ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

Ecole doctorale Écologie Géosciences Agronomie Alimentation

Alix Mas

Préparée à l'unité de recherche 6553 ECOBIO
Écosystèmes, Biodiversité, Évolution
Observatoire des Sciences de Univers de Rennes

Forecasting evolution of bacteria in a specialization context: a functional approach combining modeling, *in vitro* experiments and genomic analyses.

Thèse rapportée par :

Sara MITRI
Assistant Professor, Université de Lausanne, Département de Microbiologie Fondamentale (DMF)

George KOWALCHUK
Professor Doctor, Universiteit Utrecht, Ecology and Biodiveristy Department

et soutenue à Rennes le 27 novembre 2017

devant le jury composé de :

George KOWALCHUK
Professor Doctor, Universiteit Utrecht, Ecology and Biodiveristy Department / *Rapporteur*

Samuel CHAFFRON
Chargé de Recherche, CNRS Nantes, LS2N UMR 6004 / *Examinateur*

Thomas POMMIER
Chargé de Recherche, Ecologie Microbienne UMR 5557, INRA Lyon / *Examinateur*

Yvan LAGADEUC
Professeur, Université de Rennes 1, ECOBIO UMR 6553 / *Directeur de thèse*

Philippe VANDENKOORNHUYSE
*P*rofesseur, Université de Rennes 1, ECOBIO UMR 6553 / *Co-directeur de thèse*

Des étoiles plein la vie,
Du soleil et de la lune aussi

Remerciements

Thanks

First of all, I would like to sincerely thank the jury members for their interest in this Ph.D work and the time spent examining it. It is an honor for me to have your names here.

Merci Sara Mitri,
Merci George Kowalchuk,
Merci Samuel Chaffron,
Merci Thomas Pommier.

*

La recherche,
C'est un peu là ou nos rêves et le réel se confrontent, et parfois se rejoignent.
J'ai commencé à y apprendre la patience,
J'y ai ressenti de l'excitation; et de la frustration,
J'ai regardé les choses avec les yeux, les oreilles et l'esprit grand ouvert.
J'ai appris à accepter l'imperfection, car la vie l'est, imparfaite.
J'ai continué à rêver, par-delà les contraintes, pour ne pas laisser place à la résignation.
J'ai eu la sensation de baigner dans la synergie des pensées, des volontés et des rêves virevoltants de chacun.

Ces trois années furent un vrai chemin de vie,

Et je souhaite remercier, du fond du cœur, les gens qui m'ont accompagnés sur ce bout de chemin. Les personnes qui ont été présentes sur la course de fond, et les personnes qui ont été présentes sur le sprint de fin, qui m'ont apporté leur soutien au quotidien ou aux moments difficiles. J'ai de la chance de vous avoir auprès de moi.

Pour toujours, mes parents et ma famille. Vous êtes une ressource sans fin pour moi, un équilibre, un bonheur. Je vous embrasse de tout mon cœur.
Et spécialement, Merci Mam, je pourrais écrire une thèse entière rien que pour te remercier, de ta bienveillance, de ton soutien et du don de toi-même.

Merci à tous les gens qui ont eu foi en moi, et qui ont maintenu à flot la jauge si labile de ma confiance-en-soi.
Philippe, ton enthousiasme pour la Science est sans prix. Merci de ta présence continue, de tes encouragements, et de ta confiance. Merci aussi de m'avoir par-

fois ramené les pieds sur terre ("non mais t'as les fils qui se touchent là Alix!"), fait découvrir le vin (et accessoirement la recherche, hehe). Je suis contente d'avoir été un de tes pioupious. (le favori, mais on le dira pas à Nathan)
Yvan, merci pour ton encadrement et ta confiance depuis le master, et de m'avoir permis de parfois mieux comprendre le 'behind the scene' de la recherche. A quand la balade à poney?!
Merci Nathan et Kevin, qui avez su accepter et rire avec moi de l'ampleur de ma bulle, et de ma concentration "papillonnante"! Merci aussi pour l'aide en stats, vos encouragements, les discussions scientifiques, et les moments de haute voltige intellectuelle ('ouhhhou c'est qui le patron?!!')
Nathan merci aussi pour les nombreuses musiques partagées (et d'avoir failli me casser un bras un jour, sisi, j'avais un bleu même). Et Nivek, merci d'avoir presque réparé mes freins de vélo:), de m'avoir tiré à 30km en roller (record!) et pour ton soutien en fin de thèse. (et les brownies! mnomnomnom)
Merci Eve et Lorine, d'avoir rejoint le 216 un peu plus tard, et permit de rééquilibrer la part féminine de ce bureau (0,5 femme c'était pas assez)! Je vous souhaite une bonne fin de thèse les filles! C'était chouette, ces quelques années tous ensemble!
Et bien sûr merci à tous les quatre pour les poissons d'anniversaire! - devant lesquels j'ai parfois scotché de longues minutes..

Merci aux personnes de ma vie,
Merci Titi, de me faire rire, de me réconforter, d'être une âme-sœur(-cousine) qui est là depuis toujours.
Merci Flifloof pour tes lettres, tes petites attentions, et ta belle philosophie de vie.
Merci aussi à mes amies proches, les fifilles, c'est toujours ressourçant de partager du temps avec vous, et ces fou-rires comme j'en ai rarement ailleurs!
Merci Anne, pour ces années partagées (avec Gros Cookie) à l'appartement rue de Toulouse, je m'y suis sentie bien, j'ai aimé nos discussions, et ton soutien, empreint de réalisme qui m'a souvent fait aller de l'avant.
Sarah, tu es partie depuis un moment mais tu as été une vrai ressource quand j'étais une bébé thésard, nos longues discussions étaient toujours apaisantes ou fructueuses, et j'admirerai toujours ta force et ta persévérance! (Guèppendorf!)
Didi, tu es un vrai souffle de sympathie, de simplicité et d'énergie! C'est un plaisir à chaque fois que l'on se voit. Merci pour ton amitié, la virée en Italie, les chocolats italiens, les cafés, les discussions de tout et de rien partagées!
Merci Pampli, d'avoir apporté quelques temps un bel équilibre à ma vie, de

m'avoir permis de voir les choses avec d'autres yeux. A côté de combien de blagues ironiques suis-je passées?! -telle une poule devant une fourchette!
Merci Mewena, Guylaine, pour cette rencontre et ce lien particulier. Je suis heureuse de vous connaître.
Des pensées pour mes chères Paulines, je vous adore, et j'ai hâte de repartir en voyage avec vous!
Laure, Manue, merci de votre présence dans la dernière période bien difficile traversée.
Merci Fa, pour les moments partagés en fin de thèse, ça m'a permis de garder la tête hors de l'eau.
Merci Marie et Benji pour votre accueil chaleureux et ressourçant, je suis si excitée de partir voir les PPMR! A dream about to come true!
Gorenka, thanks girl, for those insightful evening pep-talks that helped much lately. I hope we keep in touch :)
Des pensées particulières aussi pour les personnes avec qui j'ai grimpé, couru, fais du yoga (merci Léa!), du roller, du hip-hop etc, qui sont devenus des ami(e)s au cours de ces trois ans.

Finalement, Merci aux chercheurs ayant participé à mes comités de thèse, Julie Jacquiery, Eric Petit, Alexis Dufresne, Damien Eveillard, Yvan Le Bras, Fabrice Not, Pierre Peterlongo et Achim Quaiser.

Merci Pietro de Anna, Marko Budinich, Shahrad Jashmidi, Erwan Corre et Xi Liu, Wesley Delage, Victoria Potdevin, Jean Coquet, et toutes les personnes avec qui j'ai pu travailler, longtemps ou ponctuellement, et avec qui j'ai aussi souvent partagé de nombreux moments sympathiques.

Merci à l'ensemble du laboratoire d'Ecobio, à ses chouettes doctorants, et notamment aux personnes qui ont pris part à la mise en place et à la dynamique des RJSE. Jolie expérience, comme tout le reste de cette thèse!

Et d'après une analyse statistique approfondie, il semblerait que je puisse aussi remercier Lindt pour les 360 tablettes de chocolat consommées au cours de ces trois années! Oups.

Bonne lecture!

Alix MAS

Forecasting evolution of bacteria in a specialization context: A functional approach combining metabolic modeling, *in vitro* experiments and genomic analyses.

Contents

Part I

Opening Section: Thesis'Background, Rationale and Objectives

Preamble

Since the oldest civilizations, humans have always felt a compulsion to understand the origins of things, how they unfolded in the past (the development of the universe, the emergence of life on earth), and how they might turn out to be in the future.

This later consideration has been even more compelling in the recent turns of events, when it was recognized that life conditions on earth were changing at an unprecedented pace. Indeed, environmental conditions are changing, fast, and put the human population survival at risk: the risk of not having enough space, food and water, and the risk of emergence or spreading of diseases. Be it the forecast of sea level rise, of temperature increase, of plantation yield or of infectious patterns, we want to be able to predict the direction, the rhythm and the consequences of such changes in order to protect life on earth. It sounds legitimate that a coherent way to sustain our presence is to understand and anticipate potential issues in order to prevent these issues, or to counter them. A crucial stake here is thus the phase of *prediction*. And, to be able to predict what is going to happen next in a system, it is essential to know how the system functions, and how it evolves.

