Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 2591918	Student Name: Dongliang Li
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): MSc Civil Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of **First** Supervisor: Umer Zeeshan Ijaz	Name of **Second** Supervisor: Ciara Keating
Title of Project: Whole genomes analysis of self-healing concrete bacteria Bacillus Subtilis	
Declaration of Originality and Submission Information	
I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements Signed (Student) : Dongliang Li	E N G 5 0 5 9 P
Date of Submission : August 20, 2021	

Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator	
Grade Awarded: Feedback (as appropriate to the coursework which was assessed):	
Lecturer/Demonstrator:	Date returned to the Teaching Office:

Whole genomes analysis of self-healing concrete bacteria Bacillus Subtilis

MSc Civil Engineering

Dongliang Li (Student ID: 2591918)

Supervisor: Dr Umer Zeeshan Ijaz

Co-Supervisor: Dr Ciara Keating

August 20, 2021

A report submitted in partial fulfillment of the requirements
for MSc Civil Engineering Degree
at the University of Glasgow

Acknowledgements

First, I want to thank Drs Umer Zeeshan Ijaz and Ciara Keating, my two supervisors, and I would like to express my heartfelt thanks and respect to my two supervisors for guiding, helping, and supervising us wholeheartedly during those time when they were very busy. Especially thanks for meeting us twice a week for guiding our project, which is very helpful for my own work on the project. I could successfully complete my project on time thanks to their hard coaching and aid. Secondly, I would like to thank our university, the University of Glasgow, for providing us with the learning platform and guidance as well as the opportunity to exercise and improve ourselves. During this tough and difficult time affected by the COVID-19, our university had made great efforts to ensure that our daily study life is not greatly affected, which is also the key reason for us to complete our tasks successfully. In addition, I would like to my schoolmates who are in the same project group, they also gave me lots of help and advice which are very beneficial for me to understand my own project work during the research process in the project. Last but not least, I want to thank my family. It is because of your funding and silent support that I could be able to study at the University of Glasgow and do my postgraduate study smoothly and give me power and confidence to live alone in a foreign country, your care and support are always the driving force for me to keep moving forward.

Abstract

Concrete is a composite material made up of fine and coarse aggregates that are held together by a fluid cement that hardens over time. Concrete is the most widely used building material in the world because it is strong, long-lasting, and relatively affordable. Because the aggregate efficiently supports the compression load, concrete is strong in compression. In tension, however, it is vulnerable because the cement that holds the aggregate in place might crack, causing the structure to break, and the presence of cracks may reduce the strength and durability of concrete. With the increasing maintenance cost of concrete and the pollution of the environment caused by the production of concrete, the idea of concrete self-healing has gradually been put forward and studied by more and more people.

The capacity of the concrete to repair cracks spontaneously or autonomously is commonly termed as self-healing concrete. It is also known as self-reparation concrete. The self-repair of concrete can prolong the service life of concrete structures, making concrete materials more durable and sustainable.

There are many methods for concrete self-healing. In fact, concrete self-healing is an ancient phenomenon because concrete has natural self-healing properties. Because of continuous clinker mineral hydration or calcium hydroxide ($\mathtt{Ca(OH)_2}$) carbonation, cracks will heal after a while. However, natural healing is limited to small cracks and can only be effective if there is sufficient moisture, so it is difficult to control artificially. Nevertheless, there are many ways to make concrete self-healing, such as chemical method, physical method, mechanical method, electrodeposition self-healing method, microbial self-healing method. In this thesis I will investigate the potential of the self-healing method of concrete based on microorganisms, and the Bacillus subtilis will be introduced and studied in this project as this microorganism.

Bacillus subtilis, is a gram-positive, rod-shaped bacterium found in soil and ruminant and human gastrointestinal tracts. *Bacillus* species can produce a robust, protective endospore, allowing them to withstand harsh environmental conditions.

This project aims to analyze the whole genomes of *Bacillus subtilis* strains through bioinformatic analysis that will obtain and annotate all the genes contained in the genome sequence of *Bacillus subtilis* isolates from public repositories (NCBI) and focusing on finding the key pathways related to producing molecules useful for healing concrete. For example, genes that play a key role in the calcium precipitation pathway will be introduced in detail, and their role in the self-healing process of concrete will be studied, analyzed, and evaluated.

Key Words: self-healing concrete, *Bacillus subtilis*, whole genomes, microbial self-healing.

Contents

1 Introduction..1
1.1 Background...1
1.2 Self-healing concrete...2
1.3 Biological(microbial) self-healing concrete...3
1.3.1 Precipitation of calcium carbonate for biological self-healing concrete.4
1.3.2 Precipitation of polymorphic iron-aluminum-silicate for biological self-healing concrete... 5
1.4 Bacillus subtilis...5
1.5 Related research on microbial self-healing concrete based on Bacillus subtilis 7
1.6 Aims and objectives.. 10
2 Methods..11
2.1 Preliminary preparation and primary work of the project..........................11
2.1.1 The selection of the self-healing concrete bacteria............................ 11
2.1.2 Download the genomes of the bacteria.. 12
2.2 Primary and further processing of the project..12
2.2.1 Prokka command line software..13
2.2.2 Roary software... 14
2.2.3 Metabolic software..17
2.3 Analysis and visualization of some results of the project.......................... 21
2.3.1 Coinfinder software..21
3 Results..23
3.1 Roary results...23
3.2 METABOLIC results.. 26
3.3 Coinfinder results..32
4 Discussion.. 35
5 Conclusion..37
Reference..38
Appendix..43

1 Introduction

1.1 Background

As it is strong, durable, and relatively inexpensive, concrete is the most used construction material worldwide (Jonkers, 2007). However, the brittle nature of concrete makes it prone to cracks, and the presence of cracks may reduce the strength and durability of concrete. Because of the inherent heterogeneity of concrete, micro cracks are unavoidable, and drying shrinkage, heat stress, weathering, externally imposed stresses, and reinforcing corrosion are all possible causes of cracks, if micro fractures create a continuous network, they can greatly reduce the resistance of the concrete to hostile chemicals, hence enhancing the permeability of the cement. Inorganic erosion: acid, salt, a strong alkaline and a concrete composition, a no gel effect or expansion material generated, changes the composition of the concrete structure leading to concrete corrosion; The second type of substance is organic and corroded by the micro-organism, which decomposes organic matter in the suitable environment, releases corrosive substances such as organic acid, carbon dioxide, hydrogen sulphide and degrades the concrete. And the aggressive substances permeate inside the concrete and corrode the reinforcement, which will reduce the durability of the concrete. Therefore, it is very important to fill them with more concrete.

In fact, concrete has natural self-healing properties, because the cracks will mend over time with continual clinker mineral hydration or calcium hydroxide ($\mathrm{Ca(OH)_2}$) carbonation. However, natural self-healing properties is limited to small cracks and only under sufficient water conditions can it have natural self-healing properties. In addition, during the producing of the concrete, the mixing of concrete, the mixing of cement, sand and fine stones will produce dust, which will pollute the environment when discharged into the air, and with the rising costs associated with repairs, people are considering alternatives of cracks healing. The self-healing is one approach that researchers are most interested in. Dry(Dry, 1994) started to work on the autonomous self-healing concrete in 1994. In the years after that, other researchers began examining this problem. There are several methods of self-healing. Han et al.(2014) mentioned that most include autogenous self-healing technique, self-healing method on the basis of capsules, vascular way of self-healing, self-healing method for electrical deposition, microbial self-healing method, and self-healing methods using shaped memory alloys (SMAs). For example, Jonkers et al. (Jonkers et al., 2010)investigated the ability of bacteria to act as a self-healing agent in concrete, that is, their ability to fix cracks that arise. They proved that application of bacterial spores as self-healing agent appears promising (Jonkers et al., 2010).

With the increasing amount of concrete worldwide (more and more building usage), the maintenance of concrete and the strength and sustainability of concrete materials must be more and more valued by engineers and related researchers. Microbial (bacterial) self-healing concrete, as a new methodology to enhance the strength of concrete, the sustainability of concrete use, and the reduction of maintenance costs is becoming promising research filed.

1.2 Self-healing concrete

Self-healing concrete is typically characterized as the ability of concrete to heal cracks on its own. Ghosh(2008) mentioned that self-healing concrete was inspired by natural phenomena exhibited by species such as plants and animals. Tree and animal skin that has been damaged can be restored on their own (Talaiekhozan et al., 2014). The ability of concrete to self-repair can extend the life of concrete structures, making them more durable and sustainable. Huang and Kaewunruen(2020) emphasized that calcium carbonate is generated by the method of healing cracks and then calcium carbonate is filled with fissures. During self-healing treatments, there are two ways to create calcium carbonate. The first is the utilization of unreacted cement particles to initiate hydration and the formation of $\mathrm{CaCO_3}$. The second is that $\mathrm{CaCO_3}$ is generated when $\mathrm{Ca(OH)_2}$ is dissolved (Dhir, R. K and Jones, M. R., 1999). Nijland et al.(2007) mentioned that the equation below shows that the method of forming calcium carbonate in water with different pH values is different:

When the PH value of water is between 7.5 to 8: $\mathrm{Ca^{2+} + HCO_3^- \leftrightarrow CaCO_3 + H^+}$

When the PH value of water is higher than 8: $\mathrm{Ca^{2+} + CO_3^{2-} \leftrightarrow CaCO_3}$

Previous studies have identified several elements that may influence self-healing ability. Huang and Kaewunruen (2020) mentioned that the following are the five most important factors:

1. Moisture content: samples with sufficient water content will heal more quickly.
2. Hydration time: longer periods of hydration can result in improved self-healing abilities (Teall, n.d.).
3. Pressure on cracks: applying the right amount of pressure to cracks can help them mend faster.
4. Crack width: Zhong and Yao (2008) mentioned that cracks that are less than 0.3 mm wide can be totally repaired. It is possible that cracks larger than 0.3 mm will not heal. Cracks of width 0.1 mm are completely healed after around 200 hours (Huang and Kaewunruen, 2020). Furthermore, Edvardsen, C.K(1996) emphasized that within 30 days, cracks with a width of 0.2 and 0.3 mm are mostly healed. Ahn and Kishi(2010) mentioned that Cracks with a width of 0.15 to 0.3 mm shrink

significantly in 7 days and heal entirely in 33 days.

5. Water-cement ratio: More unreacted cement particles are present in a greater water–cement ratio, which can be utilized for additional hydration to increase calcium carbonate production.

In addition, the time of cracking is also important. Van Tittelboom and De Belie (2013) emphasized because earlier cracking concrete contains more unreacted cement particles, it has a higher self-healing ability when hydration is maintained.

Self-healing concrete has many significant features and advantages, such as less pollution, lower costs, environmental friendliness, and improved toughness in difficult situations are all advantages. All these characteristics make self-healing concrete an important sustainable material in the construction engineering industry.

The self-healing methods of self-healing concrete include not only natural self-healing as mentioned above, but also chemical self-healing and biological self-healing. This paper focuses on biological self-healing concrete.

1.3 Biological(microbial) self-healing concrete

Siddique and Chahal(2011) and Jonkers(2007) and Wu et al.(2012) have classified the employment of microorganisms in the creation of self-healing concrete as a biological technique. Microbes can grow in almost any environment, and Jonkers et al.(2010) and Van Tittelboom et al.(2010) have proposed using microorganisms to create self-healing concrete. Bacteria, fungi, and viruses are the three main classifications of microorganisms. For many bacteria, the pH value, moisture content and temperature in concrete are not suitable for their growth and reproduction. Although there are a lot of ways to protect these microorganisms from bad conditions, the high complexity of the methods and the limitations of the technology make these methods unable to be applied at present.

This project aims to find a certain bacterium which is relatively suitable for living in the environment of concrete and relevant to the precipitation of specific compounds. Special bacteria strains capable of precipitating specific compounds are among the microorganisms utilized to create biological self-healing concrete. In most instances, the precipitation of polymorphic iron-aluminum-silicate ($(Fe_5AI_3)(SiAl)O_{10}(OH)_5$) and calcium carbonate ($CaCO_3$) are the two most important ways and processes which are used for designing and producing biological self-healing concrete (Talaiekhozan et al., 2014).

1.3.1 Precipitation of calcium carbonate for biological self-healing concrete

Calcium carbonate is a compound with the chemical formula $CaCO_3$. Mineral crystals of calcium carbonate come in three different forms. Talaiekhozan et al.(2014) mentioned that the chemicals they use are similar but have various structures such as calcite, aragonite and vaterite. Calcite is calcium carbonate in its most stable form. Aragonite is metastable, meaning it may change into calcite over time. Sanchez-Moral et al.(2003) mentioned that vaterite is a mineral that is extremely uncommon in nature. Therefore, calcite is more interesting for researchers working on related self-healing concrete.

There are many ways to produce calcium carbonate precipitation. For example, Berner (1975) emphasized that magnesium can increase precipitation of calcite. Biological mineral precipitation can be aided by a few microbial metabolic processes. For example, Castanier et al.(2000) emphasized that anaerobic sulphide oxidation, ammonification, denitrification and photosynthesis are the main processes to produce calcium carbonate precipitation. The schematic diagram of the sedimentation of calcium carbonate microorganisms into concrete cracks is shown in the figure below:

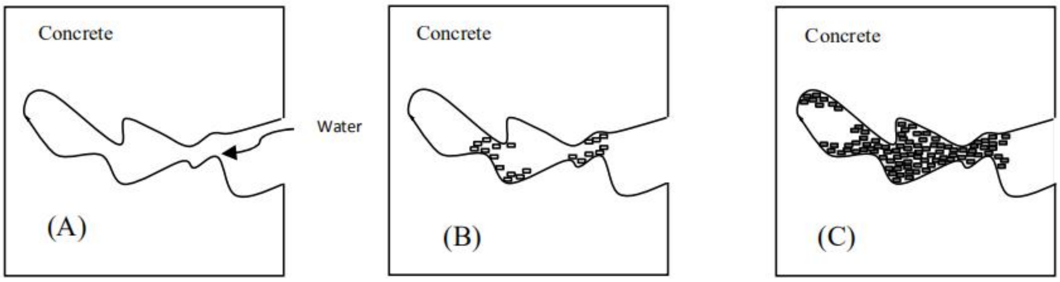


Figure 1: Schematic diagram of microbial repair of concrete cracks (Talaiekhozan et al., 2014)

(A) The break in the concrete is spreading.

(B) The microorganisms in the cracks are activated.

(C) Microorganisms can fill the fissure by growing and precipitating calcium carbonate around their cell walls.

1.3.2 Precipitation of polymorphic iron-aluminum-silicate for biological self-healing concrete

The silica accumulating bacteria have been reported in several studies (Ferris et al., 1988). Jonkers et al.(2010) emphasized that A complex iron-aluminum silicate has been found on the surface of isolated bacterial cells in a lake polluted with metal particles.

Through some related experiments and studies, it is shown that the chemical formula of polycrystalline iron aluminum silicate is $(\mathrm{Fe_5AI_3})(\mathrm{SiAl})\mathrm{O_{10}}(\mathrm{OH})_5$. At acidic pH, the bacterium *Leuconostoc mesenteroides* is crucial for silica precipitation. This bacterium uses glucose to make lactic acid, resulting in an acidic environment where the colloidal silica's reduced solubility causes precipitation. Talaiekhozan et al.(2014) mentioned that this approach is not common as acidic and does not work well for concrete durability.

1.4 Bacillus subtilis

Härtig and Jahn(2012) mentioned that as one of the most characteristic bacteria, *Bacillus subtilis* is utilized as a model organism for gram-positive bacteria. Members of this species also show a lot of genetic variation, according to recent microarray-based comparative genomic investigations (Earl et al., 2008). *Bacillus subtilis* is a rod-shaped bacterium, and it was able to grow in many environments, and the endospores it produces can survive extreme environmental conditions including high temperature and dryness. Kovács(2019) mentioned that the sleeping spores can endure tough conditions (high temperature, drying, ultraviolet and α-radiation), bacteria or macroorganisms, or even alien settings. Kovács(2019) also emphasized that bacillus subtilis has become the most studied Bacillus species due to its inherent ability to ingest extracellular DNA, which allows for easy genetic alteration and the occurrence of sporulation, one of the first known bacterial cell development processes.

