} 168 represented 66.8\% of the total number of genes predicted as essential in \textit{iPARA1056}. The remaining genes that were conflicting with experimentally determined essential genes in \textit{B. subtilis} 168 were individually checked for relevance to the strains metabolism. To do so, COG assignments of essential genes were used to see if the genes indeed fell into one of the categories that were necessary for growth maintenance, mainly ‘lipid transport and metabolism (I)’, ‘inorganic ion transport and metabolism (P)’, ‘cellular processes and signaling (M)’, ‘amino acid transport and metabolism (E)’, ‘coenzyme transport and metabolism (H)’, ‘information storage and processing (J)’, ‘energy production and conversion (C)’, and ‘nucleotide transport and metabolism (F)’. These false negative essentaility gene predictions could be due to specific metabolism patterns of the species or they could be the resultant of certain biomass components discrepancies between \textit{B. paralicheniformis} Bac48 and \textit{B. subtilis} 168.
4.2.5 Case Study: Genetic Targets for the Overproduction of Bacitracin Precursors

As case studies showing the usability of iPARA1056 predictions for the generation of \textit{in silico} metabolic engineering strategies, two precursors for the production of the bacitracin antibiotic were used as optimization targets. Specifically, FBA was used to identify engineering knock-out targets for growth-coupled production of the precursors of this secondary metabolite. Selecting bacitracin for the case study was encouraged by the presence of the complete bacitracin operon in the genome of \textit{B. paralicheniformis} Bac48 as shown in Chapter 3.

Bacitracin was first studied in the first type strain of \textit{B. licheniformis}, strain 10716, where its production reached the highest yield with the use of augmented volumes of DL-aspartic acid as a substrate in a medium with manganese [319]. Moreover, manganese was identified as a positive modulator of other secondary metabolites [320]. The use of this medium composition increased bacitracin yield from 18.5 units/ml to 25 units/ml, and cellular growth from 980 µg/ml to 1140 µg/ml. Production of bacitracin in \textit{B. licheniformis} strains was previously found to be affected by other media components reported in [319, 321-323].

Although secondary metabolites were not accounted for in metabolic models, it is still possible to model the effect of \textit{in silico} perturbations on primary metabolites acting as precursors for their biosynthesis. Two amino acids were identified as precursors for bacitracin peptide: l-isoleucine and l-cysteine. Since l-cysteine is produced directly from l-serine, the production of o-acetyl l-serine was used as the optimized reaction. For the overproduction of isoleucine, its precursor, l-aspartate was used as the end target.
Surveying the literature for experimentally verified genetic targets for the overproduction of these two precursors, showed little, if any usable results for testing in *Bacillus* species. However, a Gram-positive l-serine overproducing mutant strain of *Corynebacterium glutamicum*, a species commonly used for amino acid overproduction, was obtained through the deletion of alanine transaminase (*alaT*) resulting in the increased production of l-serine (26.23 g/L) and increased productivity (0.27 g/L/h) [324]. In *iPARA1056*, testing the deletion of *alaT* showed an increase in flux (174.9 fold increase) while maintaining biomass.

Additionally, the *in silico* predicted deletions for the overproduction of l-isoleucine in *E. coli* were also tested in *iPARA1056*. For the overproduction of aspartate - which acts as a precursor for l-isoleucine - the highest predicted production in *E. coli* was obtained with the suggested knock-out of 2-ketoglutaratedehydrogenase (rnx08094), and phosphotransacetylase (rnx00173) [325]. In the model, only phosphotransacetylase gene knock-out resulted in a growth-coupled production of l-isoleucine up to 89.3 folds when simulated under rich LB media.

One final test was conducted to identify gene knock-out targets predicted by *iPARA1056* that are not experimentally verified. Predicted metabolic engineering strategies were eliminated based on two evaluation conditions: If the strategy includes knocking-out a gene that is predicted and experimentally validated as an essential gene (true positive) or only found to be experimentally essential (false negative), and if the strategy includes knocking-out genes involved in transport reactions.