It is a beaten track, but unfortunately it is also a reality: some of the most preoccupying issues we are confronted to are to know how we will be able to feed an ever-growing population (if possible in a sustaining way), and how we are going to avoid the spreading of infectious pathogens. It might not be obvious at the first glance, but underneath these questions, and most probably underneath their answers too, lay the science of Ecology. The study of organisms and their interactions with the environment, and notably the study of plants and microbes which are at the base of many of our resources (and fragilities) is a key. By taking a deeper look at the dynamic of theses organisms, by understanding how they evolve, how they interact and respond to environmental changes, we may be on a path to sustain life on earth.

Chapter 1

A background synthesis of Evolution

1.1 History and theory of Evolution

1.1.1 History of Evolution

In its broader sense, the term evolution designates any kind of gradual and accumulated modifications over time. It can be used to define the changes of any non-living or living system, behavior or even lines of thoughts.

More specifically in biology, evolution is considered as the observable changes of organisms (at any level, from genetic to community) along generation time, these changes being adaptive, as the organisms remaining are the ones better fit to their current environment, with a better survival and reproduction rate.

When we hear the word Evolution today, first thoughts of most are for Darwin whose theory (Darwin, 1859) is inevitable. Nevertheless, Evolution is much more intricate and advanced than his sole theory. It is an interweaving of paleontology and phylogeny, of (population) genetics and ecology, of molecular evolution, developmental biology, biochemistry, physics, and even mathematics and modeling. To understand evolution is to understand the dynamic of life, to comprehend the ins and outs of the system that constitute the realm of our biological world.

Historically the beginning of evolutionary theorization is often attributed to de Lamarck (in the early 1800). If since then Darwinism has largely supplanted his work, it is de Lamarck's ideas in essence that put a foot on the ladder of evolutionary thinking. For instance, he conceived that some individuals' particular features were better adapted to the environment, and that adaptive characteristics were kept for the next generation (Ghiselain et al, 1994). These ideas are still the backbone of currently accepted theories, or are being revisited through new discoveries, such as epigenetic mechanisms which are heritable molecular mechanisms impacting gene expression and function (and thus phenotypes) without changing the DNA sequence (e.g. Richards, 2006; Holeski et al., 2012; Kawakatsu et al., 2016).

Thus, influenced by the previous work of de Lamarck, and of Malthus who formulated that populations were being limited and constrained by environmental resources (Malthus, 1826), Darwin and Wallace offered simultaneous and complementary work (Darwin & Wallace, 1858) leading to the current theory where, in a given environment, amongst the existing variants in a population, some of them would be advantaged by their particular variation and thus selected through Natural Selection (term coined by Darwin, 1859). Natural Se-

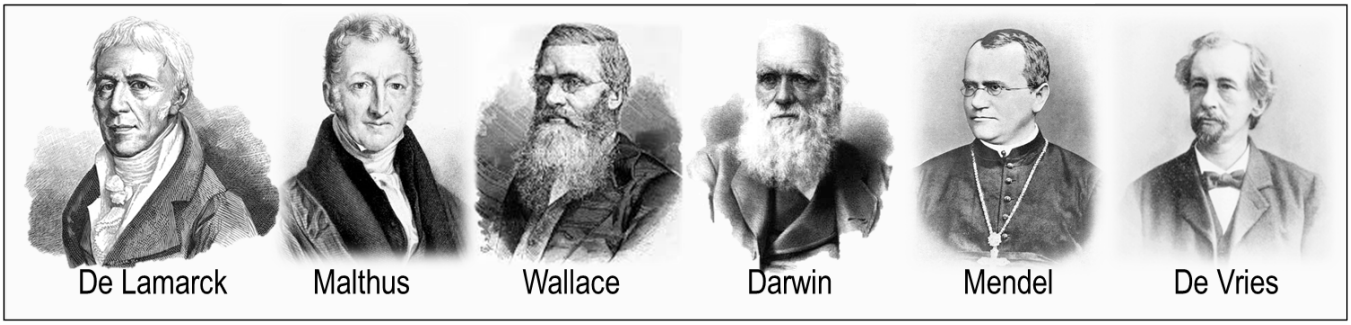


Figure 1.1: The great figures behind the early theory of evolution

lection is a force leading to the conservation of phenotypes that survive and reproduce more, logically leading to the eviction of currently under-adapted concurrent phenotypes, as the environment offers limited resources.

In a few words, some organisms which are different are conserved because the differences they harbor confer them with some advantages to exploit or survive in the environment. Thus adaptations across generations are due to the preferential retention (Natural Selection) of these variants that exist in a population. This survival of the fittest (Herbert, 1863) lead to the creation of the term and concept of *Fitness*, which is explained in section 1.1.2.

In parallel, the work of Mendel contemplated the functioning of inheritance processes (Bowler, 1989), based on what were called 'genes' afterwards. This work was rediscovered by de Vries (de Vries et al., 1901), who gave descriptions of mechanisms behind the arising of new variants: de Vries believed that the appearance of species was due to the appearance of sudden discontinuous variation that he called *mutations*, and that these mutations were heritable to the next generations (de Vries et al., 1901). These explanations (inheritance process through genes, mutations) were lacking in Darwin and Wallace's theories.

Selected organisms are thus the ones carrying specific gene variants (which were generated by mutations) that made their phenotypes advantageous, ensuring a better inheritance. These two approaches, the Darwin/Wallace-mechanisms of natural selection and the Mandelian inheriting mechanisms beneath selection, were merged together, and gave rise to the study of "population genetics" (Fisher, Wright, Haldane). Population Genetics also encompasses the notions of genetic drift and gene flow (presented in "Forces at stake", section 1.1.3) and constitute the core of the theory called Modern Evolutionary Synthesis (or Neo-Darwinism, Huxley J.S. 1942; Bock Walter, 1981; Bowler 1989). The Modern Evolutionary Synthesis integrates developments made in evolution until the mid-20's such as the notions of speciation, ecological niches, gradualism, and the role of genetic

drift.

Still today, limits of what circumvents evolutionary theories are being continuously pushed, by new findings and new understanding. As mentioned, recent key evolutionary discoveries such as epigenetics and its inheritance, but also niche constructions and eco-evolutionary feedback, developments in co-evolution, plasticity and multi-level selection (Wade, 2011; Laland and Sterelny, 2006; Pigliucci, 2006; Danchin et al., 2011; Pigliucci and Finkelman, 2014) are being gathered in a still unconventional and progressing theory named the Extended Evolutionary Synthesis (Danchin et al., 2011; Pigliucci and Finkelman, 2014; Laubichler and Renn, 2015).

1.1.2 Concepts underlying Evolution

Several concepts are inherent to the Evolutionary Synthesis: the first is the unit of (natural) selection : the phenotype, the second is the measure of natural selection called *fitness* (i.e. survival & reproduction), and the third is a reason behind the particular evolutionary trajectories existing, known as *trade-offs*. These preponderant notions in evolutionary ecology are developed here.

Phenotype

The phenotype of an organism is classically described as the morpho-phenologic features characterizing the organisms.

These features are gathered under the term trait, which consider any surrogate of organismal performance (Violle et al., 2007). More particularly, Violle et al. (2007) defined *functional* traits as being the "morpho-physiophenological traits which impact fitness indirectly via their effects on growth, reproduction and survival, the three components of individual performance".

The phenotype is considered to be the unit of selection in evolution. It is because a phenotype is more adapted to an environment than other phenotypes, that it will be able to reproduce more and extend in the population through natural selection.

Even though a phentoype refers to the ensemble of traits expressed by an organism, most of the time the study of the phentoype is narrowed down to the study of one outstanding feature of the phentoype only, such as the body coloration of the organism (e.g. Hoekestra, 2011) or the capacity of the organism to produce or utilize given metabolites (e.g., detoxyfying enzymes, carbon sources).

The phenotype of an individual can be represented as a multi-dimensional

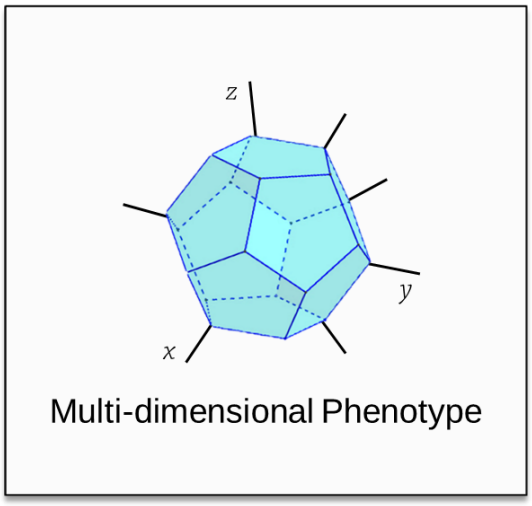


Figure 1.2: **The Phenotype seen as an hyper-volume in multidimensional space**. Each ridge of the volume is a trait of the phenotype. Ridges can take different values. Some of the traits may be related to others (functional trade-offs) and their variation will influence the variation of related trait. To ease the representation the volume is very homogeneous, but every side of the volume could be shaped and sized differently.

volume (Figure 1.2), where each dimension is one feature of the phenotype (a trait, which can be morphological, phenologic, functional, etc) which can be expressed in different values (for example, one ridge of the volume can be the coloration of an individual, varying from beige to brown). Some of the features may be tightly related to others (functional trade-offs). This representation helps understanding that a phenotype is multidimensional (morpho-physio-functional and life history traits) and that each of these features may express some variability.