Härtig and Jahn(2012) mentioned that in history, *Bacillus subtilis* was once thought to be a strictly aerobic creature. A first indication for the utilization of nitrate as an alternative electron acceptor under microaerophilic growth conditions was obtained 40 years ago (Michel et al., 1970). However, it took 25 years for Glaser et al.(1995) and Hoffmann et al.(1995) to prove that *Bacillus subtilis* has anaerobic nitrate respiration and cloned the corresponding nitrate reductase gene. Nakano et al.(1997) and Cruz Ramos et al.(2000) then clarified the various fermentation processes that maintain anaerobic growth.

Bacillus subtilis may, for a varying number of uses such as the production of Enzymes and food and plant-biocontrol, be isolated from a wide range of environments, including soil and marine habitats.

Bacillus subtilis is a model microorganism that has been used to study cell division, protein secretion, surface movement (swimming, swarming, and sliding), biofilm development, attachment of plant roots or fungal hyphae, production of secondary metabolites, and passage cytoplasmic exchange of nanotubes between cells, release of extracellular vesicles, and kinship discrimination (Kovács, 2019).

1.5 Related research on microbial self-healing concrete based on Bacillus subtilis

According to several related literatures and resources, for the studies relevant to self-healing concrete based on microorganism, which are utilizing the bacterium bacillus subtilis to produce calcium carbonate precipitation to repair cracks in concrete. Those research contents mainly include: the related research on self-healing concrete, the action mechanism and mode of microorganisms in self-healing concrete, the mechanism of *Bacillus subtilis* in the precipitation of calcium carbonate in self-healing concrete, and in-depth research on related genes or enzymes of self-healing concrete bacteria.

For the aspect of the related research on self-healing concrete, several relevant literature and research have shown that self-healing concrete is typically characterized as the ability of concrete to heal cracks on its own. Ghosh(2008) stated that natural phenomena shown by species such as plants and animals inspired self-healing concrete. Tree and animal skin that has been damaged can be restored on their own (Talaiekhozan et al., 2014). Li and Yang(2007) mentioned that certain fissures in old concrete structures are bordered by white crystalline materials which show the potential of concrete to self-sticking the cracks using chemical products, presumably by rainfall and carbon dioxide in air. Later, a number of researchers Dhir, R. K and Jones, M. R.(1999) and Reinhardt and Jooss (2003) noticed that a steady decrease of permeability over time was observed in the examination of water flow through cracked cement below the hydraulic gradient, which again suggested a capacity to self-stick cracked concrete and to slower water flow rates.

The mechanism of healing cracks is to produce calcium carbonate and then cracks can be filled with calcium carbonate (Huang and Kaewunruen, 2020). The first is the utilization of unreacted cement particles to initiate hydration and the formation of $CaCO_3$. The second is that $CaCO_3$ is generated when $Ca(OH)_2$ is dissolved (Dhir, R. K and Jones, M. R., 1999). The capacity for crack-healing in most common types of concrete, however, appears to be limited to micro-cracks, i.e., cracks with widths up to 0.1–0.2 mm (Li and Yang, 2007; Dhir, R. K and Jones, M. R., 1999). Therefore, more and more researchers are beginning to study other artificial methods to strengthen the self-healing ability of concrete, such as chemical self-healing and biological self-healing.

For the aspect of some related research on the bacteria based self-healing concrete, in 1995, Gollapudi et al. (1995) were the first to present this new technology of fracture fixation utilising environmental biological processes. Van Tittelboom et al.(2010)

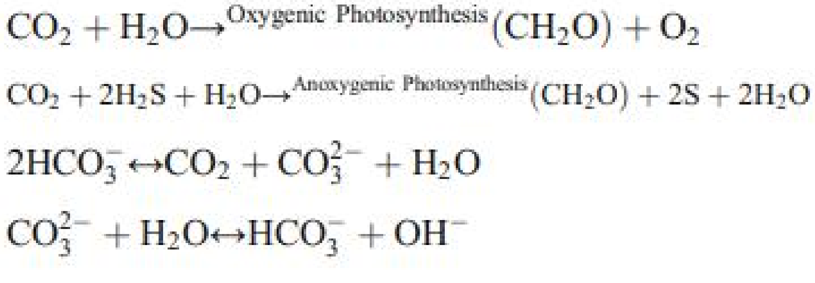
proposed that a number of parameters determines the $CaCO_3$ microbiological precipitation. These factors include: the dissolved inorganic carbon content; pH; calcium ion concentration and the existence of nuclear nuclear sites. Hammes and Verstraete*(2002) mentioned that the metabolism of bacteria provides the first three variables while the bacteria cell wall acts as a nucleation location. Seifan et al.(2016) mentioned that in the presence of a calcium supply, calcium carbonate can be precipitated through a physiologically driven mineralization process. Carbonate is created extracellularly by bacteria through two metabolic pathways, autotrophic and heterotrophic in this process. For the autotrophic pathway, in the presence of carbon dioxide, the autotrophic process occurs, in which bacteria convert carbon dioxide to carbonate in three different ways, namely (a) non-methylotrophic methanogenesis (by methanogenic archaea), (b) oxygenic photosynthesis (by Cyanobacteria) and (c) anoxygenic photosynthesis (by purple bacteria) (Castanier et al., 1999). Non-methylotrophic methanogenesis pathway converts carbon dioxide and hydrogen to methane (Eq. 1). As a result, as shown in Eq. 2, anaerobic oxidation of methane by electron acceptors like sulphate produces bicarbonate (Castanier et al., 2000). The generated carbonate precipitates as calcium carbonate in the presence of calcium ions, as described in Eq. 3.

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O \qquad (1)$$

$$CH_4 + SO_4^{2-} \rightarrow HCO_3^- + H_2O + HS^- \qquad (2)$$

$$Ca^{2+} + 2HCO_3^- \leftrightarrow CaCO_3 + H_2O + CO_2 \qquad (3)$$

In the presence of calcium ions, photosynthesis is also an autotrophic mechanism for producing calcium carbonate. Photosynthetic bacteria are divided into two groups: oxygenic and anoxygenic photosynthetic bacteria. The following is a schematic diagram of photosynthetic chemical processes for calcium carbonate formation.



$$CO_2 + H_2O \rightarrow^{\text{Oxygenic Photosynthesis}} (CH_2O) + O_2$$

$$CO_2 + 2H_2S + H_2O \rightarrow^{\text{Anoxygenic Photosynthesis}} (CH_2O) + 2S + 2H_2O$$

$$2HCO_3^- \leftrightarrow CO_2 + CO_3^{2-} + H_2O$$

$$CO_3^{2-} + H_2O \leftrightarrow HCO_3^- + OH^-$$

Figure 2: photosynthesis chemical reactions for calcium carbonate production (Seifan et al., 2016)

For the heterotrophic pathway, Seifan et al.(2016) proposed that because of their expansion in many natural settings, microbial communities may precipitate crystals. Crystal formation is a natural phenomenon that is attributable to the medium composition used to culture heterotrophic bacteria. Seifan et al.(2016) emphasized that the heterotrophic development in organic acid sales species (acetate, lactate, citrate, succinate, oxalate, malate and glyoxylate) of diverse bacterial genera, including *Bacillus*, *Arthrobacter* and *Rhodococcus*, leads to carbonate mineral production. These bacteria use organic compounds as energy sources.

In the presence of calcium acetate as a source of low molecular weight acid and calcium ion, the chemical processes to produce calcium carbonate are outlined below: (v. Knorre and Krumbein, 2000)

$$CH_3COO^- + 2O_2 \rightarrow^{\text{Heterotrophic bacteria}} 2CO_2 + H_2O + OH^-$$

$$2CO_2 + OH^- \rightarrow CO_2 + HCO_3^-$$

$$2HCO_3^- + Ca^{2+} \rightarrow CaCO_3 + CO_2 + H_2O$$

Figure 3: Chemical responses in the presence of calcium acetate as a source of calcium carbonate

Other methods for making calcium carbonate include the sulfur cycle and the nitrogen cycle. Sulphur cycle follows by dissimilatory reduction of sulphate (Seifan et al., 2016).

Production of calcium carbonate through reducing calcium sulphate ($CaSO_4$) to calcium sulfide (CaS) by sulphate reducing bacteria is shown in the equations below: (Ehrlich, 1995)

$$CaSO_4 + 2(CH_2O) \rightarrow CaS + 2CO_2 + 2H_2O$$

$$CaS + 2H_2O \rightarrow Ca(OH)_2 + H_2S$$

$$CO_2 + H_2O \rightarrow H_2CO_3$$

$$\mathrm{Ca(OH)_2} + H_2CO_3 \rightarrow CaCO_3 + 2H_2O$$

Figure 4: The chemical process of producing calcium carbonate by reducing calcium sulfate

As for the nitrogen cycle, it can produce carbonate or bicarbonate through three major pathways: (a) urea or uric acid breakdown (ureolysis), (b) amino acid ammonification, and (c) dissimilatory nitrate reduction (Castanier et al., 2000; Perito and Mastromei, 2011).

For the aspect of the mechanism of Bacillus subtilis in the precipitation of calcium carbonate in self-healing concrete and related genes or enzymes of self-healing concrete bacteria. Barabesi et al.(2007) used technical means to separate five genes (lcfA, ysiA, ysiB, etfA, and etfB) from Bacillus subtilis to observe their inducing effects on calcium carbonate precipitation. After different experiments, they found that the lack of activity observed in the absence of an inducer clearly suggests that the ysiA, ysiB, and etfB products are insufficient to support calcite precipitation. Through the result, Barabesi et al.(2007) mentioned that for the precipitation phenotype, the last gene, etfA, is crucial.

Castro-Alonso et al.(2019) discussed microbiological and molecular ideas and their relevance in bio-concrete Microbial Induced Calcium Carbonate Precipitation (MICP). MICP is the result of metabolic interactions between various microbial populations and organic and inorganic substances in the environment. Castro-Alonso et al.(2019) mentioned that the main metabolic activity of MICP on a variety of levels includes hydrolysis, denitrification, dissimilar sulphate reduction and photosynthesis. Castro-Alonso et al.(2019) also said that urea hydrolysis is best employed in concrete restoration systems and MICP is produced by urea hydrolysis by a sequence of urease (Ur) and Carbon anhydrase reactions (CA).

1.6 Aims and objectives

As mentioned above, *Bacillus subtilis* has great potential for repairing concrete cracks due to its ability to precipitate calcium carbonate. Therefore, this project aims to focus on finding the genes and enzymes that relates to the key pathways in calcium carbonate precipitation by using the Prokka, Roary and Metabolic software analysis methods; and finding out the specific reasons and chemical processes of these genes and enzymes that can produce calcium carbonate precipitation.

2 Methods

Dr Umer Zeeshan Ijaz and Ciara Keating from the University of Glasgow will support all data gathering, analysis and processing for this research. They provide me some relevant training and guidance to help me independently accomplish the further relevant research data analysis and processing. In this project, there are several programming software been used to analyze and process a lot of genomes data. The project research process can be divided primarily into various stages, described below:
1. The selection, gene data downloading and storage of the self-healing concrete bacteria.
2. Using some relevant software and algorithms to do the primary research data processing, and the results obtained by some preliminary algorithms are used for further data analysis and collection.
3. The analysis and visualization of the results data by using some relevant software and algorithms.

All the associated software and algorithms are provided by Dr Umer Zeeshan Ijaz (http://userweb.eng.gla.ac.uk/umer.ijaz/) at the University of Glasgow. And using the code ftp ftp.ncbi.nlm.nih.gov in MobaXterm to download the genomes of the selected bacteria. And some relevant code tutorials from (https://github.com/).

2.1 Preliminary preparation and primary work of the project

2.1.1 The selection of the self-healing concrete bacteria

Because this project is to study the whole gene analysis of self-healing concrete bacteria, the first step is to find some relevant bacteria according to some related literature about self-healing concrete. Using some keywords like “self-healing concrete” or “bacteria for healing cracks in concrete” to find and go through some related literatures. By adding those keywords into related academic search engines (Google Scholar https://scholar.google.com/ and online library of the University of

Glasgow https://www.gla.ac.uk/myglasgow/library/). Finding more than 10 bacteria related to the self-healing concrete, but most of them have little research value because of their few genomes. And the final choice requires a comprehensive comparison of the remaining bacteria.

In the end, for this project, the *Bacillus subtilis* is selected as self-healing concrete bacteria to do research, because according to some research literature, this bacterium has many advantages over other bacteria. Firstly, Härtig and Jahn(2012) mentioned that *Bacillus subtilis* is a bar-shaped bacterium that generates endospores that allow harsh conditions such as heat and desiccation to survive. In addition, McKenney et al.(2013) mentioned that *Bacillus subtilis* is able to split symmetrically to produce a single endospore in two daughter cells (binary fission) that can remain alive for decades and survive harmful environmental conditions such as dehydration, salt, severe pH, radiation and solvent*s*. According to another research, Feng et al.(Feng et al., 2021) mentioned that bacillus subtilis has been discovered to be suitable for the design and preparation of self-healing concrete with a crack width of 0.3 mm and a flexural strength regain capacity of 14 percent. Therefore, I selected *Bacillus subtilis* as the self-healing concrete bacteria for thus project.

2.1.2 Download the genomes of the bacteria

After choosing the bacteria, the genomes data should be downloaded from the NCBI (National Center for Biotechnology Information) database (ftp.ncbi.nlm.nih.gov). In this project, the MobaXterm software platform (https://mobaxterm.mobatek.net/) needs to be used for genomes downloading and further genomes analysis and visualization. By searching *Bacillus subtilis* in the NCBI database and downl oading its genomes, there are 477 genomes had been downloaded in the end, which will be the sample for my further research for this project. Those genomes will be analyzed by like Prokka, Roary, Metabolic and Coinfinder software for further research in this project.
By login to the Orion cluster with the platform MobaXterm and these genomes are saved in the RAW_ genomes folder.

2.2 Primary and further processing of the project

After all the genomes of the *Bacillus subtilis* have been downloaded from the NCBI and stored in the folder as mentioned above. Then several algorithms and software will used for the following analysis processing, they are Prokka software, Roary software and Metabolic software.

2.2.1 Prokka command line software

Prokka is a programming that annotates genes and identifies coding sequences in prokaryotic genomes quickly. It is suitable for annotating de novo bacterial assemblies, but not for human genomes (or any other eukaryote).

Seemann(2014) mentioned that Prokka is a Unix command line application that combines a series of current software tools to generate rich and reliable genome bacterial sequence annotations. On a quad-core desktop computer, it can annotate a normal bacterial genome in around 10 minutes using as many processing cores as possible. It’s ideal for putting iterative sequence analysis models and workflows into genome software.

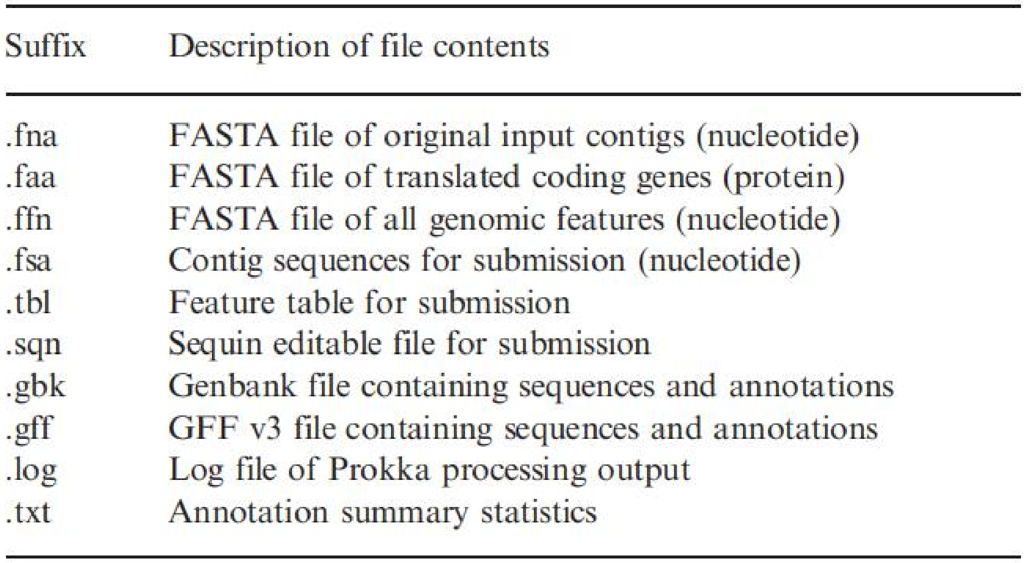
For prokka, how does it work. Prokka identifies and annotates sequence characteristics (both protein coding areas and RNA genes, such as tRNA and rRNA). Note that Prokka employs a two-step procedure to annotate protein coding areas: first, Prodigal is used to identify protein coding regions on the genome; second, the encoded protein's function is inferred based on similarities to proteins in one of many proteins or protein domain databases. Prokka is a software tool for fast annotating bacterial, archaeal, and viral genomes, with standard output files in GenBank, EMBL, and gff formats.