The model predicted that removing dihydroxy-acid dehydratase (EC 4.2.1.9) that catalyzes the conversion of 2,3-dihydroxy-isovalerate to 3-methyl-2-
oxobutanoate (rxn00898) result in 289.1 fold increase in L-serine production, while knocking-out the reaction catalyzing the conversion of acetyl-CoA to o-acetyl-L-serine (rxn00423), would result in 9.82% increase in flux for isoleucine production. These reactions are part of the L-valine biosynthesis pathway and the L-serine biosynthesis pathway, which are competing pathways with L-serine and L-isoleucine respectively.

The gene knock-out predictions made by iPARA1056 in this case study stress the applicability of the model to generate strategies for the over-production of other metabolites as needed. Accordingly, the model is presented as a testing platform for metabolic engineering strategies, adding to the aforementioned advantages of B. paralicheniformis Bac48 as an MCF.
Chapter 5

5 Concluding Remarks

Microbial inhabitants of the Red Sea niche are expected to be robust, efficient and capable of producing thermostable products, all of which are desired characteristic of MCFs. In spite of these promising potentials, very few studies have evaluated the biotechnological value of microbial systems from this environment. With the availability of ‘omics’ data, the efficient utilization of in silico tools to evaluate the metabolic capabilities of an ecosystem is now a promising path. Here, Bacilli strains, isolated from two lagoons by the Red Sea, are investigated as potentially attractive sources of novel antibiotics, due to their broad spectrum antimicrobial activities, and they are evaluated as efficient MCFs.

Reconstructed metabolic networks of 32 species (including the ten Red Sea strains) ranged from 671 to 1,398 reactions, with a median of 1,291 reactions (Chapter 2). Interestingly, the three networks with the highest number of metabolic reactions correspond to Red Sea species. The study also shows that virgibacillus species have statistically more reactions involved in the biosynthesis of the osmoprotectant ectoine and in the degradation of glucuronoarabinoxylan, indicating salt resistance properties and hydrolysis of plant cell walls respectively. Clustering metabolic reactions from the reconstructions revealed that three clades (two of which have the Red Sea strains B. foraminis Bac44, B. litoralis Bac94 and B. marisflavi Bac144) do not correspond to clades generated from phylogenetic clustering. Additionally, Red Sea strains B. foraminis Bac44, B. litoralis Bac94 and B. marisflavi Bac144, shared only two BGCs with other species. The results from the evolutionary study indicates convergent evolution of the metabolism of these species as well as the acquisition of
specialized secondary metabolism in order to adapt to the special conditions of the Red Sea environment. Future studies should aim at the detailed characterization of the bioactive compounds secreted by the strains and the identification of their role in the environment.

Given that *B. paralicheniformis* Bac48 had the highest protein secretion capability compared to other transformable Red Sea strains, the strain was selected for further evaluation (Chapter 3). Phylogenetic analysis showed that another Red Sea strain (Bac84), along with Bac48, fell within the *B. paralicheniformis* group. This motivated the inclusion of *B. paralicheniformis* Bac84 in the study. In the largest non-syntenic block in *B. paralicheniformis* Bac48, that extends over 83 Kb of DNA, resides a trans-AT PKS/NRPS cluster. This type of cluster had been mostly associated with antitumor and antimicrobial products in past examinations, suggesting a similar function, which sufficiently adds to the biosynthetic value of the strains and warrants future characterization and isolation of the product of the cluster in future studies.