The notion of individual phenotypes as being the unit of selection was challenged over time. Lately the concept of extended phenotype sustains that an organism cannot be determined by its own physico-chemicals parameters only (biological processes) but that it should also encompass the effects of its biotic and abiotic environment (e.g. symbiotic interacting species which necessarily modify the phenotype, or epigenetic modifications directly entailed by the environment). Related to the notion of extended phenotype, the concept of holobionts emerged, and suggests that an organism studied by itself is considered as an incomplete system, and one should also integrate the other species associated, as the whole is forming an ecological unit (Theis et al., 2016; Rosenberg and Zilber-Rosenberg, 2016).

Fitness

As mentioned above, the formulation of Darwin's theory lead to the creation of the concept of *Fitness*. Fitness was determined as a measure of natural selection by J.B.S Haldane in his "Mathematical Theory of Natural and Artificial Selection" (series of 10 papers published in 1924-1934, J.B.S Haldane). Concretely it describes the reproductive success of a genotype (or a phenotype) by measuring the contribution of this genotype to the pool of genes of the next generation.

If an individual of a population carries mutations that increase its reproduction in a given environment (i.e. that enhance its contribution to the next generation) the genome carrying this mutation will expand in the population along generations. It is naturally selected for. Thus, in its simplest perception, mutations (genotypic and phenotypic variants) that increase in frequency in a population suggest that they are selected for, and therefore that they confer fitness benefits.

The notion of fitness has been extensively studied, notably through theoretical work based on mathematics (e.g. Game Theory, Maynard Smith 1982) and sometimes with references to social and economic sciences (Tragedy of the Commons, Hardin 1968). It was used to explain the persistence of biological systems that diverged from classical phenotype-based natural selection, such as altruism.

Conceptually, fitness is often represented as a landscape (Wright, 1932), with peaks and valleys, and where peaks are stable high fitness geno- or pheno- types (Figure 1.3). Nevertheless, this representation has recently been criticized (Kaplan, 2008) and the notion of landscape entails a static perception of fitness which is too far from the dynamic reality of living systems, as the environment and organisms themselves are ever-changing and modifying each-other (Steinberg and Ostermeier, 2016).

Today the fitness measure is the backbone of almost every evolutionary concept.

Trade-Offs

Another important notion in evolution is the concept of trade-offs. The abstraction of a trade-off is a situation that involves losing one quality, or quantity of something in exchange for gaining another quality, or quantity of something else. In biology and evolution, the concept of trade-off is omnipresent, as both the environment offers limited resources, and the organisms have limited physiological capacities. Thus choices have to be "made" to optimize the use of available

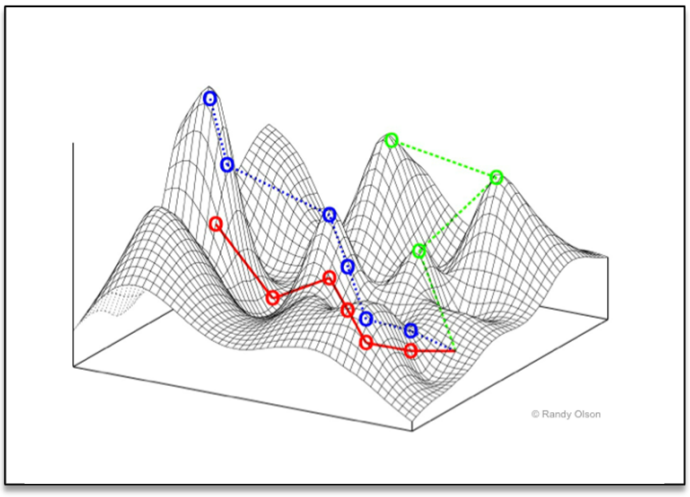


Figure 1.3: **Fitness Landscape**. The set of all possible genotypes, their degree of similarity, and their related fitness values is represented by fitness landscapes. The fitness is the 'height' of the landscape. Peaks in the fitness landscape are local or global fitness maxima. The consequence of natural selection is to maximize the individual fitness, thus peaks represent the stable attractors of the fitness landscape. Genotypes which are similar are said to be close to each other, while those that are very different are far from each other. The arrows represent various paths that the population could follow while evolving on the fitness landscape from migrations and also possibly mutations.
Image retrieved from the work of Randy Olson - CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=32274330

intrinsic and environmental resources (in line with fitness increase). As finely defined by Stearns (1989): "Trade-offs represent the costs paid in the currency of fitness when a beneficial change in one trait is linked to a detrimental change in another".

Trade-offs can be perceived at the life-history strategies, behavioral, functional, metabolic and genetic levels (Stearns, 1989). For example a simplistic and common example of trade-off is the possibility of making one big and resistant offspring (more chances to survive, but if it dies all is lost), or several smaller offspring (less resistant, but higher chance that at least one survives). Each functional trait (i.e any characteristic that influences performance and reproduction, Violle et al, 2007) is determined by its position on the spectrum of trade-offs possibilities.

The notion of fitness and trade-offs are thus tightly interrelated and fundamental to evolution.

1.1.3 Forces at stake in Evolution

Classical elements driving Evolution

Evolution is often seen as being lead by four major forces: Natural Selection, Mutation, Random Genetic Drift and Gene Flow. Depending on the context, these forces will be more or less preponderant, and can happen simultaneously. However these elements are presented altogether but refer to very different mechanisms. Half of them represent mechanisms at the origin of the apparition of variants in a population, and the other half represent mechanisms of selection of the variants.

- Mutations and Gene Flow are causes that modify the gene content in a population. Precisely, the sources of variation of the DNA present in individuals in a population can emanate from 1) either a modification appearing directly on the genome of individuals : **mutations**[1], gene duplication or insertions-deletions, and/or 2) from the appearance of a new gene or set of genes in the population via the migration of individuals (i.e exchange of genetic material between two populations) which is called **gene flow**, or admixture.

[1] see the section dedicated to Mutations and their effects

In parallel,

- Natural Selection and Genetic Drift are forces that cull out variants of a population.

  Natural Selection is the process by which some variants will expand in a population because they survive and reproduce better than other variants thanks to some specific feature (variation) they carry which makes them better adapted to their actual environment than other individuals. As a result, the advantageous variations are passed on at a higher frequency than less advantageous features.

  On the other hand **Genetic Drift** refers to a random selection, a random change in allele frequencies of a population, from a generation to the next, and is inversely proportional to the population size.

  While natural selection 'purposely' increases fitness, genetic drift operates at random, and will only occasionally confer patent, random, fitness benefits. The selection (or absence of-) for and against existing variants will usually happen through both Natural Selection and random Genetic Drift simultaneously, but according to the population size considered, either genetic drift (for small populations only) or natural selection will be the main force driving the conservation of certain alleles/variants over others. Still, in a situation where natural selection is preponderant, genetic drift cannot be excluded and some variants will be selected stochastically too.
  Anyhow, both are source of gene-frequencies variation at the population level.

An additional phenomenon can also lead to the 'passive' selection of a mutation: it is genetic hitchhiking. Genetic hitchhiking happens when two genes, one carrying a mutation conferring beneficial effects, the other gene carrying a neutral or nearly neutral mutation, are linked. Indeed, if the two genes are physically linked (close to each other on the chromosome) or functionally linked (same operon), it is possible that the selection of the beneficial mutation will drive to the 'passive' selection of the linked gene carrying a non-beneficial mutation (Figure 1.4).

Molecular mechanisms and effects of mutations

The mechanisms behind evolution are multiples and act either as sources of selection (which, as we explained already, are mostly natural selection and genetic

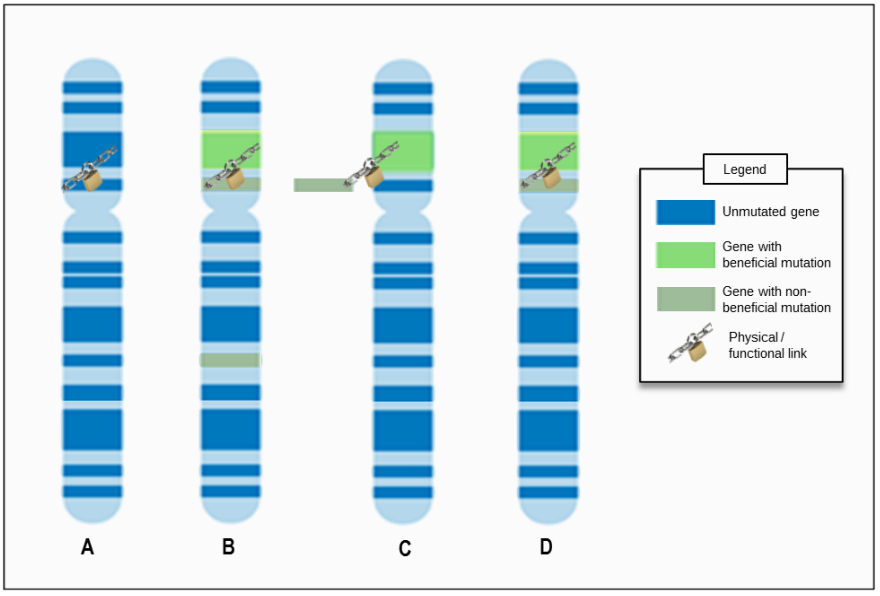


Figure 1.4: **Schematisation of genetic hitchhiking.**
(A) represents an unmutated chromosome, the deep-blue bands are genes, **(B)** mutations appear : green bands are beneficial mutations, grey bands are non-beneficial mutations, **(C)** the beneficial mutation is selected along evolution, and physically linked genes carrying non-neutral mutations are also passively selected, **(D)**, the selected genome thus carries actively-selected-for mutations, and other mutations linked via genes interactions. Note that other non-beneficial mutations that emerged were not selected for (grey band in the lower part of the chromosome).

drift) or they act as sources of modifications of the genome.