For the input files for prokka, Seemann(2014) mentioned that it expects FASTA-formatted preassembled genomic DNA sequences. Although finished sequences with no gaps are desirable, a collection of scaffold sequences generated by de novo assembly software is predicted to be the most common input. The software's sole required parameter is the sequence file.

Genes that code for proteins are annotated in two phases. The coordinates of potential genes are identified by Prodigal, but the putative gene product is not described. Traditional methods for establishing what a gene codes for include comparing it to a large database of known sequences, usually at the protein sequence level, and then transferring the annotation of the most significant match.

Seemann(2014) mentioned that Prokka employs this strategy in a tiered fashion, beginning with a smaller, more reliable database, progressing to medium-sized but domain-specific databases, and finally to curated models of protein families.

Prokka creates ten files in the output directory, each with the same prefix. The detailed description of the prokka output files as below:



Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

Figure 5: Display of Prokka output file name(Seemann, 2014)

The .gff files are very important, because they are the key files that make the further processing work. On multicore machines, prokka employs parallel processing to reduce the amount of time it takes to execute. Seemann(2014) BLAST+ and hmmscan, which both support multiple CPUs natively, are the most time-consuming phases. However, Prokka is more efficient when it uses GNU parallel to execute several single CPU threads on subsets of the data. (Tange, n.d.).

2.2.2 Roary software

Sitto and Battistuzzi (2020) mentioned that Roary is a Linux-native application that may be installed in a variety of ways on Linux, MacOSX, and Windows machines. Roary is a stand-alone pan genome pipeline that generates the pan genome using annotated assemblies in GFF3 format (generated by Prokka (Seemann, 2014)). It can analyze datasets with thousands of samples using a basic desktop PC, which is computationally impossible with present approaches, without affecting the quality of the results. With 1 GB of RAM and a single processor, 128 samples may be analyzed in less than an hour. Using conventional methods, this study would take weeks and hundreds of GB of RAM to complete. Roary isn't designed for meta-genomics or comparing large groups of genomes.

Because Roary's input files are in GFF3 format (General Feature Format version 3). This format contains a sequence of data in a certain order that must be precisely followed in order for Roary to accept the input file (see https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md;) (Sitto and Battistuzzi, 2020). GFF3 files can be obtained in one of two ways: through the NCBI website or by converting .fna files to GFF3 files using the software Prokka. In this project, the GFF3 files are be obtained by converting .fna files from the software Prokka as mentioned above. And before running Roary commands, all the .gff files will be moved into a folder named roary which has been created by the user, then activating the Conda environment by using code [source activate pangenome] every time using the terminal window for the first time. Then after running for about 20 hours for the 477 genomes in this project, the output files of Roary have been gained.

For the interpretation of the output files of Roary, normally a run of Roary will produce 17 output files, and for all the output files, the summary_statistics.txt and the gene_presence_absen-ce.csv are the most important. The summary statistics text file lists the overall number of genes in the pangenome, as well as the number of genes in each of four categories (core, soft, shell, and cloud). Sitto and Battistuzzi (2020) emphasized that these numbers accurately depict the pangenome of the species under consideration .

Table 2: The output file summary_statistics.txt of Roary

Core genes	(99% <= strains <= 100%)	596
Soft core genes	(95% <= strains < 99%)	1450
Shell genes	(15% <= strains < 95%)	2935
Cloud genes	(0% <= strains < 15%)	43943
Total genes	(0% <= strains <= 100%)	48924

As for the gene_presence_absen-ce.csv file, it contains additional data, such as the individual gene IDs of sequences that fall into each of the summary statistic's categories (although it is not mentioned directly, the ratio of the number of genes contained in each cluster to the total number of genomes studied can be estimated).
In addition, 3 figures are obtained after a run of Roary. The first one is illustrating the tree in comparison to a matrix with core and auxiliary genes present and absent. The second is a pie chart depicting the distribution of genes and the number of isolates in which they are found. The last graph shows the frequency of genes in relation to the number of genomes. The 3 figures are shown below:

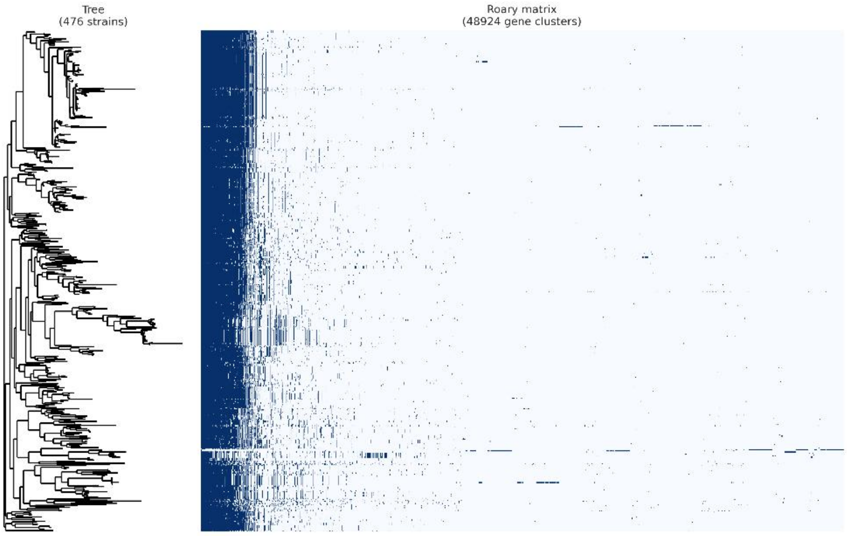


Figure 6: Pan genome matrix

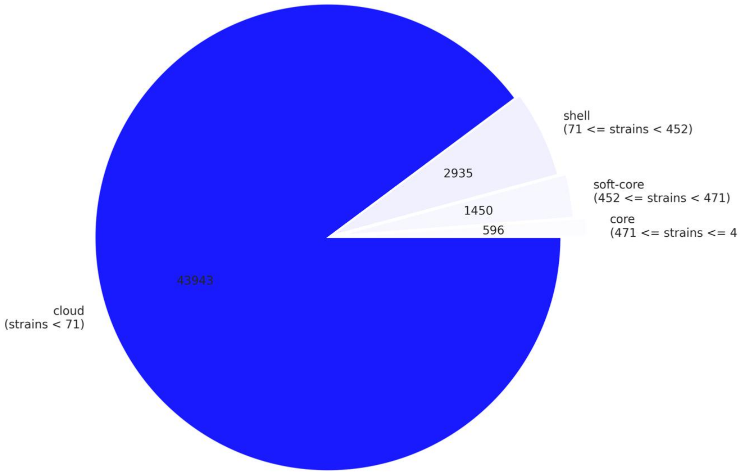


Figure 7: Pie chart of the number of genes in each of four categories

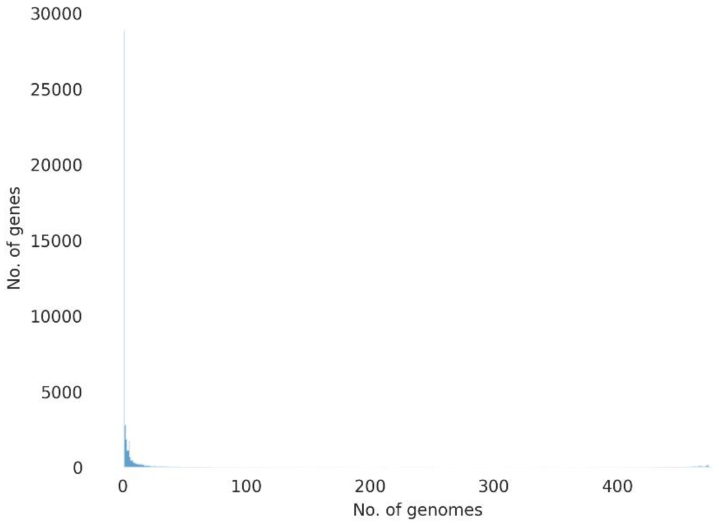


Figure 8: The frequency of genes versus the number of genomes

Page et al.(2015) mentioned that the Roary publication's additional material contains a thorough description of all of Roary's output files, in a less detailed manner, on the github page (https://sanger-pathogens.github.io/Roary/).

2.2.3 Metabolic software

METABOLIC (Metabolic and Biogeochemistry Analysis in Microbes) is a scalable programming that uses genomes at the level of individual organisms and/or microbial communities to enhance microbial ecology and biogeochemistry. The genome-scale approach includes annotation of microbial genomes, motif validation of biochemically proven conserved protein residues, metabolic indicator identification, metabolic pathway analyses, and contribution estimations to specific biogeochemical transformations and cycles (Zhou et al., 2019). Kanehisa (2002)mentioned that the METABOLIC is also a set of tools for analysing metabolic and biogeochemical functional characteristics in microbial genomes. METABOLIC combines KEGG annotation with METABOLIC, TIGRfam (Selengut et al., 2007), Pfam (Finn et al., 2014), and custom hidden Markov model (HMM) databases (Anantharaman et al., 2016), contains a motif validation phase for reliably identifying proteins based on biochemical validation, KEGG modules are used to assess the existence or lack of metabolic pathways, and creates user-friendly outputs such as tables and graphs, for individual genomes and at the community level, including an overview of functional profiles, biogeochemically significant pathways, and metabolic networks.

Detailed application and introduction of METABOLIC by other scholars
METABOLIC is a Perl and R script that should run on Unix, Linux, or MacOSX. On Metabolic GitHub website (https://github.com/AnantharamanLab/METABOLIC), the prerequisites are listed. Microbial genome sequences in FASTA format, as well as a collection of genomic or the metagenomic reads that can be used to reassemble those genomes if desired (Figure 4), are required in the input folder.

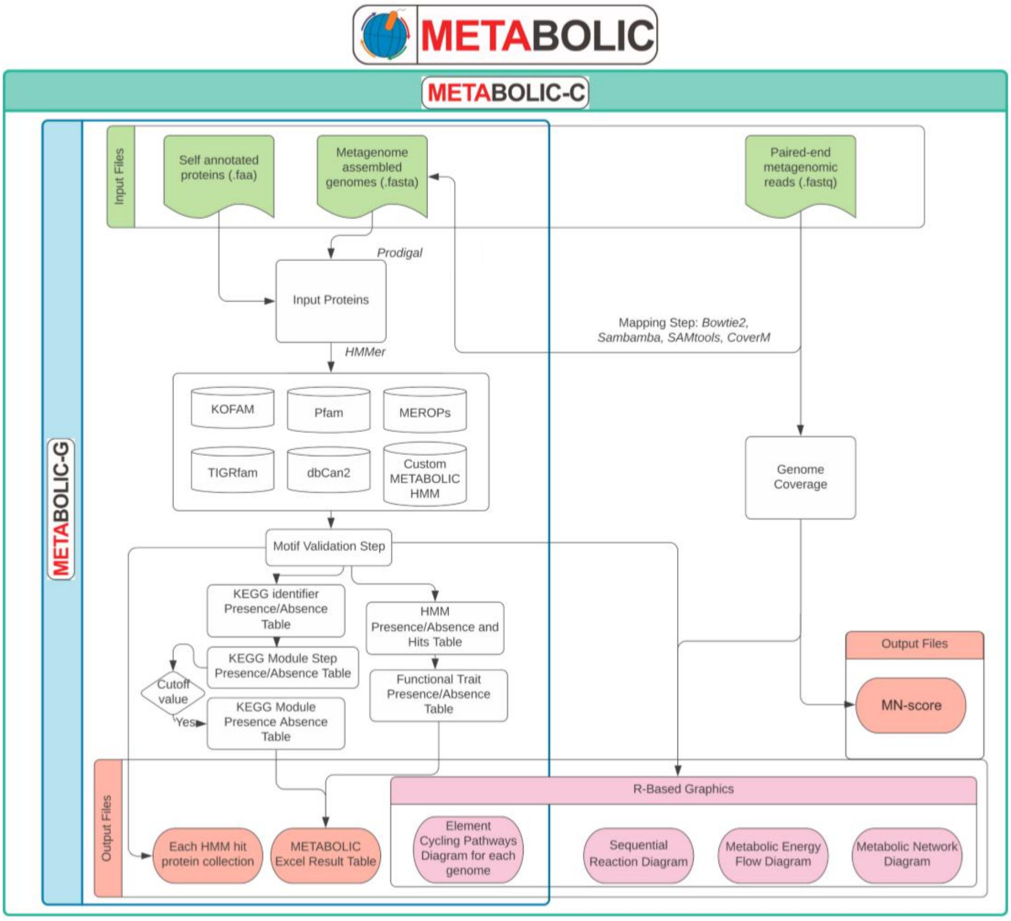

METABOLIC
METABOLIC-C
METABOLIC-G
Input Files
Self annotated proteins (.faa)
Metagenome assembled genomes (.fasta)
Paired-end metagenomic reads (.fastq)
Prodigal
Input Proteins
HMMer
Mapping Step: Bowtie2, Sambamba, SAMtools, CoverM
KOFAM
Pfam
MEROPs
TIGRfam
dbCan2
Custom METABOLIC HMM
Genome Coverage
Motif Validation Step
KEGG identifier Presence/Absence Table
HMM Presence/Absence and Hits Table
KEGG Module Step Presence/Absence Table
Cutoff value
Yes
KEGG Module Presence Absence Table
Functional Trait Presence/Absence Table
Output Files
MN-score
R-Based Graphics
Output Files
Each HMM hit protein collection
METABOLIC Excel Result Table
Element Cycling Pathways Diagram for each genome
Sequential Reaction Diagram
Metabolic Energy Flow Diagram
Metabolic Network Diagram

Figure 9: An overview of METABOLIC's workflow (From https://github.com/AnantharamanLab/METABOLIC)

Prodigal (Hyatt et al., 2010) annotates genomic sequences, or a user can submit self-annotated proteins (with extensions of '.faa') to make integration into current processes easier. We've also supplied an optional Perl script that may be used to extract gene and protein sequences from input genomes using Prodigal-generated '.gff' files. These files are used in the mapping of genomic/metagenomic reads in the downstream phases.

Finn et al.(2011) using the HMMER-implemented hmmsearch, and HMM databases are used to search for proteins (KEGG KOfam, Pfam, TIGRfam, and custom HMMs), which uses techniques for the most sensitive and efficient detection of remote homologues. METABOLIC checks the principal outputs after the hmmsearch stage with a motif-checking process for a subset of protein families; only those protein hits that pass this step are considered significant hits (Zhou et al., 2019).

To infer the presence of certain metabolic pathways in microbial genomes, METABOLIC depends on matches to the following databases. For a better understanding of metabolic pathways, in the context of KEGG modules, individual KEGG annotations are inferred. Each step in a KEGG module represents a different metabolic activity. For the project of KEGG annotation into the metabolic network, which is designed to better reflect the metabolic planes of the input genomes by means of separate components - modules, Zhou et al.(Zhou et al., 2019) parsed the KEGG module database (Muto et al., 2013) to link the existing relationship of KO identifiers to KEGG module identifiers. In most case, for KEGG module assignments, Zhou et al.(Zhou et al., 2019) used KOfam HMM profiles. Zhou et al.(2019) also used the TIGRfam/Pfam/custom HMM profiles and motif-validation processes for a collection of key metabolic marker proteins and often misannotated proteins. When targeting the same set of proteins, the software offers adjustable options for raising or lowering the priority of database. This is primarily intended to improve annotation confidence by preferring bespoke HMM databases over KEGG KOfam.