The aforementioned advantages of *B. paralicheniformis* strains (Chapter 3), specifically of the Red Sea strain *B. paralicheniformis* Bac48 (Chapter 2 and 3), motivated the reconstruction of a genome-scale metabolic model to theoretically explore its metabolism (Chapter 4). First, the strain growth rate was simulated in rich LB media with glucose as the main carbon source, yielding a rate of 0.9434 h⁻¹, a value that is comparable to the experimentally-determined growth rate. Additionally, in silico simulations resulted in the identification of 145 essential genes in both minimal media and in LB media. Future studies on *B. paralicheniformis* should aim at utilizing the model for the generation of in silico strain design methods for the overproduction of different metabolic targets as needed.
The results from the described paradigm in this research, provides an excellent example of one of the advantages of using Systems biology approaches for the evaluation of newly sequenced genomes. Particularly, it shows how computational analysis assists in reducing the high cost and laborious tasks of relying exclusively on *in vitro* experiments for the evaluation of the biotechnological capabilities of microbial strains.
REFERENCES


Wang, Y., et al., *Complete Genome Sequence of Bacillus paralicheniformis MDJK30, a Plant Growth-Promoting Rhizobacterium with Antifungal Activity.* Genome Announc, 2017. 5(25).

Ma, J., et al., *Complete Genome Sequence of Bacillus subtilis GQJK2, a Plant Growth-Promoting Rhizobacterium with Antifungal Activity.* Genome Announc, 2017. 5(22).


110


112


231. May, J.J., T.M. Wendrich, and M.A. Marahiel, The dhh operon of bacillus subtilisEncodes the biosynthetic template for the catecholic siderophore 2,


119


**SUPPLEMENTARY**

**Supplementary 1. Composition of Simulated Growth Media**

**Table S1.1. Chemical composition of complex, LB (Luria-Bertani) medium**

<table>
<thead>
<tr>
<th></th>
<th>Yeast extract Bacto™ a</th>
<th>In silico</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Glucose</strong></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>Amino acids</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Histidine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Leucine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Methionine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Proline</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Threonine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Tryosine</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Arginine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Cystine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glycine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Lysine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Serine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Valine</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>Vitamins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiamine (B1)</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Riboflavin (B2)</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Calcium pantothenate (B5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyridoxine (B6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biotin (B8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Folic acid (B9)</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Cyanocobalamin (B12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Niacin (PP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other chemicals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sodium</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Chloride</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Sulfate</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Potassium</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Phosphorus (phosphate)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Calcium</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Magnesium</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Selenium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zinc</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Arsenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cadmium</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Mercury</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Lead</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nucleotides/nucleosides</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inosine</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
Hypoxanthine +
Deoxycytidine +
Thymidine +
Uracil +
Uridine +
Deoxyadenosine +
Adenosine +
Other Chroimate +
Fe3 +
O2 +
Cu2 +
Fe2 +
H+ +
Chorismate +
Mn2+ +

Based on [318]

Supplementary 2. Biolog Phenotype Data

S2.1. M9 Media Composition

M9 minimal media was used to test for growth before the Biolog test. For the carbon source negative control, glucose was excluded from the media and for nitrogen source negative control, NH4Cl was excluded from M9 media salt.

Table S2.1 Composition of M9 media used for Biolog experiments

<table>
<thead>
<tr>
<th>Salt for positive control, sterilization by autoclaving</th>
<th>1000ml H2O</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64g Na2HPO4-7H2O</td>
</tr>
<tr>
<td></td>
<td>15g KH2PO4</td>
</tr>
<tr>
<td></td>
<td>2.5g NaCl</td>
</tr>
<tr>
<td></td>
<td>5.0g NH4Cl</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M9 media complete</th>
<th>1000 ml of sterile distilled H2O</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200ml of M9 salts</td>
</tr>
<tr>
<td></td>
<td>2ml of 1M of sterile MgSO4</td>
</tr>
<tr>
<td></td>
<td>20 ml of 20% filtrated glucose</td>
</tr>
<tr>
<td></td>
<td>100µl of sterile 1M CaCl2</td>
</tr>
</tbody>
</table>
S2.2. Growth Phenotypes Images

Figure S2.1 Carbon sources utilization

Figure S2.2. Nitrogen sources utilization
Figure S2.3 Peptide nitrogen sources utilization

Figure S2.4. Growth under different osmolytes and pH levels
Supplementary 3. Urease Christensen Urea Agar Media

Figure S3.1. Change in color in media inoculated with *B. paralicheniformis* Bac48 (red), compared to the control (yellow)