The sources of variations of the genomes themselves are quite diverse, and the phenomenon at stake are being better and better understood. Several of them can happen simultaneously, some will be specific to diploids, haploids, eucaryotes or procaryotes.

After a quick recall on genes structure of procaryotes (Box 1) the type of mutation that can emerge and their potential consequences are described below.

Recall on genes:
A gene is a coding sequence of nucleotides of the genome. It is transcribed from DNA into RNA, which can either be non-coding (ncRNA) with a direct function, or an intermediate messenger (mRNA) that is then translated into protein. The beginning and the end of a genes are determined by a start-codon and a stop-codon respectively. Genes are preceded by promoters which initiate transcription, and the expression of a given gene can be modulated by regulatory sequences (enhancers and silencers,Wiper-Bergeron and Skerjanc 2009). Genes can also be pooled together in operons, which are sets of genes (and related regulatory sequences) encoding for particular functions. Most of the times genes belonging to an operon are physically close on the genome. Importantly, genes have pleiotropic effect, i.e. they are implied in different functions.

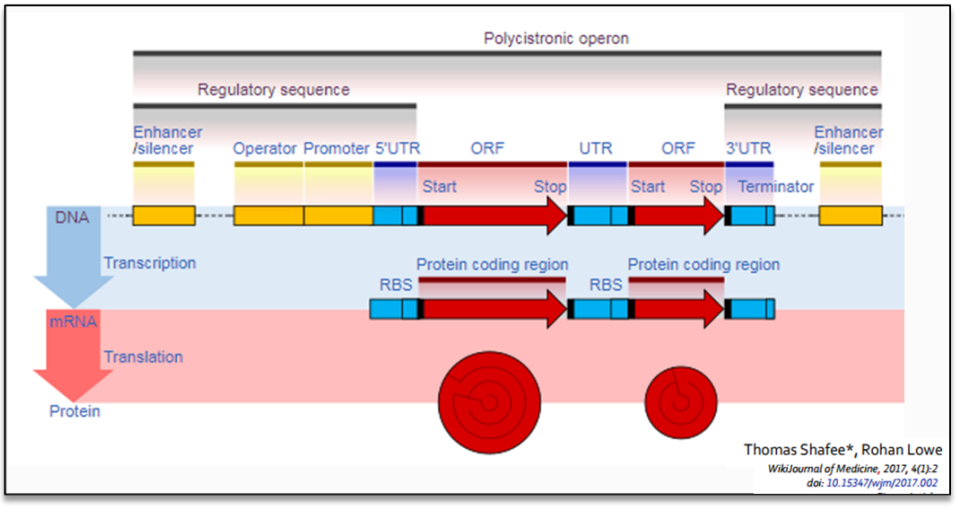


Figure 1.5: **Structure of a prokaryotic gene, from (Shafee & Lowe 2017)** Regulatory sequence controls when expression occurs for the multiple protein coding regions (red). Promoter, operator and enhancer regions (yellow) regulate the transcription of the gene into an mRNA. The mRNA untranslated regions (blue) regulate translation into the final protein products.

Box 1. Recall on procaryote genes structure.

A mutations is a punctual change in the DNA sequence, resulting from the addition, the deletion or the substitution of nitrogen bases (Adenine, Guanine, Cytosine, Thymine). Mutations are the result of incorrect base pairings during cell division. Two types of mutations are usually considered:

- **Point mutations** such as nucleotide substitution or nucleotide inversion. Substitution of nucleotides are also called Single Nucleotide Polymorphism (SNP) and can generate either synonymous or non synonymous mutations.

  Due to the redundancy of the genetic code (Figure 1.6), different triplets of nucleotides can encode the same amino-acid, consequently the amino-acid sequence is not modified, and the mutation is presumed to have no effect on the structure and function of the crafted protein, we thus talk about synonymous, or silent mutations.

  On the other hand if the new triplet encode a different amino-acid, the mutation can modify the amino-acid sequence, with small or important biological effects on the phenotype of the organism. Furthermore particular modifications can also give rise to stop-codons, leading to a premature stop of the translation into the amino-acid sequence and thus impeaching the creation of the protein.

  However it is exciting to note here that it was recently demonstrated that such said 'silent mutation' could also imply effects on protein folding and even fitness (Bailey et al., 2014). Indeed, as detailed by Goymer (2007) synonymous mutations can affect transcription, splicing, mRNA transport, and translation (any of which could alter phenotype, rendering the synonymous mutation non-silent).

- **Frameshift mutations** such as insertions or deletions. Frameshift mutations are called so because the insertion/deletion can change the reading frame by changing the grouping of the codons. If there is indeed a frameshift it can result in a different translation than should be (the revised codons after the mutation will code for different amino acids) and lead to abnormal and usually dysfunctional polypeptides (proteins), especially if a stop-codon appears prematurely.

Mutations can appear anywhere on the genome. On coding sequences, on regulatory sequences, on the promoting region of the gene or at the end of its sequence. Depending on the location of the mutation more or less strong

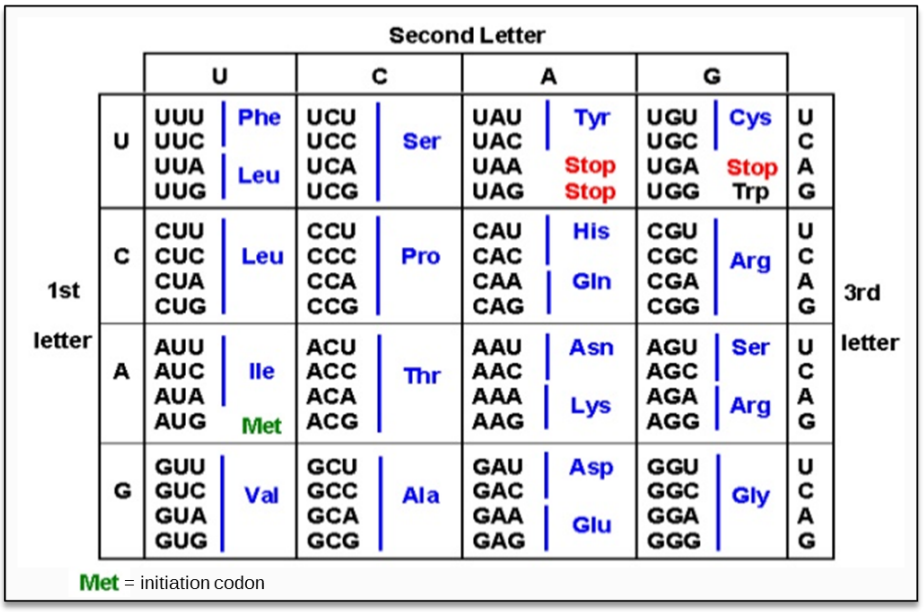


Figure 1.6: **Table of the genetic code.**
Each mRNA codon codes for an amino acid with the exception of UAA, UAG and UGA (the three stop codons). The AUG codon is also a start-codon. Several nucleotide combination can encode the same amino-acid, this known as the redundancy - or degeneracy, of the genetic code.

repercussions will come into effect. For example, if a mutation appears on a cis-regulatory coding region, then the level of expression of the gene will vary, but if the mutation appears on the promoting region of a an operon, then the whole operon may be shut off. (Gompel and Prud'homme, 2009).

At yet another level, according to the effect of the mutation on the phenotype they can be identified as advantageous mutation if they lead to a fitness increase of the phenotype carrying it, deleterious if they are counter-selected, or neutral if they have no visible effect. This way, a stop codon may be considered as beneficial as long as it entails a fitness benefit of the organism carrying it.

The importance of interactions in evolution

About a century after the Natural Selection theory was proposed, the Red Queen Hypothesis (RQH) emerged (Van Valen, 1973). This theory sustains that species of organisms must constantly adapt to survive when facing ever-evolving opposing organisms (resulting in an arm's race where each actor of the interaction tries to overcome the other).

Behind this hypothesis is recognized the extensive role of species interactions in evolution. Indeed, (direct and indirect) interactions are omnipresent between

organisms, and if they are stable enough along generations time, they will significantly impact evolutionary dynamics. It was first acknowledged through the term 'coevolution', created by Ehrlich and Raven (1964), to define the situation when the evolutionary trajectories of two species is influenced by their interaction.

The arm's race generated by opposing species and inducing their coevolution, in the RQH, is only an example of how antagonistic interaction can shape evolution, but of course the concept is valid for any kind of interactions existing, from antagonistic to beneficial. For example the coevolution stemming from beneficial interactions between butterflies and plants (Ehrlich and Raven, 1964), or as more commonly studied the host-parasites interactions (e.g. Schulte et al. 2010), and evidently also for competitive interactions (Connell, 1980). The coevolution between organisms is ubiquitous, and co-evolved couples of organisms can be found both within or across biological kingdoms.

From a mechanistic point of view, the co-evolutionary phenomenon can be both straightforward and complex:

Straightforward because it is intuitive to consider that interactions such as competition, parasitism or mutualism will affect the direct survival, growth and reproduction of the organisms implied, and the organisms that are most adapted to the interactions (either to reduce or enhance it) will be the ones conserved by natural selection.