Zhou et al. (Zhou et al., 2019) offer a means to measure metabolism completeness since metagenomes and single-cell genomes can usually include incomplete metabolism (or a module here). The completeness of a particular module is estimated using a user-defined threshold (default cut-off is a 75 percent metabolic step/gene presence inside a particular module), the KEGG module presence/absence table is then utilised to produce. All modules that pass the cutoff are considered complete. In the process, the presence/absence data is summarised in an overall output table for each module phase to facilitate further extensive study.

Normally for the output files of METABOLIC, 6 different results which are reported in excel and lots of pdf files. For the six distinct findings which are provided in an excel spreadsheet as outputs. Zhou et al.(2019) mentioned that these contain details of protein hits which include both existence and absence and protein names, the presence and absence of functionality features and the presence or absence of KEGG modules.

For the pdf files of community-scale biogeochemical cycling processes, which are a summary scheme of community cycling activities at biogeochemical scale for each genome in the bacterium. Each arrow is depicted as a single transformation/step in a cycle. Each arrow is labelled (from top to bottom): step number, reaction, number of genomes capable of conducting these reactions, metagenomic coverage of genomes capable of conducting these reactions (expressed as a percentage within the community). Like the example below:

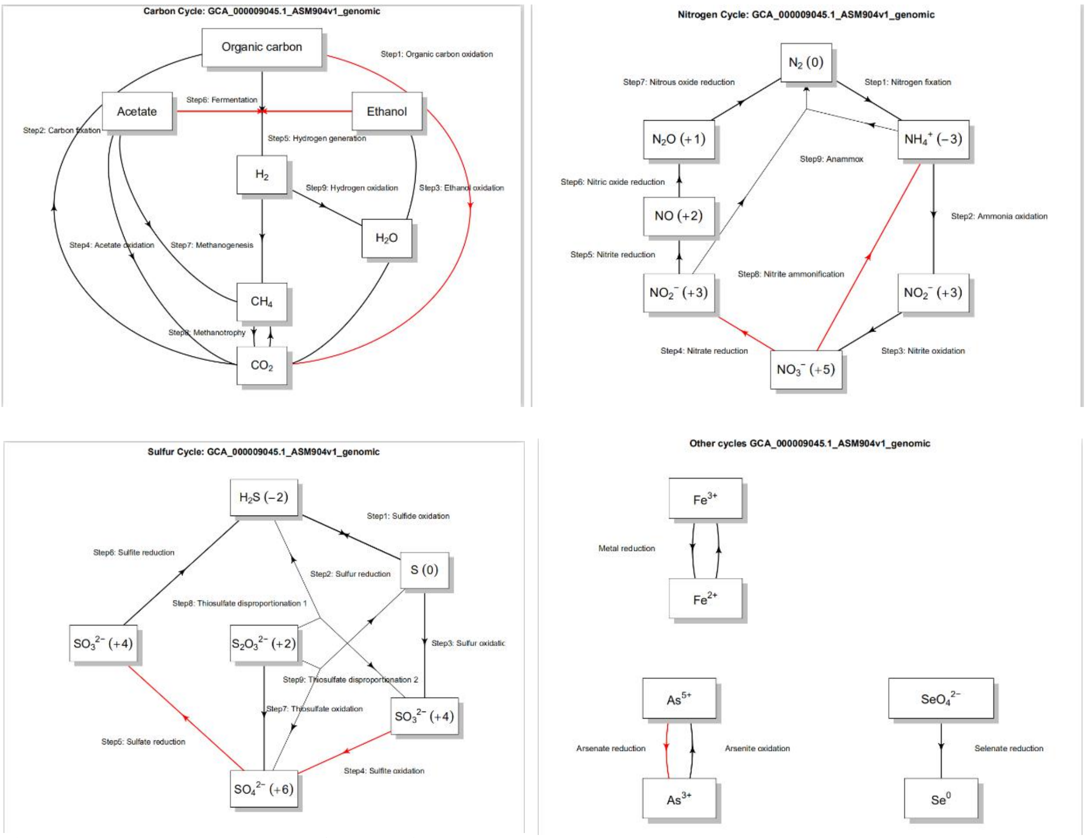


Figure 10: Process of Carbon, Nitrogen,
Sulfur, and other cycle for one certain genome

All my workflow/commands and the screenshot of the process of PROKKA, ROARY and METABOLIC will be showed in the part of appendix.

2.3 Analysis and visualization of some results of the project

As for this part work of the project, the software Coinfinder will be used for the further analysis and visualization of some results which are obtained by the previous process.

2.3.1 Coinfinder software

Coinfinder is a command-line tool that finds genes that are related (associating or dissociating) across a group of genomes. Coinfinder may operate in any Unix environment with any number of processor cores provided by the user. Whelan et al.(2020) emphasized that Coinfinder is not restricted to genomic input of prokaryote or eukaryote and may be utilised for the construction of pangenomes of strains or species.

For the input files of Coinfinder, Coinfinder accepts data on genetic content in two formats: (a) the gene_presence_absence.csv output from Roary (Page et al., 2015); or (b) as a tab-delimited list of the genes present in each strain (Whelan et al., 2020). If option (b) is chosen, Whelan et al.(Whelan et al., 2020) mentioned that prior to utilizing Coinfinder, genes should be grouped into orthologous groups/gene clusters (for example, using blast (AltschuP et al., n.d.) and a clustering algorithm, such as MCL (van Dongen, n.d.). Furthermore, Yarza et al.(2008) mentioned that Coinfinder requires a Newick-formatted phylogeny of genomes in the data set that is proposed that it is built with the core genes of input genomes, which are made using programmes like Roary or Ribosomal RNA or a similar manner.

For the output of Coinfinder, it will visualize its analysis results in two ways. For the first one, Coinfinder establishes a network with each node reflecting the families and edges of a genetic link or major gene dissociation statement (adjusted for lineages). The size of a node is proportional to the gene's D value (Whelan et al., 2020). For the second one, Coinfinder creates a heatmap highlighting the presence or absence of coincident genes in the context of the supplied phylogeny. The genes in the heatmap are colored according to coincident patterns and sorted by D value (from most lineage-independent to least). Examples of the output files are shown below:

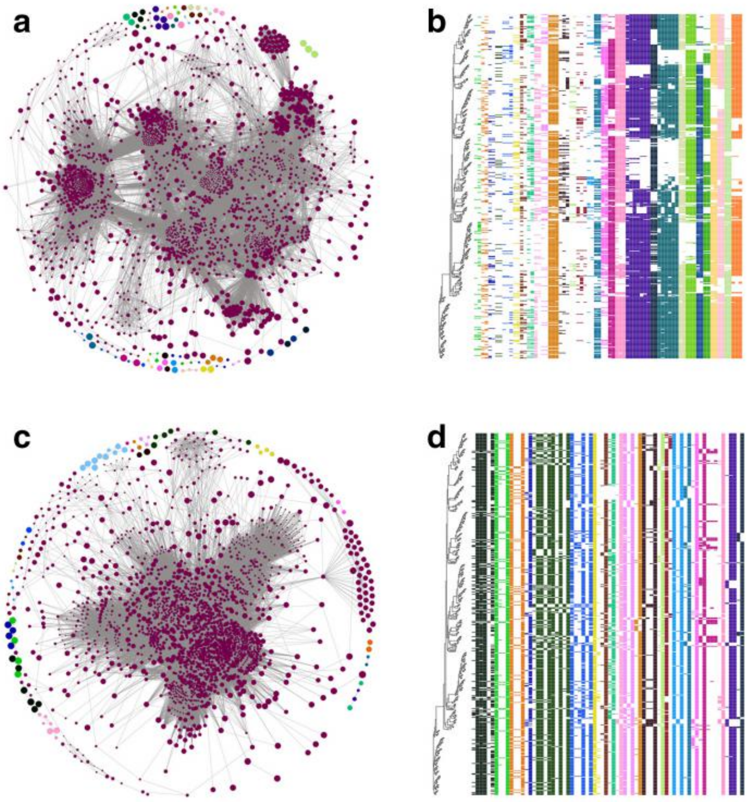


Figure 11: Example of the output files of Coinfinder (Whelan et al., 2020)

3 Results

Above mentioned some of the results will be listed in this section, the results figure included some genes and genomes information of the bacterium *Bacillus subtilis*, the information about the relationship between different genes and between different genes and different genomes, this project is aimed at by analyzing and summarizing these results information to find out and analyze some genes and enzymes of *Bacillus subtilis* which play a key role in some important chemical processes in self-healing concrete.

3.1 Roary results

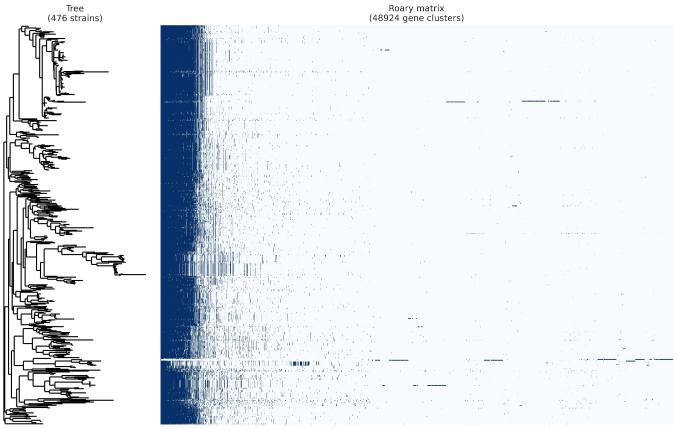


Figure 12: Pan genome matrix

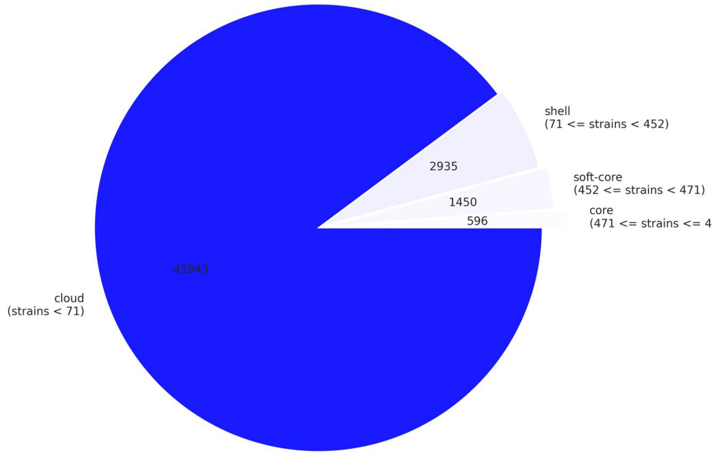


Figure 13: Pie chart of the number of genes in each of four categories

Table 3: Statistical table of the number of four types of genes

Core genes	(99% <= strains <= 100%)	596
Soft core genes	(95% <= strains < 99%)	1450
Shell genes	(15% <= strains < 95%)	2935
Cloud genes	(0% <= strains < 15%)	43943
Total genes	(0% <= strains <= 100%)	48924

As shown above, from figure 11 it shows that there are 476 genomes and approximately 48924 genes in the bacterium *Bacillus subtilis*. From figure 12 and table 3, the proportions and specific numbers of the four types of genes are clearly shown. As it shows that the proportion of the core genes whose strains are between 99% and 100% is at a lower amount, the number of which is only 596 and accounting for approximately one percent of the total. The number of the soft-core genes whose strains are between 95% and 99% is 1450 and accounting for approximately three percent of the total. The number of cloud genes is relatively the largest among the four types of genes, the number of which is 43943 and accounting for nearly 90% of the total.

One of the core tasks of this paper is to find out the core genes that play a role in the process of calcium carbonate precipitation in self-healing concrete, and through some other results later combined with relevant literature to complete the search for key genes.

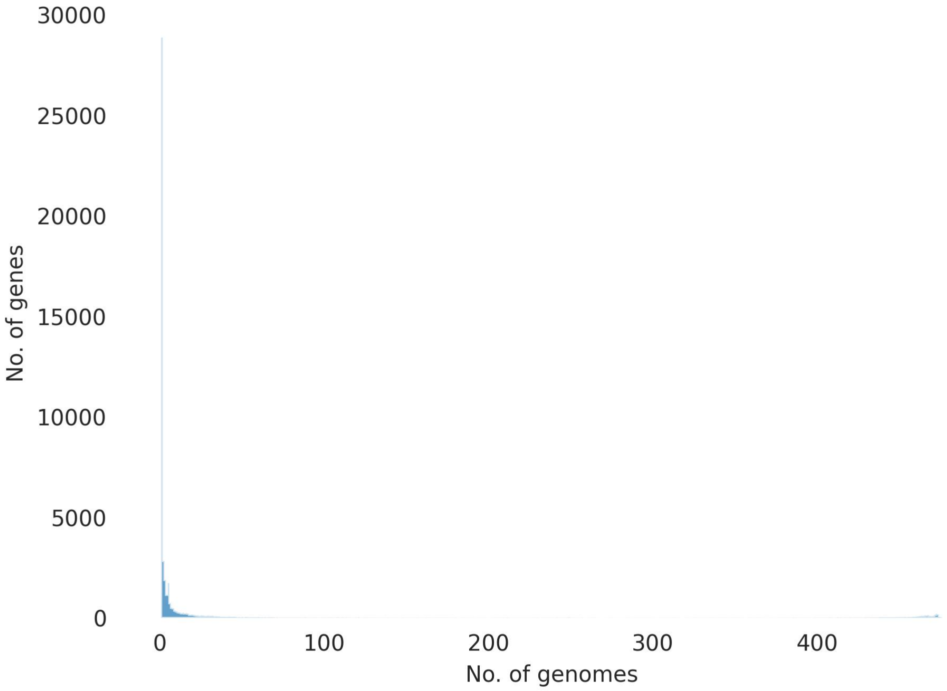


Figure 14: The frequency of the number of genes versus the number of genomes

As shown in the figure 13, the number of genes in which genomes and the relative proportion of genes contained in which genomes are clearly displayed. As what the histogram above shows, there are a lot of genes roughly in No.1 to No.20 genomes, and there are also some genes approximately in the last 10 sets of genomes of the whole 476 genomes. Especially in the first few sets of genomes there are many genes, the number of genes in several sets of genomes is roughly close to 3000, and the number of genes in the other sets of genomes is also between 500 to 1000. From these results, it can be concluded that the frequency of genes appearing in the genome is not high. For all 476 genomes, there are few genomes containing a huge number of genes, and only a few genomes contain a huge number of genes.