On the other hand, the coevolution process can be more subtle, in situations where the interactions are indirect, for example through the sharing of metabolites. In this case, the coevolution in place may be harder to detect as the visible phenotypes of organisms were not necessarily modified. Yet their genotype and metabolism underwent changes that can be detected through -omic analyses. This type of indirect coevolution was notably theorized in the Black Queen Hypothesis (BQH, Morris et al. 2012a). This evolutionary hypothesis gives explanations on the evolution of dependencies, through adaptive gene loss in free-living organism. The name refers to the card game Hearts, where the winning strategy is to avoid the Queen of Spade. In the BQH context, free living organisms "avoid" having a function to optimize their adaptation to the environment (these organisms are called beneficiaries). This loss of function is permitted because other organisms in their close environment (the helpers) publicly provide for the function. The underlying mechanism of the loss of function in the BQH is genome reduction through gene loss. This type of evolutionary process is of major importance in the context of coevolution, as the beneficiary

species develops a strong dependency on the helper species.

In fact any biotic interaction that results in a variation of the fitness of organisms implied will be subjected to natural selection, and can thus drive co-evolution. In cases of beneficial interactions, the notion of co-evolution is closely related to the one of specialization as co-evolving organisms specialize toward each-others. It is perceivable for bird's beak shapes that are uniquely fit to consume the nectar of one type of flower, and flowers that express traits favoring the pollination by specific birds instead of other organisms (Castellanos et al., 2004).

Living organisms are known as the biotic constituent of the environment, thus organisms specialize to this component through coevolution as well as they would specialize to an environmental physico-chemical parameter (abiotic constituents).

1.1.4 Relevant complementary theories of Evolution

What is commonly accepted in the scientific community is that mutations in a population arise randomly in organisms independently of the environment, some of these mutations will confer phenotypic advantages to the organisms in the current environment, and thus will be selected. These selected variants will expand in the population as they can reproduce better, and this is how the population evolves. This is Natural Selection and Mendelian mechanisms in their most classical view.

Nevertheless, other theories and concepts have emerged and are still of actuality. A few of the most interesting are the following:

Genes, the substratum of evolution?

While until the mid-sixties the focus was made on the evolution of organisms, George C. Williams in his book "Adaptation and Natural Selection" and subsequently Richard Dawkins in "The Selfish Gene" (Dawkins, 1990) proposed that genes themselves (instead of whole organisms) should be the core units considered when dealing with evolution. This gene-centered view of evolution states that the conservation and expansion of specific genes is at the core of the evolutionary dynamics. It considers phenotypes and fitness of individuals only as a way (a vehicle) to enable the persistence of the genes they carry. This theory permits to elucidate, *inter alia*, the altruist behaviors sometimes expressed by organisms of a group.

However, in a more familiar tone, genes are only the ingredients of the cake,

and a gene by itself will not be selected, it is the effects it induces at the organism level that will (or will not) undergo Natural Selection.

Revisiting the unit of Selection

As presented in the section on "Phenotype" (part 1.1.2), in the light of the latest discoveries, the unit of selection for evolution through Natural Selection was revisited. Initially and for a long time, the unit of selection was the individual's phenotype, their morphological or functional traits. But as it was understood that those traits could modify the environment and also depended on the interactions organisms are involved in, more accurate definitions emerged. Indeed, the notion of Extended Phenotype (Dawkins, 1999) suggests that selection acts on a wider domain than just individual's phenotypes, by including their interactions with the biotic and abiotic components of the system. These more biologically accurate approaches are also more complex to encompass, and may blur the delineation of selection.

Neutrality of mutations and non-adaptive phenotypes

Alternatively, in 1983, Kimura proposed a "Neutral Theory of Molecular Evolution" (Kimura, 1983). In this monograph, it is explained how, at the *molecular* level, the vast majority of mutations are not selectively beneficial but only neutral or slightly deleterious (circum-neutral, as deleterious mutations are purged immediately). Thus most of the phenotype modifications observed are in fact non-adaptive changes. This puts the notion of adaptationism into perspective.

Mechanistically, as suggested by Kimura (but also in King and Jukes 1969), this theory relies upon the redundancy of the genetic code (Figure 1.6), according to which many mutations (i.e. changes at the molecular level) can arise without influencing the fitness of organisms (but see Bailey et al. 2014). If they have no beneficial effects on the phenotype [for a given environment], mutations cannot be the leverage for natural selection, and therefore genetic drift (defined in section "Forces at stake") is the major force driving the retention of mutations and evolution.

To wrap up, evolution encompasses behaviors, phenotypes, functional traits, physiology, genetics, epigenetics, interactions and goes behind what is commonly taught today. As long as discoveries will be made in any of the fields of biology (and others), the theory of evolution will keep evolving itself. Evolution is an ubiquitous phenomenon, and as assessed by Dobzhansky: " Nothing in

biology makes sense except in the light of evolution " (Dobzhansky, 1973). It can be added that things make much more sens in biology (and maybe especially in evolution) when they are contextualized and reintegrated to their whole system. Necessarily to study a system it is easier to detail each of its component (genetic, physiological, functional, interactions with the environment, etc) but the need to bring back everything together is still essential to enable an overview of pervasive phenomenon such as evolution.

1.1.5 The study of evolution

Studying evolution, from then to now.

Paleontology

During the first 150 years, the study of evolution was essentially based on the investigation of surrounding species (naturalism) and of traces of ancient species via fossils (paleontology). Studying evolution was studying the rise and extinction of species, it was inferring phylogenetic trees based on the similitude of morphological or phenotypical traits in order to determine common ancestors (see Figure 1.7, A). Such a retracing of the history of species over long period of time is refereed to as macro-evolution and is still of actuality.

Molecular Biology

However the understanding and study of evolution has known an extensive shift at the discovery of DNA (around 1870). The advent of molecular biology (1940's) allowed for an increasing comprehension of genome modifications and heredity processes. Instead of only observing the outcomes of evolution 'once it had happened' (e.g. morphological features), we can look deeper into the origins behind the differences of features, in order to understand the possible underlying molecular mechanisms allowing for evolution.

Sequencing

The scrutiny of DNA (and thus Evolution) has known another huge leap at the development of genome sequencing (in the 70s), mostly with the apparition of automated whole genome sequencing (90s). Being able to retrieve whole genomes of organisms (but also of populations or communities) gave another dimension to our perception of life kingdoms and their relatedness (Figure 1.7, B) and, importantly, mutations started to be better understood.

The rise of -omic approaches (DNA, RNA, proteins scales study) enlarged our comprehension of Evolution, by enabling the study of focused mutations on single organisms as well as the study of co-evolution of species within commu-

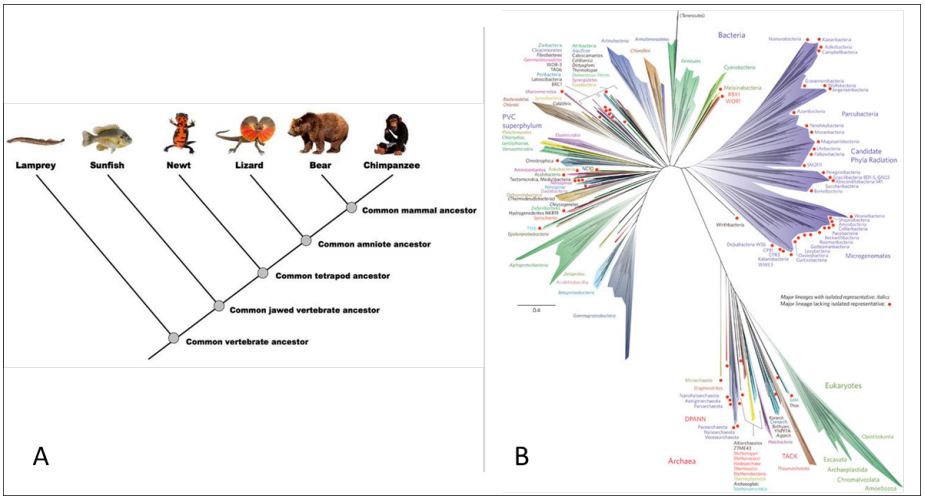


Figure 1.7: **Early and recent Trees of Life.**
(A) Tree of life determined in the early study of evolution. The comparison between organisms was principally based on their morphological traits. **(B)** Recent (2016) view of the Tree of Life as determined through analysis of DNA sequences of organisms. From Hug et al. (2016). The bacterial kingdom which was barely considered initially is now preponderant in the tree.

nities, thereby showcasing the fact that evolutionary events (can) occur at every level of ecological study.

Despite the tremendous advances genetic allowed for, the genome is only a part of an organism, and it cannot explain everything. The DNA sequence carried by an organism is a potential proxy of what an organism can be (Goldman and Landweber, 2016), because a genome will be expressed differently over time, over situations, over events. The environment, biotic and abiotic, has influence on both the genetic, metabolic, functional and behavioral scales of any organism, thus on the phenotype, which is determinant of evolution.

Evolutionary Experiments

Alongside with the development of new technologies to study DNA, it appeared that evolutionary events were happening not only over geological times but could also occur on a timescale of centuries, years or months (e.g. the evolution of plant domesticaion and of bacterial populations, as presented in Gaut 2015). Subsequently, the use of *in situ* and *in vitro* evolutionary experiments greatly complemented the comprehension of evolutionary dynamics and mechanisms. Since the first experiments, dating back to the end of the 20^{th} century (e.g. Dykhuizen 1990; Lenski et al. 1991; Travisano and Lenski 1996), the enthusiasm for *in vitro* experimenting has been restless (Riley et al., 2001; Barrett et al., 2005; Nilsson et al., 2005; Maughan et al., 2006; Gravel et al., 2011; Brockhurst

and Koskella, 2013; Bailey et al., 2014; Jerison and Desai, 2015; Ross-Gillespie et al., 2015; Tenaillon et al., 2016; Bono et al., 2017; Finn et al., 2017). One of the most cited and most dense work that can be mentioned here is the LTEE (Long Term Evolutionary Experiments realized in Pr. Lenski's laboratory) which followed the parallel evolution of several populations of *E.coli* along 30 years (to date). This work triggered cascades of parallel researches.

Evolution is thus faster than previously assumed, and it is now common to study the evolution of bacteria on relatively short-term experiments (as in Blank et al. 2014; Toll-Riera et al. 2016).

Modelling Evolution

Modelling in evolution is omnipresent. Be it for the reconstruction of phylogenetic trees to determine species relatedness or traits emergence, or be it modelling of population dynamics in a context of predation (Lotka-Voltera types of Models), or again the multitude of simulations of evolutionary trajectories on fitness landscapes (Lobkovsky et al., 2011; de Visser and Krug, 2014; Aguilar-Rodríguez et al., 2017), a high quantity of models exists. Everyday models for the study of evolution are refined and integrate transcdisciplinary concepts progressively developed (e.g. Morozov 2013; Louca and Doebeli 2015; Ribeck and Lenski 2015; Budinich et al. 2017). Modeling is evidently of great interest to overcome the complexity and the limits (technical, ethical, and mostly temporal) of simulating evolution in laboratories.

Studying evolution at different level of integration.

Temporal scales Evolution can be considered at different timescales: the 'geological' timescale where we consider the rise and extinction of species (sometimes defined as macroevolution), and current timescales, where we study the local functional or genetic changes a population or a species is undergoing, over a few generations (microevolution). This micro-evolution and the apparitions of mutations can be much faster than what was admitted until recently (Yoshida et al., 2003; Hairston et al., 2005; Schulte et al., 2010; Matthews et al., 2011; Maddamsetti et al., 2017). Microevolution is also interesting in that it can be observed at the various levels of integration considered in ecology (Figure 1.8).

Biological scales Ecology considers the interactions between the organisms and their environments, and here 'environment' can be both biotic and abiotic, meaning that it can refer to physico-chemicals parameters only, and to other living organisms too. The tiniest level considered is the one of genes and genomes, which is the source of individual differences between organisms. The next level, and

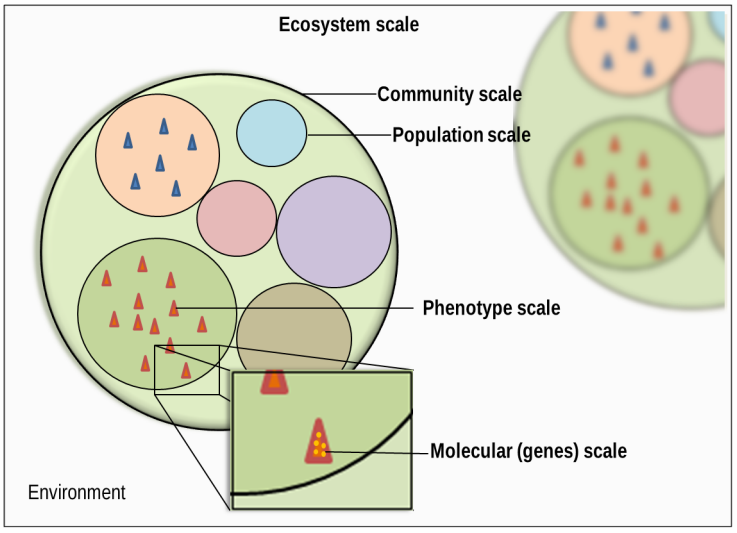


Figure 1.8: **Ecological levels of integration.** The various level of ecological systems are represented (genomic, individual(=phenotype), population and communities and ecosystem. The functional level is also considered in ecology, it corresponds to the activity and roles in the ecosystem expressed by an organism or a population.

a very classic one, is the phenotypic level, which has long been the sole system studied in evolution. The phenotype corresponds to the features expressed by an organism. These traits can be physiological, morphological, even behavioral, and are often presented as functions of the organisms (see also functional traits, in section 1.1.2). In between genetic and phenotypic levels, can be considered the metabolic level, which is usually not considered as a level by itself but can still be subjected to evolutionary forces in action. Classically, above the individual phenotypical level is the populational level, consisting of a group of individuals of the same species, living in the same spatial and time frame. Involving various populations leads to the level of community, which considers the present populations in a given environment. Finally, the wider level considered in ecology is the one of Ecosystems, which integrates all of the previous level parameters.

Let's emphasis here that, as for Matryoshka ("Russian") dolls, each smaller level is embedded into the next one. But considering a more functional point of view, these levels have to be regarded as being 'interwoven' as they have reciprocal effects on one another. Each level studied on its own will give information, but they will never make as much sense as when they can be merged together. For example, some metabolic constraints can forge genetic changes that would be perceived as random if only considering the genetic level. And respectively, some genetic features can influence higher levels such as the population level: if a gene is implied in shared functions, such as biofilm production, then these

genes will influence the dynamic of the population. So to understand evolution is to understand the dynamic of living systems from their molecular state to their 'interaction' state, taking into account the reciprocal action between these different levels.

1.2 Evolution and the Bacterial World

Invisible for a long time, the empire of bacteria started to be recognized along the 19th century, notably through the study of diseases. Since then, bacteria have been categorized as ubiquitous and are supposed to represent a colossal cumulative biomass on earth (Whitman et al., 1998). Bacteria show a large panel of tolerance to various environments and are found to survive even in the harshest conditions. Mostly known by the general public through the prism of disease, bacteria nonetheless take part in many diverse biogeochemical activities and ecosystemic processes as they contribute to the carbon and nitrogen cycle, to the recycling of organic matter, and to the detoxification of polluted environments (e.g. Balser et al, 2000,Schimel et al. 2007; Fuhrman 2009; Venail and Vives 2013). Because of this prominence, micro-organisms are now taking an increasing importance in the field of ecology, and macroecology theories are transferred and propagated toward the microbial world (e.g. Beaumont et al. 2009).

Bacteria are also of great interest in our current civilization: they are used in biotechnologies for the production of molecules used in pharmacology, they are used to increase the yield of plantations, for wine and cheese production, for waste degradation (sewage and industries, oils) and importantly, they are used as model system for fundamental and medical research.

1.2.1 Advantage for the study of evolution

On a more fundamental level, bacteria are also a tremendous source of investigation of biological and ecological processes, and particularly of evolution. Due to their genetic, functional and behavioral specificities, bacteria are great model organisms for scientific research. Because of their small size, malleability and fast generation time, it is somewhat easy (and more ethically justifiable) to cultivate bacteria under laboratory conditions, and to experiment on them. Their large population size and the short generation time they express (i.e. about 20 minutes for the classically studied *E.coli*, according to Ludewig and Fehlhaber 2009) enables evolutionary studies that would take years or decades otherwise.

Even though the major part of the bacterial kingdom cannot be grown under laboratory conditions, there has been great improvements regarding the study of bacteria: the development of sequencing and -omics (e.g. Hall 2007; Vandenkoornhuyse et al. 2010), to investigate bacterial diversity and genomes, but also the development of visualization tools which became so performing that it is now possible to observe distinctly a single bacterium.

One of the drawbacks of studying evolution on bacteria in the laboratory is that most of the bacteria are not actually cultivable at will (e.g. Whitman et al. 1998), there is thus a potential risk that the knowledge accumulated is biased toward the cultivable portion of bacteria, if for example they present different mechanisms, or lack mechanisms present in other bacterial populations.

1.2.2 Evolutionary mechanisms distinction

Bacteria can live freely in the environment. They are also involved in a spectrum of biotic interactions spanning from mutualism to pathogenic, passing by commensalism and ammensalism. These interactions will not only influence the ecology of the organisms implied, but also their reciprocal evolution (coevolution) in order to optimize the interactions.

The clonal mode of reproduction of bacteria is particularly interesting in evolutionary studies, as it implies that modifications are always transferred to next generation. Additionally, since the reproduction is asexual, fitness measures can be directly assigned to genotypes. Also, as their genome is haploid, the traits observed are the direct consequences of genes expressed (and their modifications). Yet one has to keep in mind that knowledge acquired for bacteria might not soundly be extrapolated to sexually reproducing organisms.

Another interesting mechanisms present in bacteria is the Horizontal Gene Transfers (Ochman, 2001): Some genes acquired directly from the environment, or from other organisms (bacterial or distantly related organisms such as animals or plants) and are naturally integrated to 'living' bacterial genomes through transformation, transduction, or conjugation mechanisms (Ochman, 2001). This bacterial recombination is frequent (McDaniel et al., 2010; Koonin et al., 2001), and acquired genes are transmitted to the next generations.

Chapter 2

Rationale of the predictive approach developed

A new approach for evolutionary predictions

Alix Mas, Yvan Lagadeuc, Philippe Vandenkoornhuyse.
Université de Rennes 1, ECOBIO UMR 6553.