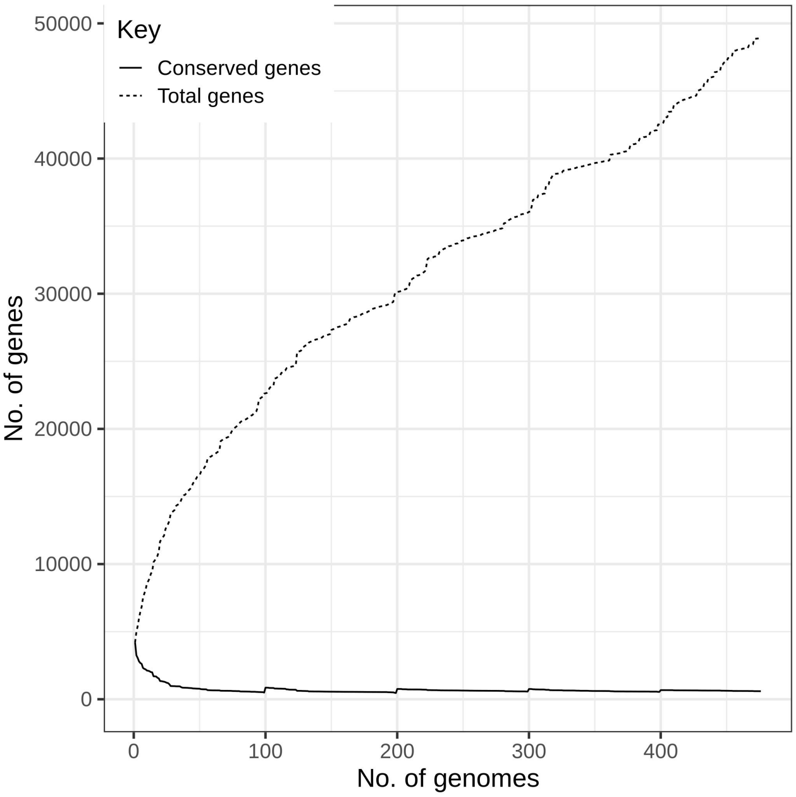
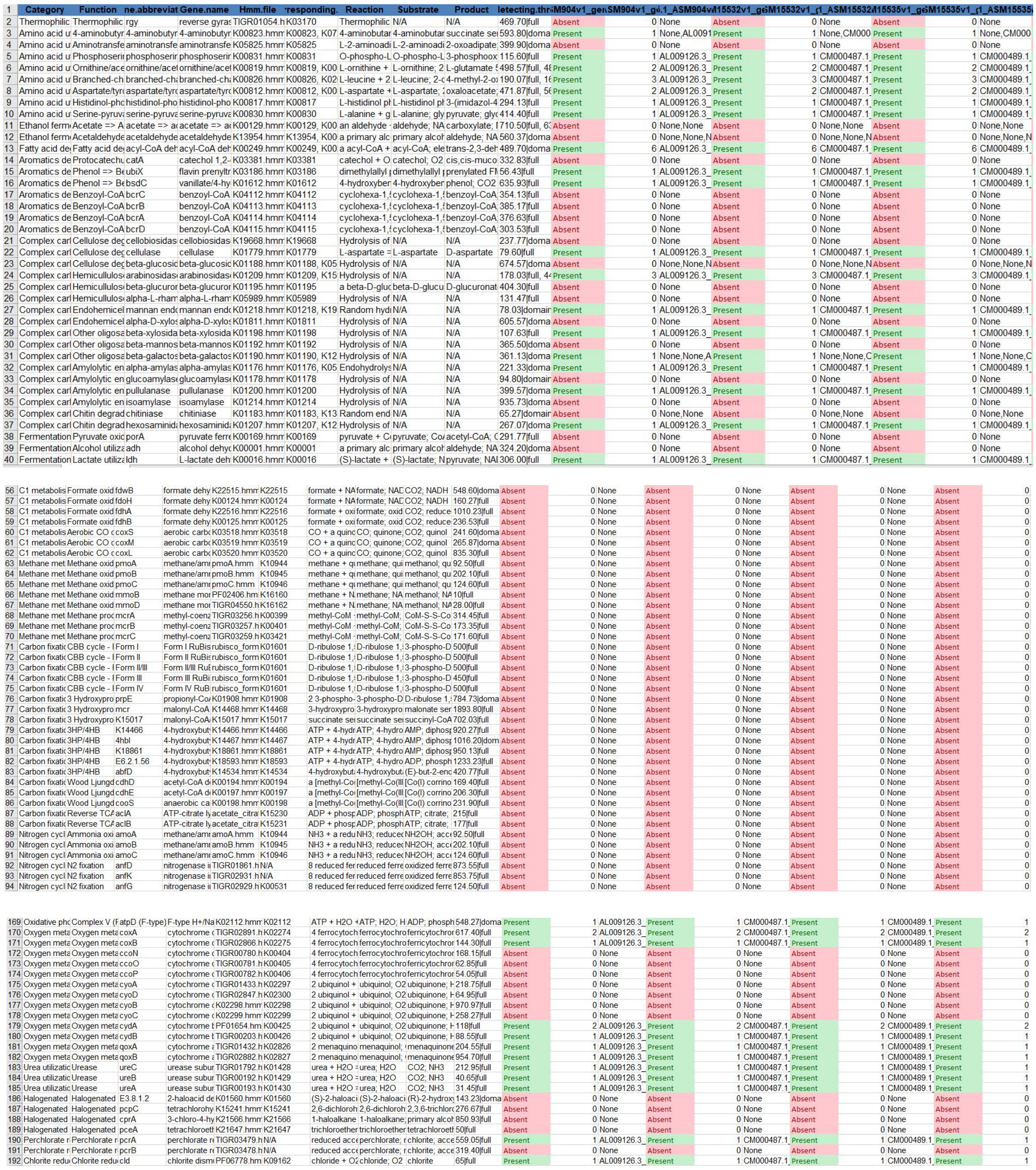


Figure 15: The line chart of the conserved genes vs total genes

As shown in Figure 14, the chart shows that in the first few sets of genomes, conservative genes occupy a large number in the genomes; with the further execution of genome statistics, the total number of genes keeps increasing; however, the number of conservative genes remain at a stable value and close to 0, indicating that conservative genes exist in a few specific genomes, in most genomes do not contain conservative genes.

3.2 METABOLIC results

As for the output files of the METABOLIC software, there are two important kinds of files worthy of attention. They are 'gene _ presence _ absence of METABOLIC result' and 'Nutrient _ Cycling _ Diagrams' files. The first one is showing which genes are present or not present in which genomes, and the second one shows which specific genome is responsible for which specific steps in the carbon cycle, sulfur cycle and nitrogen cycle.



1	Category	Function	ne.abbreviat	Gene.name	Hmm.file	rresponding.	Reaction	Substrate	Product	letecting.thr	M904v1_gen	SM904v1_g	.1_ASM904v	l15532v1_g	M15532v1_	1_ASM15532	l15535v1_g	M15535v1_	1_ASM15535
2	Thermophilic	Thermophilic	rgy	reverse gyras	TIGR01054.h	K03170	Thermophilic	N/A	N/A	469.70|full	Absent	0	None	Absent	0	None	Absent	0	None
3	Amino acid u	4-aminobutyr	4-aminobutyr	4-aminobutyr	K00823.hmm	K00823, K07	4-aminobutar	4-aminobutar	succinate se	593.80|doma	Present	1	None,AL0091	Present	1	None,CM000	Present	1	None,CM000
4	Amino acid u	Aminotransfe	aminotransfe	aminotransfe	K05825.hmm	K05825	L-2-aminoadi	L-2-aminoadi	2-oxoadipate;	399.90|doma	Absent	0	None	Absent	0	None	Absent	0	None
5	Amino acid u	Phosphoserir	phosphoserir	phosphoserir	K00831.hmm	K00831	O-phospho-L	O-phospho-L	3-phosphoox	115.60|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
6	Amino acid u	Ornithine/ace	ornithine/acet	ornithine/acet	K00819.hmm	K00819, K00	L-ornithine +	L-ornithine; 2	L-glutamate 5	498.57|full, 48	Present	2	AL009126.3_	Present	2	CM000487.1_	Present	2	CM000489.1_
7	Amino acid u	Branched-ch	branched-cha	branched-cha	K00826.hmm	K00826, K02	L-leucine + 2	L-leucine; 2-c	4-methyl-2-ox	190.07|full, 16	Present	3	AL009126.3_	Present	3	CM000487.1_	Present	3	CM000489.1_
8	Amino acid u	Aspartate/tyr	aspartate/tyrc	aspartate/tyrc	K00812.hmm	K00812, K00	L-aspartate +	L-aspartate; 2	oxaloacetate;	471.87|full, 56	Present	2	AL009126.3_	Present	2	CM000487.1_	Present	2	CM000489.1_
9	Amino acid u	Histidinol-pho	histidinol-pho	histidinol-pho	K00817.hmm	K00817	L-histidinol pl	L-histidinol pl	3-(imidazol-4	294.13|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
10	Amino acid u	Serine-pyruva	serine-pyruva	serine-pyruva	K00830.hmm	K00830	L-alanine + g	L-alanine; gly	pyruvate; glyc	414.40|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
11	Ethanol ferm	Acetate => A	acetate => ac	acetate => ac	K00129.hmm	K00129, K00	an aldehyde ·	aldehyde; NA	carboxylate; I	710.50|full, 63	Absent	0	None,None	Absent	0	None,None	Absent	0	None,None
12	Ethanol ferm	Acetaldehyde	acetaldehyde	acetaldehyde	K13954.hmm	K13954, K00	a primary alc	primary alcoh	aldehyde; NA	560.37|doma	Absent	0	None,None,N	Absent	0	None,None,N	Absent	0	None,None,N
13	Fatty acid de	Fatty acid de	acyl-CoA deh	acyl-CoA deh	K00249.hmm	K00249, K00	a acyl-CoA +	acyl-CoA; ele	trans-2,3-deh	489.70|doma	Present	6	AL009126.3_	Present	6	CM000487.1_	Present	6	CM000489.1_
14	Aromatics de	Protocatechu	catA	catechol 1,2-	K03381.hmm	K03381	catechol + O	catechol; O2	cis,cis-muco	332.83|full	Absent	0	None	Absent	0	None	Absent	0	None
15	Aromatics de	Phenol => Be	ubiX	flavin prenyltr	K03186.hmm	K03186	dimethylallyl p	dimethylallyl p	prenylated FM	56.43|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
16	Aromatics de	Phenol => Be	bsdC	vanillate/4-hy	K01612.hmm	K01612	4-hydroxyber	4-hydroxyber	phenol; CO2	635.93|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
17	Aromatics de	Benzoyl-CoA	bcrC	benzoyl-CoA	K04112.hmm	K04112	cyclohexa-1,5	cyclohexa-1,5	benzoyl-CoA;	354.13|full	Absent	0	None	Absent	0	None	Absent	0	None
18	Aromatics de	Benzoyl-CoA	bcrB	benzoyl-CoA	K04113.hmm	K04113	cyclohexa-1,5	cyclohexa-1,5	benzoyl-CoA;	385.17|full	Absent	0	None	Absent	0	None	Absent	0	None
19	Aromatics de	Benzoyl-CoA	bcrA	benzoyl-CoA	K04114.hmm	K04114	cyclohexa-1,5	cyclohexa-1,5	benzoyl-CoA;	376.63|full	Absent	0	None	Absent	0	None	Absent	0	None
20	Aromatics de	Benzoyl-CoA	bcrD	benzoyl-CoA	K04115.hmm	K04115	cyclohexa-1,5	cyclohexa-1,5	benzoyl-CoA;	303.53|full	Absent	0	None	Absent	0	None	Absent	0	None
21	Complex carl	Cellulose deg	cellobiosidas	cellobiosidas	K19668.hmm	K19668	Hydrolysis of	N/A	N/A	237.77|doma	Absent	0	None	Absent	0	None	Absent	0	None
22	Complex carl	Cellulose deg	cellulase	cellulase	K01779.hmm	K01779	L-aspartate =	L-aspartate	D-aspartate	79.60|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
23	Complex carl	Cellulose deg	beta-glucosid	beta-glucosid	K01188.hmm	K01188, K05	Hydrolysis of	N/A	N/A	674.57|doma	Absent	0	None,None,N	Absent	0	None,None,N	Absent	0	None,None,N
24	Complex carl	Hemicullulos	arabinosidas	arabinosidas	K01209.hmm	K01209, K15	Hydrolysis of	N/A	N/A	178.03|full, 44	Present	3	AL009126.3_	Present	3	CM000487.1_	Present	3	CM000489.1_
25	Complex carl	Hemicullulos	beta-glucuror	beta-glucuror	K01195.hmm	K01195	a beta-D-gluc	beta-D-glucu	D-glucuronat	404.30|full	Absent	0	None	Absent	0	None	Absent	0	None
26	Complex carl	Hemicullulos	alpha-L-rham	alpha-L-rham	K05989.hmm	K05989	Hydrolysis of	N/A	N/A	131.47|full	Absent	0	None	Absent	0	None	Absent	0	None
27	Complex carl	Endohemicel	mannan endo	mannan endo	K01218.hmm	K01218, K19	Random hydi	N/A	N/A	78.03|domain	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
28	Complex carl	Endohemicel	alpha-D-xylos	alpha-D-xylos	K01811.hmm	K01811	Hydrolysis of	N/A	N/A	605.57|doma	Absent	0	None	Absent	0	None	Absent	0	None
29	Complex carl	Other oligosa	beta-xylosida	beta-xylosida	K01198.hmm	K01198	Hydrolysis of	N/A	N/A	107.63|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
30	Complex carl	Other oligosa	beta-mannos	beta-mannos	K01192.hmm	K01192	Hydrolysis of	N/A	N/A	365.50|doma	Absent	0	None	Absent	0	None	Absent	0	None
31	Complex carl	Other oligosa	beta-galactos	beta-galactos	K01190.hmm	K01190, K12	Hydrolysis of	N/A	N/A	361.13|doma	Present	1	None,None,A	Present	1	None,None,C	Present	1	None,None,C
32	Complex carl	Amylolytic en	alpha-amylas	alpha-amylas	K01176.hmm	K01176, K05	Endohydrolys	N/A	N/A	221.33|doma	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
33	Complex carl	Amylolytic en	glucoamylase	glucoamylase	K01178.hmm	K01178	Hydrolysis of	N/A	N/A	94.80|domain	Absent	0	None	Absent	0	None	Absent	0	None
34	Complex carl	Amylolytic en	pullulanase	pullulanase	K01200.hmm	K01200	Hydrolysis of	N/A	N/A	399.57|doma	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
35	Complex carl	Amylolytic en	isoamylase	isoamylase	K01214.hmm	K01214	Hydrolysis of	N/A	N/A	935.73|doma	Absent	0	None	Absent	0	None	Absent	0	None
36	Complex carl	Chitin degrad	chitiniase	chitiniase	K01183.hmm	K01183, K13	Random end	N/A	N/A	65.27|domain	Absent	0	None,None	Absent	0	None,None	Absent	0	None,None
37	Complex carl	Chitin degrad	hexosaminida	hexosaminida	K01207.hmm	K01207, K12	Hydrolysis of	N/A	N/A	267.07|doma	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_
38	Fermentation	Pyruvate oxid	porA	pyruvate ferre	K00169.hmm	K00169	pyruvate + C	pyruvate; Co	acetyl-CoA; C	291.77|full	Absent	0	None	Absent	0	None	Absent	0	None
39	Fermentation	Alcohol utiliza	adh	alcohol dehyc	K00001.hmm	K00001	a primary alc	primary alcoh	aldehyde; NA	324.20|doma	Absent	0	None	Absent	0	None	Absent	0	None
40	Fermentation	Lactate utiliza	ldh	L-lactate deh	K00016.hmm	K00016	(S)-lactate +	(S)-lactate; N	pyruvate; NAI	306.00|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_