Two hundred years into the exploration of evolution, and the question of predictability is still actively debated. The beliefs have consecutively shifted, from a deterministic Lamarckist view to the all-random insights of the Neutral Theory (Kimura, 1983), and the questions of determinism, and thus of predictability, have made ink and ideas flow, as much as they have triggered many research (Stern and Orgogozo, 2008; Szendro et al., 2013; Kryazhimskiy et al., 2014; Blank et al., 2014; **?**; de Visser and Krug, 2014; Lapidot and Conley, 2015; Duarte et al., 2015; Bank et al., 2016). The first part of this paper reviews some of the arguments weighing for the unpredictability of evolution, while the second part present alternative arguments that could invert the balance. A third part presents a new approach experimentally testable which could give further insights into the constraints and predictability in the particular evolutionary trajectory of specialization.

2.1 Introduction

Since the beginning of the theorization of Evolution by de Lamarck in 1800 until now, the underlying question of predictability has been tightly imbricated with the investigation of evolution itself. In a deterministic view of evolution, the environment is the cause and determinant of evolutionary trajectories followed by organisms which attempt to adapt to these environments. If the environmental parameters are well enough apprehended, the evolution of organisms present should tend to some defined expected form. This view was challenged and progressively refuted by the subsequent theories of Darwin, and Kimura: Darwin's evolutionary theory states that the environment itself is not the causal parameter behind the change of phenotypes observed (traits, functions) in organisms, but that it only acts as a filter selecting the best adapted phenotypes. Further understanding of molecular mechanisms of DNA lead Kimura to state that the random mutations causing the observable differences in phenotypes mostly had circum-neutral effects on the phenotype and thus were conserved at random (Neutral Theory of Molecular Evolution).

While the first approach (Lamarckism) should make it potentially possible to predict evolution at the phenotypic level, Darwin's theory makes predicting more intricate as the purpose and thus trajectory of evolution is not especially determined. Finally Kimura's theory makes it plainly impossible to predict evolution, as randomness is the angular stone of his work, and random is by essence not predictable.

It is nevertheless important to precise that all these theories are not set at the same biological level, which can lead to a misconception of prediction. Indeed, the actual widely accepted theory of evolution (Modern Synthesis) states that natural selection is a driving force of evolution, acting (passively in fact) as a filter conserving only the fittest individuals. This filter operates at the phenotypic (and extended-phenotype) level only. Natural selection acts on phenotypes and thus indirectly acts at the genomic, or metabolic level, through the assimilation of genes (and else) into the phenotype.

Prediction, on the other hand, could concern any of the biological levels known (from genotype to community evolution) : one could want to predict patterns at the molecular level (i.e. amino-acid or gene(s) modified during evolution), the functional level (i.e. traits modified), the individual's phenotype level (i.e. the phenotype selected) but also at the population or community level (i.e. modification of alleles frequencies, species overcoming others).

In the following, the potential to predict evolution at these different levels is considered.

2.2 The Unpredictability of Evolution

2.2.1 Mutations' randomness and effects

In the context of the Modern Evolutionary Synthesis, which accept natural selection as its basal theory and combines it with complementary explanations of mechanisms, the current dogma is that mutations are random.

Genetic variants occur randomly, as a result of DNA replication mistakes which are eventually integrated in the genome and transmitted to the new generations. Since these mistakes are not driven by any underlying force and are strictly random, it is impossible to forecast where on the genome (on which nucleotides, codons, or (set of) genes) and which type of mutations (SNP, Insertion, deletion, synonymous or not) will arise.

Additionally, the modification of a gene does not necessarily have linear effects on the modification of the phenotype. Some genetic modifications will not affect phenotypic traits (neutral evolution), while others can have a cascade of consequences at the phenotypic level if they occur on a regulatory sequence of an operon, on pleiotropic genes or on epistatic genes. Operons are cluster of genes that can be transcribed differentially to encode for different gene prod-

ucts. Therefore the effects of mutations affecting the regulatory sequence of an operon are difficult to predict. When mutations occur on pleiotropic genes

Pleiotropy, is defined as a genetic phenomenon in which a single gene impact different phenotypic traits at the same time. Thus if a mutation affects a pleiotropic gene, it can modify different traits simultaneously. Pleiotropic genes generally encode for products targeting cells sharing the same signaling mechanisms or more generally products used by a range of cells. Therefore the effects of mutations emerging on these pleiotropic genes are difficult to encompass, and unpredictable. However pleiotropy seems to have very important functional consequences for cellular cooperation in multicellular organisms (e.g. Aktipis et al, 2015) but also important consequences for microorganisms communities interactions, especially when considering the evolution of cooperation with the molecular network, making [...] pleiotropy an effective way to stabilize cooperation evolution [...] (Mitri and Foster, 2016), thus confirming particular evolutionary trajectories on these pleiotropic genes.

Beside pleiotropy, other genetic modifications are difficult to assess to predict evolution, it is the case for mutations appearing on epistatic genes. Epistasic genes are genes that act in interaction with other genes, and of which the effect by itself is different than the effect when in combination with other genes. Thus epistasis is a phenomenon of genes interactions that affects traits expression, and thus phenotype. Genetic modification occurring on such epistatic genes are only phenotypically observable because they affect the presence of other genes. Understanding epistasis is important in predicting phenotype from genotype for a given individual (e.g. Miton and Tokuriki 2016, Sackton & Hartl, 2016). Recent advances on the understanding of these 'genetic interactions' suggest that they contribute to many complex traits (Bloom et al. 2013; Weinreich et al., 2013; Taylor and Ehrenreich 2015). Of course the predictability of such complex epistatic interaction is currently out of reach but recent models to explain heritable variation of the traits seems successful when including the additive contribution to a given phenotype (i.e. the known epistasis /genetic interactions) (Forsberg et al., 2017).

It can also be added that it is relatively rare that only a single mutation will transform entirely a phenotype, it is often a combination of temporally successive modifications (neutral or not) of the genome that will result in the modification of a feature or give rise to the appearance of a new feature. This adds to the complexity of predicting mutations occurrence and effects on phenotypes.

In this context, the only possible prediction enabled are the ones relying on

the concept of fitness: if we know a mutation confers the organism a better fitness, then it will be naturally selected for. Yet, predictions on mutations *per se* are not directly accessible.

2.2.2 Phenotypes Complexity and Neutral Theory

Forecasting the emergence of mutations is so far impossible, but could we forecast the emergence of particular phenotypes?

Considering we could know what traits or features would need to be adjusted to increase the fitness of an organisms in an environment (for example change of the color of the organism to make it more cryptic and thus less prone to be predated, as in Arendt and Reznick 2008) we could predict such phenotype variation to arise. But in reality, phenotypes are multi-dimensional systems, and thus encompassing the evolution of only one particular trait might be far from the actual constraints at play on the organism. For example, we cannot necessarily determine which trade-off such modifications will imply, what can be the functions diminished or impaired, and what evolutionary consequences it may have. It is true that often, when studied, the systems are oversimplified to particular traits of the phenotypes (traits supposed to affect fitness the most) while a phenotype is a synergy of traits and functions expressed by an organisms and where a feature can influence another. Understanding the evolutionary dynamic of a single trait is an interesting but incomplete perception of the evolution of the organism, making accurate predictions difficult.

From an other perspective, under the Neutral Theory framework (Kimura, King, Jung), as the vast majority of mutations have circum neutral effects, natural selection cannot act on the phenotypes, which are then mainly conserved stochastically. In consequence, if phentoypes are picked randomly, there is no way one could predict the functional traits that will be kept in a population for a given environment, since evolution is not adaptationist. In this case, evolution is not directed and thus also unpredictable.

Both the bottom-up (genomes to phenotypes) and a top-down (environment to phenotype) considerations suggest, by their respective randomness and complexity, that forecasting the precise emergence of a multi-dimension phenotype in given environment is out of reach.

2.2.3 The Environment of Evolution

Additionally to these straightforwards limits to evolution forecasting, it is also important to underline that similarly as for phenotypes, environment is a dynamic and multidimensional system. Consequently, focusing on one or a few environmental parameters at a given time point is somehow restricted and biases the view of the systems.

Usually when studies define adaptation to an environment they focus on relatively few parameters of the environment, such as the temperature, the light, the toxicity, the parasites or some resources present, and rarely experiments are made with multiple variables at the same time. These studies can give interesting insights on the evolutionary response of the organisms toward those particular variables, but these insights are limited, as the environment is constituted of interactions between biotic and abiotic components themselves variable and having effects on each others. The interactions between the variables are often overlooked, and evidently it is not possible to integrate all the existing variables and their interactions into *in vitro* experiments or *in silico* model (Morozov, 2013). Additionally, environments are dynamic and constantly vary over space and time, and the predictability of these variation themselves are very intricate. On an evolutionary time scale (which differs according to the generation time of organisms) it is not possible to predict with certitude environmental variations, and thus anticipate the adequate traits organisms will develop to respond to these changes. Furthermore there is an import part of the local environments that is modified by the organisms themselves as they evolves (eco-evolutionary feedbacks: Fussmann et al. 2007; Ferriere and Legendre 2012), calling for the need of continually refining new adaptations. This eco-evolutionary feedback loop, because it is not fully recognized yet, is hard to integrate into predictive models.

As being depicted, natural selection itself is based on ever-changing and complex environmental parameters, which cannot all be appraised simultaneously, and the overall complexity of each systems makes them unpredictable.