56	C1 metabolis	Formate oxid	fdwB	formate dehy	K22515.hmm	K22515	formate + NA	formate; NAD	CO2; NADH	548.60|doma	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
57	C1 metabolis	Formate oxid	fdoH	formate dehy	K00124.hmm	K00124	formate + NA	formate; NAD	CO2; NADH	160.27|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
58	C1 metabolis	Formate oxid	fdhA	formate dehy	K22516.hmm	K22516	formate + oxi	formate; oxid	CO2; reduce	1010.23|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
59	C1 metabolis	Formate oxid	fdhB	formate dehy	K00125.hmm	K00125	formate + oxi	formate; oxid	CO2; reduce	236.53|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
60	C1 metabolis	Aerobic CO c	coxS	aerobic carbo	K03518.hmm	K03518	CO + a quinc	CO; quinone;	CO2; quinol	241.60|doma	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
61	C1 metabolis	Aerobic CO c	coxM	aerobic carbo	K03519.hmm	K03519	CO + a quinc	CO; quinone;	CO2; quinol	265.87|doma	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
62	C1 metabolis	Aerobic CO c	coxL	aerobic carbo	K03520.hmm	K03520	CO + a quinc	CO; quinone;	CO2; quinol	835.30|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
63	Methane met	Methane oxid	pmoA	methane/amr	pmoA.hmm	K10944	methane + qu	methane; qui	methanol; qu	92.50|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
64	Methane met	Methane oxid	pmoB	methane/amr	pmoB.hmm	K10945	methane + qu	methane; qui	methanol; qu	202.10|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
65	Methane met	Methane oxid	pmoC	methane/amr	pmoC.hmm	K10946	methane + qu	methane; qui	methanol; qu	124.60|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
66	Methane met	Methane oxid	mmoB	methane mor	PF02406.hm	K16160	methane + N.	methane; NA	methanol; NA	10|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
67	Methane met	Methane oxid	mmoD	methane mor	TIGR04550.h	K16162	methane + N.	methane; NA	methanol; NA	28.00|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
68	Methane met	Methane proc	mcrA	methyl-coenz	TIGR03256.h	K00399	methyl-CoM ·	methyl-CoM;	CoM-S-S-Co	314.45|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
69	Methane met	Methane proc	mcrB	methyl-coenz	TIGR03257.h	K00401	methyl-CoM ·	methyl-CoM;	CoM-S-S-Co	173.35|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
70	Methane met	Methane proc	mcrC	methyl-coenz	TIGR03259.h	K03421	methyl-CoM ·	methyl-CoM;	CoM-S-S-Co	171.60|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
71	Carbon fixatic	CBB cycle - I	Form I	Form I RuBis	rubisco_form	K01601	D-ribulose 1,5	D-ribulose 1,5	3-phospho-D	500|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
72	Carbon fixatic	CBB cycle - I	Form II	Form II RuBis	rubisco_form	K01601	D-ribulose 1,5	D-ribulose 1,5	3-phospho-D	500|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
73	Carbon fixatic	CBB cycle - I	Form II/III	Form II/III Rub	rubisco_form	K01601	D-ribulose 1,5	D-ribulose 1,5	3-phospho-D	500|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
74	Carbon fixatic	CBB cycle - I	Form III	Form III RuBi	rubisco_form	K01601	D-ribulose 1,5	D-ribulose 1,5	3-phospho-D	450|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
75	Carbon fixatic	CBB cycle - I	Form IV	Form IV RuB	rubisco_form	K01601	D-ribulose 1,5	D-ribulose 1,5	3-phospho-D	500|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
76	Carbon fixatic	3 Hydroxypro	prpE	propionyl-Co	K01908.hmm	K01908	2 3-phospho-	3-phospho-D	D-ribulose 1,5	784.73|doma	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
77	Carbon fixatic	3 Hydroxypro	mcr	malonyl-CoA	K14468.hmm	K14468	3-hydroxypro	3-hydroxypro	malonate ser	1893.80|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
78	Carbon fixatic	3 Hydroxypro	K15017	malonyl-CoA	K15017.hmm	K15017	succinate se	succinate se	succinyl-CoA	702.03|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
79	Carbon fixatic	3HP/4HB	K14466	4-hydroxybut	K14466.hmm	K14466	ATP + 4-hydr	ATP; 4-hydro	AMP; diphosp	920.27|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
80	Carbon fixatic	3HP/4HB	4hbl	4-hydroxybut	K14467.hmm	K14467	ATP + 4-hydr	ATP; 4-hydro	AMP; diphosp	1016.20|dom	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
81	Carbon fixatic	3HP/4HB	K18861	4-hydroxybut	K18861.hmm	K18861	ATP + 4-hydr	ATP; 4-hydro	AMP; diphosp	950.13|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
82	Carbon fixatic	3HP/4HB	E6.2.1.56	4-hydroxybut	K18593.hmm	K18593	ATP + 4-hydr	ATP; 4-hydro	ADP; phosph	1233.23|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
83	Carbon fixatic	3HP/4HB	abfD	4-hydroxybut	K14534.hmm	K14534	4-hydroxybut	4-hydroxybut	(E)-but-2-enc	420.77|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
84	Carbon fixatic	Wood Ljungd	cdhD	acetyl-CoA de	K00194.hmm	K00194	a [methyl-Co	[methyl-Co(III	[Co(I) corrino	169.40|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
85	Carbon fixatic	Wood Ljungd	cdhE	acetyl-CoA de	K00197.hmm	K00197	a [methyl-Co	[methyl-Co(III	[Co(I) corrino	206.30|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
86	Carbon fixatic	Wood Ljungd	cooS	anaerobic ca	K00198.hmm	K00198	a [methyl-Co	[methyl-Co(III	[Co(I) corrino	231.90|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
87	Carbon fixatic	Reverse TCA	aclA	ATP-citrate ly	acetate_citra	K15230	ADP + phosp	ADP; phosph	ATP; citrate;	215|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
88	Carbon fixatic	Reverse TCA	aclB	ATP-citrate ly	acetate_citra	K15231	ADP + phosp	ADP; phosph	ATP; citrate;	177|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
89	Nitrogen cycl	Ammonia oxi	amoA	methane/amr	amoA.hmm	K10944	NH3 + a redu	NH3; reducec	NH2OH; acce	92.50|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
90	Nitrogen cycl	Ammonia oxi	amoB	methane/amr	amoB.hmm	K10945	NH3 + a redu	NH3; reducec	NH2OH; acce	202.10|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
91	Nitrogen cycl	Ammonia oxi	amoC	methane/amr	amoC.hmm	K10946	NH3 + a redu	NH3; reducec	NH2OH; acce	124.60|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
92	Nitrogen cycl	N2 fixation	anfD	nitrogenase ir	TIGR01861.h	N/A	8 reduced fer	reduced ferre	oxidized ferre	873.55|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
93	Nitrogen cycl	N2 fixation	anfK	nitrogenase ir	TIGR02931.h	N/A	8 reduced fer	reduced ferre	oxidized ferre	853.75|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
94	Nitrogen cycl	N2 fixation	anfG	nitrogenase ir	TIGR02929.h	K00531	8 reduced fer	reduced ferre	oxidized ferre	124.50|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0

169	Oxidative phc	Complex V (F	atpD (F-type)	F-type H+/Na	K02112.hmm	K02112	ATP + H2O +	ATP; H2O; H	ADP; phosph	548.27|doma	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
170	Oxygen meta	Oxygen meta	coxA	cytochrome c	TIGR02891.h	K02274	4 ferrocytoch	ferrocytochro	ferricytochror	617.40|full	Present	2	AL009126.3_	Present	2	CM000487.1_	Present	2	CM000489.1_	Present	2
171	Oxygen meta	Oxygen meta	coxB	cytochrome c	TIGR02866.h	K02275	4 ferrocytoch	ferrocytochro	ferricytochror	144.30|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
172	Oxygen meta	Oxygen meta	ccoN	cytochrome c	TIGR00780.h	K00404	4 ferrocytoch	ferrocytochro	ferricytochror	168.15|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
173	Oxygen meta	Oxygen meta	ccoO	cytochrome c	TIGR00781.h	K00405	4 ferrocytoch	ferrocytochro	ferricytochror	62.85|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
174	Oxygen meta	Oxygen meta	ccoP	cytochrome c	TIGR00782.h	K00406	4 ferrocytoch	ferrocytochro	ferricytochror	54.05|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
175	Oxygen meta	Oxygen meta	cyoA	cytochrome c	TIGR01433.h	K02297	2 ubiquinol +	ubiquinol; O2	ubiquinone; H	218.75|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
176	Oxygen meta	Oxygen meta	cyoD	cytochrome c	TIGR02847.h	K02300	2 ubiquinol +	ubiquinol; O2	ubiquinone; H	64.95|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
177	Oxygen meta	Oxygen meta	cyoB	cytochrome c	K02298.hmm	K02298	2 ubiquinol +	ubiquinol; O2	ubiquinone; H	970.97|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
178	Oxygen meta	Oxygen meta	cyoC	cytochrome c	K02299.hmm	K02299	2 ubiquinol +	ubiquinol; O2	ubiquinone; H	258.27|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
179	Oxygen meta	Oxygen meta	cydA	cytochrome b	PF01654.hm	K00425	2 ubiquinol +	ubiquinol; O2	ubiquinone; H	118|full	Present	2	AL009126.3_	Present	2	CM000487.1_	Present	2	CM000489.1_	Present	2
180	Oxygen meta	Oxygen meta	cydB	cytochrome b	TIGR00203.h	K00426	2 ubiquinol +	ubiquinol; O2	ubiquinone; H	88.55|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
181	Oxygen meta	Oxygen meta	qoxA	cytochrome a	TIGR01432.h	K02826	2 menaquino	menaquinol; (	menaquinone	204.55|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
182	Oxygen meta	Oxygen meta	qoxB	cytochrome a	TIGR02882.h	K02827	2 menaquino	menaquinol; (	menaquinone	954.70|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
183	Urea utilizatio	Urease	ureC	urease subur	TIGR01792.h	K01428	urea + H2O =	urea; H2O	CO2; NH3	212.95|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
184	Urea utilizatio	Urease	ureB	urease subur	TIGR00192.h	K01429	urea + H2O =	urea; H2O	CO2; NH3	40.65|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
185	Urea utilizatio	Urease	ureA	urease subur	TIGR00193.h	K01430	urea + H2O =	urea; H2O	CO2; NH3	31.45|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
186	Halogenated	Halogenated	E3.8.1.2	2-haloacid de	K01560.hmm	K01560	(S)-2-haloaci	(S)-2-haloaci	(R)-2-hydroxy	143.23|doma	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
187	Halogenated	Halogenated	pcpC	tetrachlorohy	K15241.hmm	K15241	2,6-dichloroh	2,6-dichloroh	2,3,6-trichlorc	276.67|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
188	Halogenated	Halogenated	cprA	3-chloro-4-hy	K21566.hmm	K21566	1-haloalkane	1-haloalkane;	primary alcoh	850.93|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
189	Halogenated	Halogenated	pceA	tetrachloroeth	K21647.hmm	K21647	trichloroether	trichloroether	tetrachloroeth	50|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
190	Perchlorate r	Perchlorate r	pcrA	perchlorate re	TIGR03479.h	N/A	reduced acce	perchlorate; r	chlorite; acce	559.05|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1
191	Perchlorate r	Perchlorate r	pcrB	perchlorate re	TIGR03478.h	N/A	reduced acce	perchlorate; r	chlorite; acce	319.40|full	Absent	0	None	Absent	0	None	Absent	0	None	Absent	0
192	Chlorite redu	Chlorite redu	cld	chlorite dism	PF06778.hm	K09162	chloride + O2	chloride; O2	chlorite	65|full	Present	1	AL009126.3_	Present	1	CM000487.1_	Present	1	CM000489.1_	Present	1

Figure 16: Several parts of the ‘gene_presence_absence’ output files

As shown in the figure 15 above, the excel files present the presence and absence of genes in the whole 476 genomes. As seen in the figure, the leftmost column is the category of each gene, and the third and fourth columns are the abbreviations and names of the genes, respectively. The top line shows the name of the genome, the suffixes are Hmm.presence, Hit.number and Hits, these three represent the information of one genome. In the column with the suffix Hmm.presence, green means that the gene is present in the genome, and red means that the gene is not present which shows absence in the genome. In the column with the suffix Hit.numbers, 0 means that the gene does not exist in the genome, and a number other than 0 means that the gene exists.

As mentioned before in this paper, in combination with the relevant literature of other scholars and researchers, it is understood that the key step in self-healing concrete is the precipitation of calcium carbonate, and the urease in *Bacillus subtilis* plays a major role in the calcium carbonate precipitation process. Unsurprisingly, the relevant information about the urease gene was found in this excel file, which is shown in the third picture of the figure 15. As shown in the picture and shown in the document, urease contains three genes, namely ureC, ureB and ureA, and the corresponding gene names are urease subunit alpha, urease subunit beta and urease subunit gamma. These three genes appear in almost all genomes except for genome 'GCA _ 000815405.1 _ ASM81540v1', genome 'GCA _ 001286785.1 _ *Bacillus* _ JRS10', genome 'GCA _ 001698525.1 _ ASM169852v1' and genome 'GCA _ 008764245.1 _ FS357'.

This shows that the gene to which urease belongs almost exists in all the genomes of *Bacillus subtilis* and indicates that the fundamental reason why *Bacillus subtilis* can contribute to the precipitation of calcium carbonate is that urease is abundantly present in this bacterium.

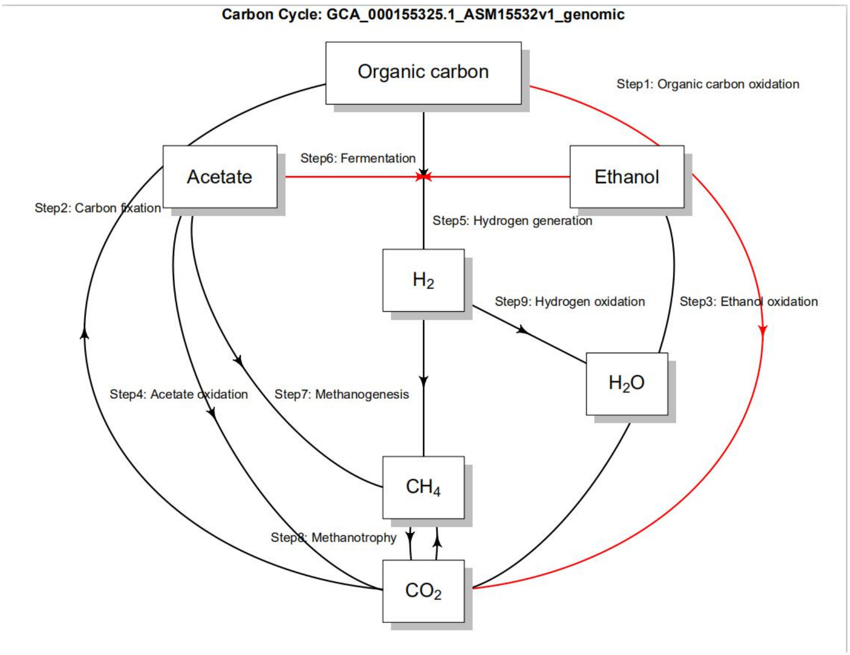


Figure 17: Schematic diagram of carbon cycle for genome GCA_000155325.1_ASM15532v1

As mentioned above, taking a genome where the urease gene is located as an example. As shown in the figure 16, this is a schematic diagram of the carbon cycle involved in this genome GCA_000155325.1_ASM15532v1. As it shows that this genome is mainly involved in the organic carbon oxidation process, which converts organic carbon into ethanol; and involved in the ethanol oxidation process, and the oxidation of the ethanol produces carbon dioxide gas; and it is also mainly involved in the ethanol fermentation process.

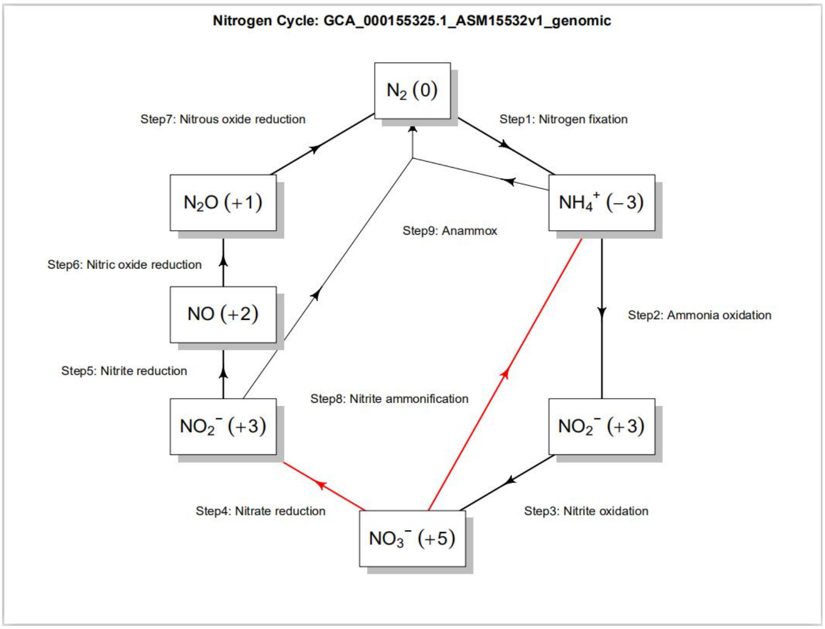


Figure 18: Schematic diagram of nitrogen cycle for genome GCA_000155325.1_ASM15532v1

As shown in the figure 17, this is a schematic diagram of the nitrogen cycle involved in this genome GCA_000155325.1_ASM15532v1. As it shows that this genome is mainly involved in the nitrite ammonification process, which converts nitrate ion into ammonium ion; and involved in the nitrate reduction process, which reduces nitrate to nitrite.

As mentioned in the above two diagrams and their meanings, the enzyme urease exists in the above genome, indicating that the genes ureA, ureB and ureC are included in this genome. Combined with the relevant research data mentioned above, it is confirmed that urease is involved in carbon cycle and nitrogen cycle during calcium carbonate precipitation in self-healing concrete.

For the sulfur cycle, the metabolic results also include a schematic diagram of this process. Combined with the metabolic excel results, several example genes such like sdo gene and sat gene and several example enzymes involved in the sulfur cycle have been selected.

Taking a genome where the sat gene is located as an example to show the schematic diagram of the sulfur cycle. The excel results of the metabolic show that the sat gene is also included in the GCA_000155325.1_ASM15532v1 genome.

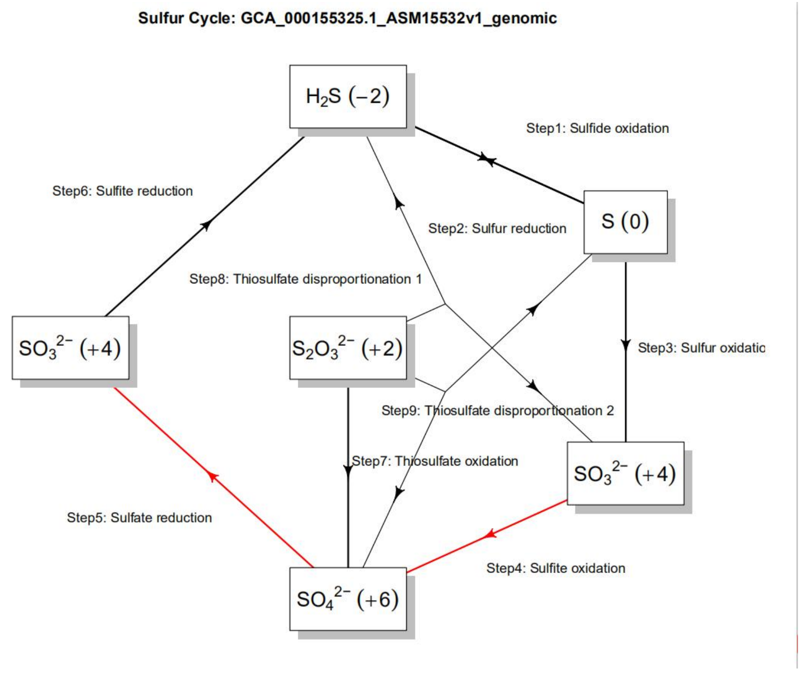


Figure 19: Schematic diagram of sulfur cycle for genome GCA_000155325.1_ASM15532v1

As shown in the figure 18, this is a schematic diagram of the sulfur cycle involved in this genome GCA_000155325.1_ASM15532v1. As it shows that this genome is mainly involved in the sulfite oxidation process, which converts sulfite ion into sulfate ion; and involved in the sulfate reduction process, which reduces sulfate ion to sulfite ion.

According to the above analysis process, some significant genes which are related to some important chemical processes in the self-healing concrete were selected from the excel results of metabolic. The following are their respective gene names and the names of the corresponding enzymes.

Table 4: Several significant genes selected from Metabolic excel result

Gene	Enzyme
ureA	urease subunit gamma
ureB	urease subunit beta
ureC	urease subunit alpha
sdo	sulfur dioxygenase

sat	sulfate adenylyl transferase
nirB	nitrite reductase (NADH) large subunit
nirD	nitrite reductase (NADH) small subunit
cysC	adenylyl sulfate kinase

As shown in the table 4, combined with the previous results, some of the genes are related to the carbon cycle of the genome that contains them such like ureA, ureB and ureC, which are mainly involved in the oxidizable organic carbon process, Ethanol oxidation process and ethanol fermentation process in the carbon cycle; and some genes are related to the nitrogen cycle such like nirB and nirD (except for the three genes corresponding to urease), which are mainly involved in the nitrate reduction process and nitrite ammonification process in the nitrogen cycle. The remaining genes like sdo, sat and cysC are related to the sulfur cycle of the genome that contains them, which are mainly involved in the sulfite oxidation process and sulfate reduction process in the sulfur cycle.

3.3 Coinfinder results

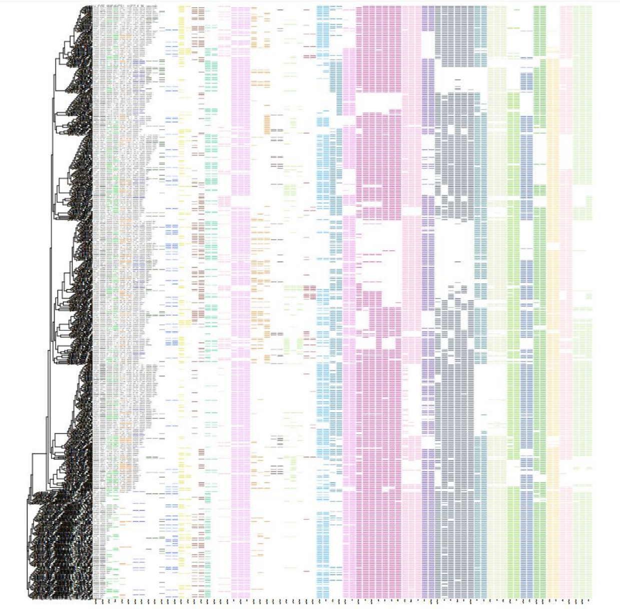


Figure 20: A portion of the heatmaps from Coinfinder of the presence/absence patterns of the associating gene sets for *Bacillus subtilis* pan genomes

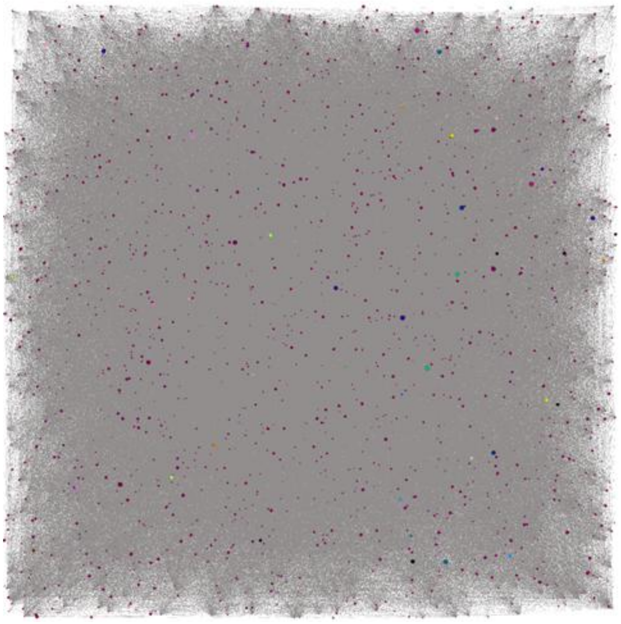


Figure 21: Gene association network output from Coinfinder for *Bacillus subtilis* pan genomes

As shown above in figure 19 and figure 20, each gene (node) is connected to (edge) another gene if they are statistically related to each other (associate with each other) in the pan-genome. Connected components (i.e., overlapping gene sets) color the nodes, and the colors correlate to the heat map output colors. Coinfinder generates a network file with all nodes and edge coloring.

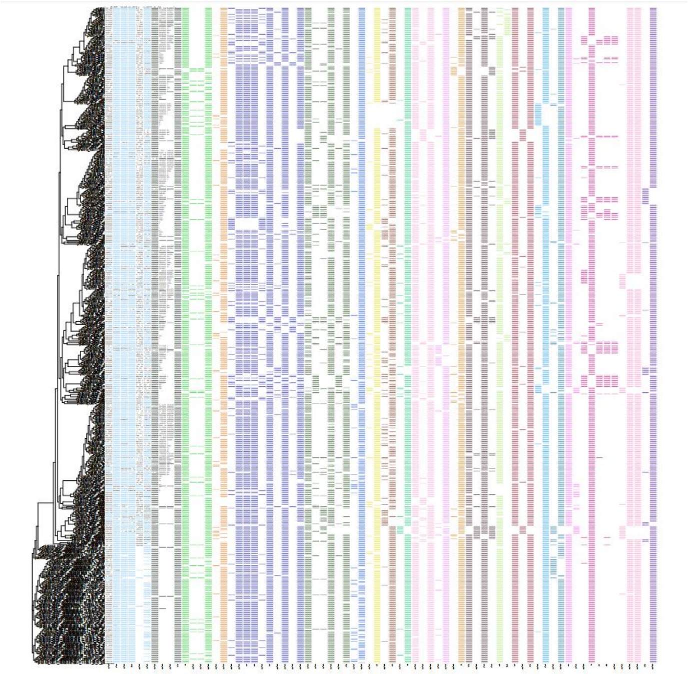


Figure 22: A portion of the heatmaps from Coinfinder of the presence/absence patterns of the dissociating gene sets for *Bacillus subtilis* pan genomes

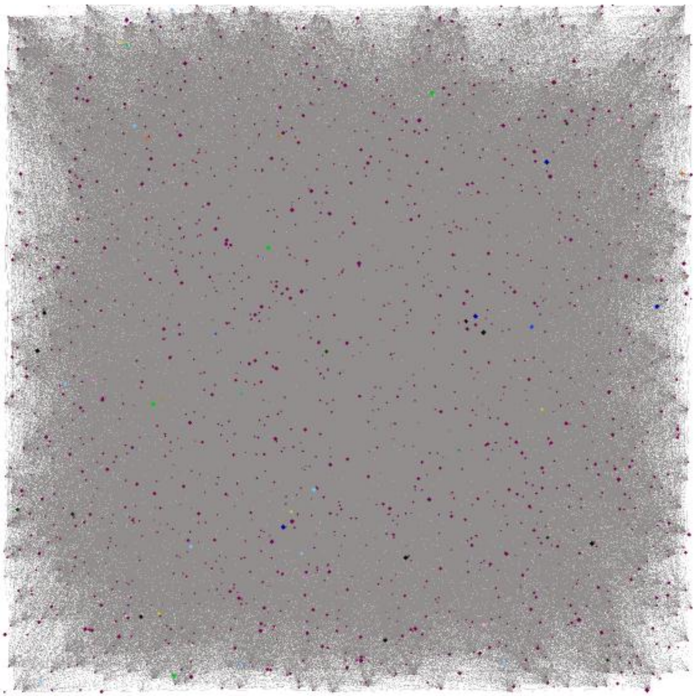


Figure 23: Gene dissociation network output from Coinfinder for *Bacillus subtilis* pan genomes

As shown above in figure 21 and figure 22, each gene (node) is connected to (edge) another gene if they are statistically separated from each other (dissociate with each other) in the pan-genome. Connected components (i.e., overlapping gene sets) color the nodes, and the colors correlate to the heat map output colors.

4 Discussion

After analyzing the whole genome of *Bacillus subtilis* through several related software, combined with all the results in the above results section, it can be clearly understood that *Bacillus subtilis* can be used as a bacterium that meets the requirements of biological self-healing concrete. The fundamental reason why *Bacillus subtilis* can be used as a bacterium that promotes concrete self-healing is that many of its genomes contain enzymes and genes that play a key role in important chemical processes in self-healing concrete (such as calcium carbonate precipitation). The bacterium *Bacillus subtilis*, which is a self-healing concrete microorganism in this article, can produce urease through metabolism. The urease of most species contains 7 or more genes such as ureA, B, C, D, E, F, G, etc. It has been proved that the urease of *Bacillus subtilis* is a typical trimer structure (Christians and Kaltwasser, 1986). To begin, the ureA, B, and C genes encode the three subunits (α, β, γ) of urease, and the α, β, γ and other three subunits combine to create a monomer, which is then compounded to produce the Trimer structure (Karplus et al., 1997; Benini et al., 2013). In addition, ureD, E, F, and G respectively encode several auxiliary proteins necessary for the degradation of urea.

Urease plays a very important role in the precipitation process of calcium carbonate in self-healing concrete. When *Bacillus subtilis* is put into concrete, it can produce a large amount of urease through metabolism. After the urea in the environment enters the bacterial cell, it can catalyze hydrolysis under the action of a large amount of urease inside the bacterial cell. The hydrolysis process can be roughly divided into two steps: (1) Under the catalysis of urease, 1 molecule of urea is hydrolyzed into 1 molecule of carbamate and 1 molecule of ammonia; (2) 1 molecule of carbamate is spontaneously hydrolyzed to form 1 molecule of carbonic acid and the second molecule of ammonia. In the end, 1 molecule of urea hydrolysate is 2 molecules of ammonia and 1 molecule of carbonic acid. The chemical equation of the whole process is as follows:

$$\mathrm{CO(NH_2)_2} + \mathrm{H_2O} \rightarrow \mathrm{NH_3} + \mathrm{NH_2COOH}$$

$$\mathrm{NH_2COOH} + \mathrm{H_2O} \rightarrow \mathrm{NH_3} + \mathrm{H_2CO_3}$$

$$\mathrm{H_2CO_3} \rightarrow 2\mathrm{H^+} + \mathrm{CO_3^{2-}}$$

$$\mathrm{NH_3} + \mathrm{H_2O} \leftrightarrow \mathrm{NH_4^+} + \mathrm{H^+}$$

Calcium carbonate precipitation is a fairly simple chemical process, which is mainly controlled by 4 key factors (Hammes and Verstraete*, 2002): calcium concentration, soluble inorganic carbon concentration, pH value, and whether there are nucleation sites. The calcium carbonate precipitation process is very slow in natural conditions, but microorganisms can influence or induce calcium carbonate precipitation by modifying any of the essential parameters impacting calcium carbonate precipitation listed above (Hammes and Verstraete*, 2002), thereby producing a large amount of calcium carbonate in a short time.

$$\mathrm{Ca}^{+} + \mathrm{CO}_3^{2-} \rightarrow \mathrm{CaCO}_3$$

Bacillus subtilis can induce calcium carbonate precipitation, mainly due to two characteristics: First, bacillus subtilis can secrete a large amount of highly active urease, which can quickly catalyze the hydrolysis of urea in the environment, thus increasing the concentration of carbonate and the value of pH in the environment of bacteria; Secondly, studies have shown that The surface of *Bacillus subtilis* has more negative surface charge than some other non-mineralized bacteria, and has stronger adsorption effect on cations. It can adsorb a large amount of calcium ions in the environment, to precipitate calcium carbonate crystals in the environment with high concentration of carbonate and high pH value.

In addition, for the precipitation of specific compounds in self-healing concrete to repair cracks, in addition to calcium carbonate precipitation method, there are precipitation methods such as multistate ferroaluminum silicate; But this kind of crack repairing method is not suitable for self-healing concrete because of some problems such as not adapting to acidic environment and reducing concrete durability.

Through statistical analysis of the whole genome of *Bacillus subtilis*, this project found that urease and its genes are abundantly present in the genome of *Bacillus subtilis*, thereby inferring that urease and its genes are important for the calcium carbonate precipitation process in self-healing concrete. This fact is confirmed by consulting other research data, that is, *Bacillus subtilis* has an inducing effect on calcium carbonate precipitation because it can produce a large amount of urease. However, there are many types of bacteria that help calcium carbonate precipitation in self-healing concrete, which are not covered in this project. There are many other enzymes and genes in other bacteria that are helpful to this process, which will be mentioned and analyzed in other literature experimental data or in future research.

5 Conclusion

As the main building material in the world, concrete has a huge amount of usage and economic effects worldwide. Since the buildings involved in the construction of concrete materials are closely related to human life activities, the stability and durability of concrete materials must be the focus of consideration by related researchers. Due to the inherent heterogeneity of concrete, micro-cracks in concrete are inevitable. The existence of cracks poses a potential threat to the safety and stability of building structures. Therefore, it is very important to repair concrete cracks in a timely and efficient manner. Due to the increasing cost and difficulty of man-made maintenance, as well as the hazards of the production of concrete to the natural environment, the idea of using microorganisms to heal concrete cracks has been put forward and analyzed by more and more researchers.

The *Bacillus subtilis* mentioned in this project is the representative bacteria that uses microorganisms to heal concrete cracks. Through the statistical analysis and research of this project, it is concluded that the key reason for *Bacillus subtilis* to heal concrete cracks is that many genomes contain urease and its genes in the whole genome, which has a strong inducing effect on calcium carbonate precipitation, which can be used for concrete Cracks are repaired.

With the increasing use of concrete in the world in the future, the concept of self-healing concrete will also be paid attention to by more and more researchers; as a low-cost and high-efficiency method of healing concrete, microorganisms Self-healing concrete will be paid more and more attention, and it will be used more and more in actual engineering.

Reference

Ahn, T.-H., Kishi, T., 2010. Crack Self-healing Behavior of Cementitious Composites Incorporating Various Mineral Admixtures. J. Adv. Concr. Technol. 8, 171–186. https://doi.org/10.3151/jact.8.171

AltschuP, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., n.d. Basic Local Alignment Search Tool 8.

Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., Thomas, B.C., Singh, A., Wilkins, M.J., Karaoz, U., Brodie, E.L., Williams, K.H., Hubbard, S.S., Banfield, J.F., 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. 7, 13219. https://doi.org/10.1038/ncomms13219

Barabesi, C., Galizzi, A., Mastromei, G., Rossi, M., Tamburini, E., Perito, B., 2007. *Bacillus subtilis* Gene Cluster Involved in Calcium Carbonate Biomineralization. J. Bacteriol. 189, 228–235. https://doi.org/10.1128/JB.01450-06

Berner, R.A., 1975. The role of magnesium in the crystal growth of calcite and aragonite from sea water. Geochim. Cosmochim. Acta 39, 489–504. https://doi.org/10.1016/0016-7037(75)90102-7

Castanier, S., Le Métayer-Levrel, G., Perthuisot, J.-P., 1999. Ca-carbonates precipitation and limestone genesis — the microbiogeologist point of view. Sediment. Geol. 126, 9–23. https://doi.org/10.1016/S0037-0738(99)00028-7

Castanier, S., Métayer-Levrel, G.L., Perthuisot, J.-P., 2000. Bacterial Roles in the Precipitation of Carbonate Minerals, in: Riding, R.E., Awramik, S.M. (Eds.), Microbial Sediments. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 32–39. https://doi.org/10.1007/978-3-662-04036-2_5

Castro-Alonso, M.J., Montañez-Hernandez, L.E., Sanchez-Muñoz, M.A., Macias Franco, M.R., Narayanasamy, R., Balagurusamy, N., 2019. Microbially Induced Calcium Carbonate Precipitation (MICP) and Its Potential in Bioconcrete: Microbiological and Molecular Concepts. Front. Mater. 0. https://doi.org/10.3389/fmats.2019.00126

Christians, S., Kaltwasser, H., 1986. Nickel-content of urease from Bacillus pasteurii. Arch. Microbiol. 145, 51–55. https://doi.org/10.1007/BF00413026

Cruz Ramos, H., Hoffmann, T., Marino, M., Nedjari, H., Presecan-Siedel, E., Dreesen, O., Glaser, P., Jahn, D., 2000. Fermentative Metabolism of *Bacillus subtilis* : Physiology and Regulation of Gene Expression. J. Bacteriol. 182, 3072–3080. https://doi.org/10.1128/JB.182.11.3072-3080.2000

Dhir, R. K, Jones, M. R., 1999. Innovation in concrete structures: design and construction. Thomas Telford Ltd. https://doi.org/10.1680/iicsdac.28241

Dry, C., 1994. Matrix cracking repair and filling using active and passive modes for smart timed release of chemicals from fibers into cement matrices. Smart Mater. Struct. 3, 118–123. https://doi.org/10.1088/0964-1726/3/2/006

Earl, A.M., Losick, R., Kolter, R., 2008. Ecology and genomics of Bacillus subtilis. Trends Microbiol. 16, 269–275. https://doi.org/10.1016/j.tim.2008.03.004

Feng, J., Chen, B., Sun, W., Wang, Y., 2021. Microbial induced calcium carbonate precipitation study using Bacillus subtilis with application to self-healing concrete preparation and characterization. Constr. Build. Mater. 280, 122460. https://doi.org/10.1016/j.conbuildmat.2021.122460

Ferris, F.G., Fyfe, W.S., Beveridge, T.J., 1988. Metallic ion binding by Bacillus subtilis: Implications for the fossilization of microorganisms 4.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230. https://doi.org/10.1093/nar/gkt1223

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37. https://doi.org/10.1093/nar/gkr367

Ghosh, S.K., 2008. Self-Healing Materials: Fundamentals, Design Strategies, and Applications, in: Ghosh, S.K. (Ed.), Self-Healing Materials. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 1–28. https://doi.org/10.1002/9783527625376.ch1

Glaser, P., Danchin, A., Kunst, F., Zuber, P., Nakano, M.M., 1995. Identification and isolation of a gene required for nitrate assimilation and anaerobic growth of Bacillus subtilis. J. Bacteriol. 177, 1112–1115. https://doi.org/10.1128/jb.177.4.1112-1115.1995

Gollapudi, U.K., Knutson, C.L., Bang, S.S., Islam, M.R., 1995. A new method for controlling leaching through permeable channels. Chemosphere 30, 695–705. https://doi.org/10.1016/0045-6535(94)00435-W

Hammes, F., Verstraete*, W., 2002. Key roles of pH and calcium metabolism in microbial carbonate precipitation. Rev. Environ. Sci. Biotechnol. 1, 3–7. https://doi.org/10.1023/A:1015135629155

Han, B., Yu, X., Ou, J., 2014. Challenges of Self-Sensing Concrete, in: Self-Sensing Concrete in Smart Structures. Elsevier, pp. 361–376. https://doi.org/10.1016/B978-0-12-800517-0.00011-3

Härtig, E., Jahn, D., 2012. Regulation of the Anaerobic Metabolism in Bacillus subtilis, in: Advances in Microbial Physiology. Elsevier, pp. 195–216. https://doi.org/10.1016/B978-0-12-394423-8.00005-6

Hoffmann, T., Troup, B., Szabo, A., Hungerer, C., Jahn, D., 1995. The anaerobic life of *Bacillus subtilis* : Cloning of the genes encoding the respiratory nitrate reductase system. FEMS Microbiol. Lett. 131, 219–225. https://doi.org/10.1111/j.1574-6968.1995.tb07780.x

Huang, X., Kaewunruen, S., 2020. Self-healing concrete, in: New Materials in Civil Engineering. Elsevier, pp. 825–856. https://doi.org/10.1016/B978-0-12-818961-0.00027-2

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Jonkers, H.M., 2007. Self Healing Concrete: A Biological Approach, in: van der Zwaag, S. (Ed.), Self Healing Materials, Springer Series in Materials Science. Springer Netherlands, Dordrecht, pp. 195–204. https://doi.org/10.1007/978-1-4020-6250-6_9

Jonkers, H.M., Thijssen, A., Muyzer, G., Copuroglu, O., Schlangen, E., 2010. Application of bacteria as self-healing agent for the development of sustainable concrete. Ecol. Eng. 36, 230–235. https://doi.org/10.1016/j.ecoleng.2008.12.036

Kanehisa, M., 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30, 42–46. https://doi.org/10.1093/nar/30.1.42

Karplus, P.A., Pearson, M.A., Hausinger, R.P., 1997. 70 Years of Crystalline Urease: What Have We Learned? Acc. Chem. Res. 30, 330–337. https://doi.org/10.1021/ar960022j

Kovács, Á.T., 2019. Bacillus subtilis. Trends Microbiol. 27, 724–725. https://doi.org/10.1016/j.tim.2019.03.008

Li, V.C., Yang, E.-H., 2007. Self Healing in Concrete Materials, in: van der Zwaag, S. (Ed.), Self Healing Materials, Springer Series in Materials Science. Springer Netherlands, Dordrecht, pp. 161–193. https://doi.org/10.1007/978-1-4020-6250-6_8

McKenney, P.T., Driks, A., Eichenberger, P., 2013. The Bacillus subtilis endospore: assembly and functions of the multilayered coat. Nat. Rev. Microbiol. 11, 33–44. https://doi.org/10.1038/nrmicro2921

Michel, J.F., Piechaud, M., Schaeffer, P., 1970. [Nitrate-reductase constitutivity for nitrate in early asporogenous mutants of Bacillus subtilus. Ann. Inst. Pasteur 119, 711–718.

Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., Kanehisa, M., 2013. Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. J. Chem. Inf. Model. 53, 613–622. https://doi.org/10.1021/ci3005379

Nakano, M.M., Dailly, Y.P., Zuber, P., Clark, D.P., 1997. Characterization of anaerobic fermentative growth of Bacillus subtilis: identification of fermentation end products and genes required for growth. J. Bacteriol. 179, 6749–6755. https://doi.org/10.1128/jb.179.21.6749-6755.1997

Nijland, T.G., Larbi, J.A., de Rooij, M., 2007. SELF HEALING PHENOMENA IN CONCRETES AND MASONRY MORTARS: A MICROSCOPIC STUDY 10.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Reinhardt, H.-W., Jooss, M., 2003. Permeability and self-healing of cracked concrete as a function of temperature and crack width. Cem. Concr. Res. 33, 981–985. https://doi.org/10.1016/S0008-8846(02)01099-2

Sanchez-Moral, S., Canaveras, J.C., Laiz, L., Saiz-Jimenez, C., Bedoya, J., Luque, L., 2003. Biomediated Precipitation of Calcium Carbonate Metastable Phases in Hypogean Environments: A Short Review. Geomicrobiol. J. 20, 491–500. https://doi.org/10.1080/713851131

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Seifan, M., Samani, A.K., Berenjian, A., 2016. Bioconcrete: next generation of self-healing concrete. Appl. Microbiol. Biotechnol. 100, 2591–2602. https://doi.org/10.1007/s00253-016-7316-z

Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., White, O., 2007. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 35, D260–D264. https://doi.org/10.1093/nar/gkl1043

Siddique, R., Chahal, N.K., 2011. Effect of ureolytic bacteria on concrete properties. Constr. Build. Mater. 25, 3791–3801. https://doi.org/10.1016/j.conbuildmat.2011.04.010

Sitto, F., Battistuzzi, F.U., 2020. Estimating Pangenomes with Roary. Mol. Biol. Evol. 37, 933–939. https://doi.org/10.1093/molbev/msz284

Talaiekhozan, A., Keyvanfar, A., Shafaghat, A., Andalib, R., Majid, M.Z.A., Fulazzaky, M.A., Zin, R.M., Lee, C.T., Hussin, M.W., Hamzah, N., Marwar, N.F., Haidar, H.I., 2014. A Review of Self-healing Concrete Research Development 2, 12.

Tange, O., n.d. GNU Parallel: The Command-Line Power Tool 6.

Teall, O.R., n.d. Crack Closure and Enhanced Autogenous Healing of Structural Concrete Using Shape Memory Polymers 278.

v. Knorre, H., Krumbein, W.E., 2000. Bacterial Calcification, in: Riding, R.E., Awramik, S.M. (Eds.), Microbial Sediments. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 25–31. https://doi.org/10.1007/978-3-662-04036-2_4

van Dongen, S., n.d. Performance criteria for graph clustering and Markov cluster experiments. . Introduction 38.

Van Tittelboom, K., De Belie, N., 2013. Self-Healing in Cementitious Materials—A Review. Materials 6, 2182–2217. https://doi.org/10.3390/ma6062182

Van Tittelboom, K., De Belie, N., De Muynck, W., Verstraete, W., 2010. Use of bacteria to repair cracks in concrete. Cem. Concr. Res. 40, 157–166. https://doi.org/10.1016/j.cemconres.2009.08.025

Whelan, F.J., Rusilowicz, M., McInerney, J.O., 2020. Coinfinder: detecting significant associations and dissociations in pangenomes. Microb. Genomics 6. https://doi.org/10.1099/mgen.0.000338

Wu, M., Johannesson, B., Geiker, M., 2012. A review: Self-healing in cementitious materials and engineered cementitious composite as a self-healing material. Constr. Build. Mater. 28, 571–583. https://doi.org/10.1016/j.conbuildmat.2011.08.086

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., Rosselló-Móra, R., 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31, 241–250. https://doi.org/10.1016/j.syapm.2008.07.001

Zhong, W., Yao, W., 2008. Influence of damage degree on self-healing of concrete. Constr. Build. Mater. 22, 1137–1142. https://doi.org/10.1016/j.conbuildmat.2007.02.006

Zhou, Z., Tran, P.Q., Breister, A.M., Liu, Y., Kieft, K., Cowley, E.S., Karaoz, U., Anantharaman, K., 2019. METABOLIC: High-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks (preprint). Bioinformatics. https://doi.org/10.1101/761643

Appendix

All my workflow and commands will be shown as below:

GENOME DOWNLOAD:

# Make the folder in which *Bacillus subtilis* genomes will be downloaded into

```
mkdir Bacillus_subtilis
cd Bacillus_subtilis
```

# Get genomes from NCBI

```
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/archaea/Methanococcus_maripaludis/assembly_summary.txt
genomeNum=$(grep -c "." assembly_summary.txt)
if [ $genomeNum -gt 700 ]; then grep "Complete\|Chromosome" assembly_summary.txt | cut -f20 > var.txt; else cut -f20 assembly_summary.txt | sed '1,2d' > var.txt; fi
for f in `cat var.txt`; do name=$(grep -w "$f" assembly_summary.txt | cut -f9 | cut -f2 -d'=' | sed 's/ /_/g' | sed 's/\//_/g' | sed 's/\:/_/g' | sed 's/)/_/g' | sed 's/(/_/g'); xx=$(grep -w "$f" assembly_summary.txt | cut -f20 | cut -f10 -d'/'); wget --tries=75 -c $f/$xx\_genomic.fna.gz; done
gzip -d *.gz
```

# Move raw data to RAW_genomes folder

```
cp -r Bacillus_subtilis/ RAW_genomes/
cd /Bacillus_subtilis/ RAW_genomes/
```

PROKKA workflow:

# Set up environment by enabling miniconda and perl

```
export PATH=/home/opt/miniconda2/bin:$PATH
unset PERL5LIB
source activate pangenome
```

# Run prokka

```
for i in $(ls *.fna); do echo "Processing $i"; prokka $i --locustag ${i%.fna} --outdir ${i%.fna} --quiet; done
```

ROARY workflow:

```
# Make a folder named ‘roary’ to store the .gff files

mkdir roary

# Move the .gff files from prokka output into roary folder via for loop

for i in $(ls */*.gff); do cp $i roary/$(echo $i | sed 's!/.*!!').gff; done

# Set up roary environment by enabling miniconda and exporting perl

export PATH=/home/opt/miniconda2/bin:$PATH
source activate pangenome
export PERL5LIB=/usr/local/lib/perl5/site_perl/5.22.0/

# Run roary and put the output into a new folder named ‘roary_tree’

roary -f ./roary_tree -e -n -p 8 -v -r -i 80 --group_limit 100000 ./roary/*.gff

# Now run the bespoke roary scripts

python /home/opt/roary_scripts/roary_plots.py roary_tree/accessory_binary_genes.fa.newick roary_tree/gene_presence_absence.csv
```

METABOLIC workflow:

```
# Change the filesname from .fna to .fasta

for i in $(ls *.fna); do mv $i ${i%.fna}.fasta; done

# Now enable METABOLIC on Orion Cluster

export PATH=/home/opt/miniconda2/bin:$PATH
source activate metabolic
export PATH=/home/opt/METABOLIC:$PATH

# Now run the METABOLIC
```

```
METABOLIC-G.pl -in-gn METABOLIC -o METABOLIC_OUTPUT
```

COINFINDER workflow:

```
# Enable coinfinder on Orion cluster

export PATH=/home/opt/miniconda2/bin:$PATH
source activate coinfinder-env

# Go to my directory ‘Dongliang_Li’ and download the data associated with the
coinfinder manuscript using "git clone" command

mkdir coinfinder-test
cd coinfinder-test
git clone https://github.com/fwhelan/coinfinder-manuscript.git

# Copy the gene_presence_absence.csv output and core-gps_fasttree.newick output to
coinfinder-manuscript folder

cp coinfinder-manuscript/gene_presence_absence.csv .
cp coinfinder-manuscript/core-gps_fasttree.newick .

# Now run coinfinder

coinfinder -i gene_presence_absence.csv -I -p core-gps_fasttree.newick -o output
```