2.2.4 The Chaotic Dynamics of Biological Systems

Besides, on a more theoretical line, even if we could integrate all the dimensions of a phenotype, a study on the long term predictability of evolution of complex phenotypic systems affirmed that chaos-like dynamics were more likely to determine evolution (Doebeli and Ispolatov, 2014). This means that even if ev-

ery single parameters possible and their respective effects on each other were known, a single minuscule modification could lead to the rise of major changes over generation times (known as the butterfly effect), which we are currently not able to predict. This study and others (e.g. Huneman 2012) state that natural selection may be deterministic, but not necessarily predictable nonetheless. If the biological system is chaotic, it would require an omniscient understanding of every variable at stake to be able to make predictions of organisms evolution in this context.

Because every mechanisms are intricate, with each hierarchical ecological level (Figure 1.8) having repercussions onto the others, and also because evolution is dynamic and perpetual, it is thus impossible to embrace all these considerations holistically. As a result of this complexity and because of the apparent randomness of several mechanisms, evolution is appraised as being unpredictable.

2.3 A door ajar on the predictability of evolution

Yet, when narrowed down to simpler systems, distinct evolutionary patterns emerge, suggesting the existence of constrained mechanisms, which could enable some predictability. Such patterns are presented below.

2.3.1 Convergence of Phenotypic features

The most striking patterns of evolution that emerged are the ones of repeated evolution at the phenotypic level, where analogous phenotypes appear in similar environments, sometimes to respond to similar constraints (Arendt and Reznick, 2008; Gompel and Prud'homme, 2009; Conway Morris, 2010; Lee and Marx, 2012; Bailey et al., 2016) .

At the phenotypic level organisms living in similar environments will regularly evolve the same features. This phenomenon has been widely studied for the evolutionary radiations of cichlid fishes of the African Rift lakes (Sturmbauer et al., 2010; Muschick et al., 2012) and for stickleback fishes (Rundle et al., 2000) , or again for the radiation of *Anolis* lizards (Losos, 1998)), where in each cases a set of similar phenotype emerged in each ecosystems they were subjected to. This kind of repeated evolution is also compelling with the example of emergence of similar environmentally coherent coloration-phenotypes in mice and fishes (Gompel and Prud'homme, 2009) and stick insects (Comeault et al., 2016).

Parallel evolution also works at the functional level: as in the repeatable experiments of Rainey (Rainey and Travisano, 1998), where a population of bacteria repeatedly evolves several phenotypes to exploit better the experimental environment. Herron and Doebeli (2013) thus argued that [...] *parallel genetic changes underlying similar phenotypes in independently evolved lineages provide empirical evidence of adaptive diversification as a predictable evolutionary process* [...].

From all these accumulated observations and even without necessarily knowing the cause of phenotypic convergence, predicting evolution at the phenotypic level sounds reasonable for particular environments.

The emergence of similar phenotypic features is not omnipresent, but frequent enough to have attracted evolutionary biologists' attention (e.g. (Harmon et al., 2005; Losos, 2011; Stern, 2013)). Mechanistically, it is not impossible that such repeated evolution could exclusively be the result of fortuitous events (random mutation leading to emergence of same traits, either randomly or naturally selected for) but alternative options also exist.

Indeed, the fact that similar phenotypes emerge distantly in space and time,

and repeatedly in either closely related or highly different species (e.g loss of pigmentation, (Gompel and Prud'homme, 2009)) suggests that certain types of transformation are favored in evolution. It could be possible that the emergence of new phenotypes is extremely constrained (for example at the molecular or metabolic levels), and that these repeated traits observed are the unique answers possible to respond to an environmental variable, thus they are the only ones to emerge.

Yet a more flexible option can be that some constraints effectively shape potential outcomes, leading to the emergence of a finite set of phenotypes matching the environment, and amongst these possibilities, the one selected are the one offering the optimal answer for the given environment, often (but not always) resulting in the emergence of similar features in organisms.

2.3.2 Predictability in the evolution of genomes?

On a relatively frequent basis, the emergence of equivalent phenotypic features are lead by similar genetic modifications. It is the case for the mutation of genes encoding pigmentation that is shared by very different species and that mutate to entail a loss of pigmentation (Gross et al., 2009; Gompel and Prud'homme, 2009), or also for (closely related or not) insects developping a resistance to the toxicity of plants they consume (Dobler et al., 2012) and similarly for snakes resistance to poison (Feldman et al., 2012), or again for the evolution of compensatory mutation to reduce the cost of antibiotic resistance in *Bacillus spp* (Levin et al., 2000), to mention only a few of the compelling examples existing.

This parallel evolution at the genetic level have emerged either because the same families of genes were recruited, because orthologous genes (genes of identical origin) were modified, and even sometimes because the same nucleotides on a gene were affected (Cooper et al., 2001; Wood et al., 2005; Woods et al., 2006; Arendt and Reznick, 2008; Gross et al., 2009; Christin et al., 2010; Stern, 2013; Rosenblum et al., 2014; Signor et al., 2016; Conte et al., 2012; Lee and Marx, 2012; Le Gac et al., 2013; Bailey et al., 2016).

With the amount of DNA data collected and analyzed, repeatable and convergent patterns of molecular modifications are being unveiled, unveiling the potential for predictable patterns of evolution at this level.

Comparative analyses permit to consider the question of the similarity of mechanisms behind repeatable phenotypique evolution (Arendt and Reznick, 2008; Gompel and Prud'homme, 2009; Christin et al., 2010; Lobkovsky and Koonin,

2012; Toll-Riera et al., 2016).

According to species studied it was shown that (similar) variation could stem from different genetic sources but that often it stems from identical genetic modifications (Gompel and Prud'homme, 2009). At the genetic level, clear types of repeated events are characterized. Regions of the genomes more suceptible to mutate are well describe (Stern, 2013) and some genes are known to mutate more often than others (this is the case for duplicated genes for example (Toll-Riera et al., 2016). For others genes, it will be their position in the gene regulatory network or their pleiotropic effects that will determine their propensity to mutate (Gompel and Prud'homme, 2009). Blank et al. (2014) also determined that depending on the cellular context of the metabolic innovation, the mutations would preferentially affect protein structure itself, or protein expression level only if for example the development of the novel trait is spatially or temporally limited (Blank et al., 2014).

The frequency of convergent evolution is too conspicuous to be only attributed to stochasticity, at least in the selection of conserved modifications. But one also has to keep in mind that it is virtually impossible to detect mutations that are not conserved, as they do not improve fitness or are maladaptive. Thus with a static view, it appears that some genes are repeatedly modified, while it might be that other genes were also mutated but the changes were not conserved by natural selection. In that sense, it is tricky to know if mutations are random or constrained molecular events, and maybe there are some still unraveled mechanisms for which, currently, we can only see the outcome as random. Indeed the convergence and repeatability of observed features (Wood et al., 2005; Woods et al., 2006; Arendt and Reznick, 2008; Gross et al., 2009; Conway Morris, 2010; Lee and Marx, 2012; Stern, 2013; Rosenblum et al., 2014) suggest that genetic modifications are constrained (to optimize the effect of mutations with the fewer collateral effects as possible) and thus restricted to a subset of solutions. Yet these conspicuous patterns may be explained by the fact that observed solutions are the ones entailing the overall fittest phenotypes, and thus selected for.

Even if it still seems impossible today to precisely forecast the apparition of a precise mutation on a single nucleotide to determine adaptative evolution, as described above, recent understanding behind the recurrence of mutations and genes modification patterns gages of mechanisms still unraveled that would enhance predictability. We introduce hereafter new ideas to predict evolution.

2.3.3 Evolution, a process constrained at several levels

As presented above, convergent and repeated evolution, at the phenotype and genomic levels, suggest the existence of repeated mechanisms in action, mechanisms which if understood could permit the prediction of their outcome. Very often only the environment is considered as a filter in which the organisms will thrive or decline (natural selection is according to the adaptation to the *environment* through which phenotypes are sieved and conserved or eventually discarded). However, every biological level (genetic, metabolic, functional, etc..) of a living system is constrained in some way, which imposes a reduction of the set of valid possibilites for adaptive evolution (Figure 2.1). Yet the ensemble is rarely considered altogether.

The conceptual 'evolutionary funnel' (Figure 2.1) explains how from the virtual set of potential modifications that could emerge at the individual level, these variations are sieved by the different constraints existing at the different biological levels. Environmental constraints act on the expressed phenotypes, the phenotype being a consequence of genome expression and functioning, itself being determined by the genetic variations. From this conceptual understanding, the predictability of evolution (and thus related retained genetic changes) is rooted in the understanding of the consequences of the environmental filtering of the phenotypes expressed. In this 'evolutionary funnel', the possibility of convergent traits can be triggered both by similar genetic modifications and underlying molecular events or by changes on dissimilar genes. The accuracy of the predictability is likely to be positively linked with the stringency of the constraints.

Evolution is therefore the result of several different forces and constraints exerted on each biological levels considered (Figure 1.8). The virtual set of every potential mutations and phenotypes that could come to existence are reduced to a subset of possibilities, sometimes driving to the emergence of similar evolutionary trajectories.

And if perfect forecasting of evolutionary trajectories is not accessible yet, understanding the mechanisms and functional constraints at each biological level could permit to narrow down the potential genetic modification, and corresponding phenotypic features, emerging in given environments. In respect to this, we developed a new approach which could enhance prediction by focusing on metabolic constraints existing on individuals.

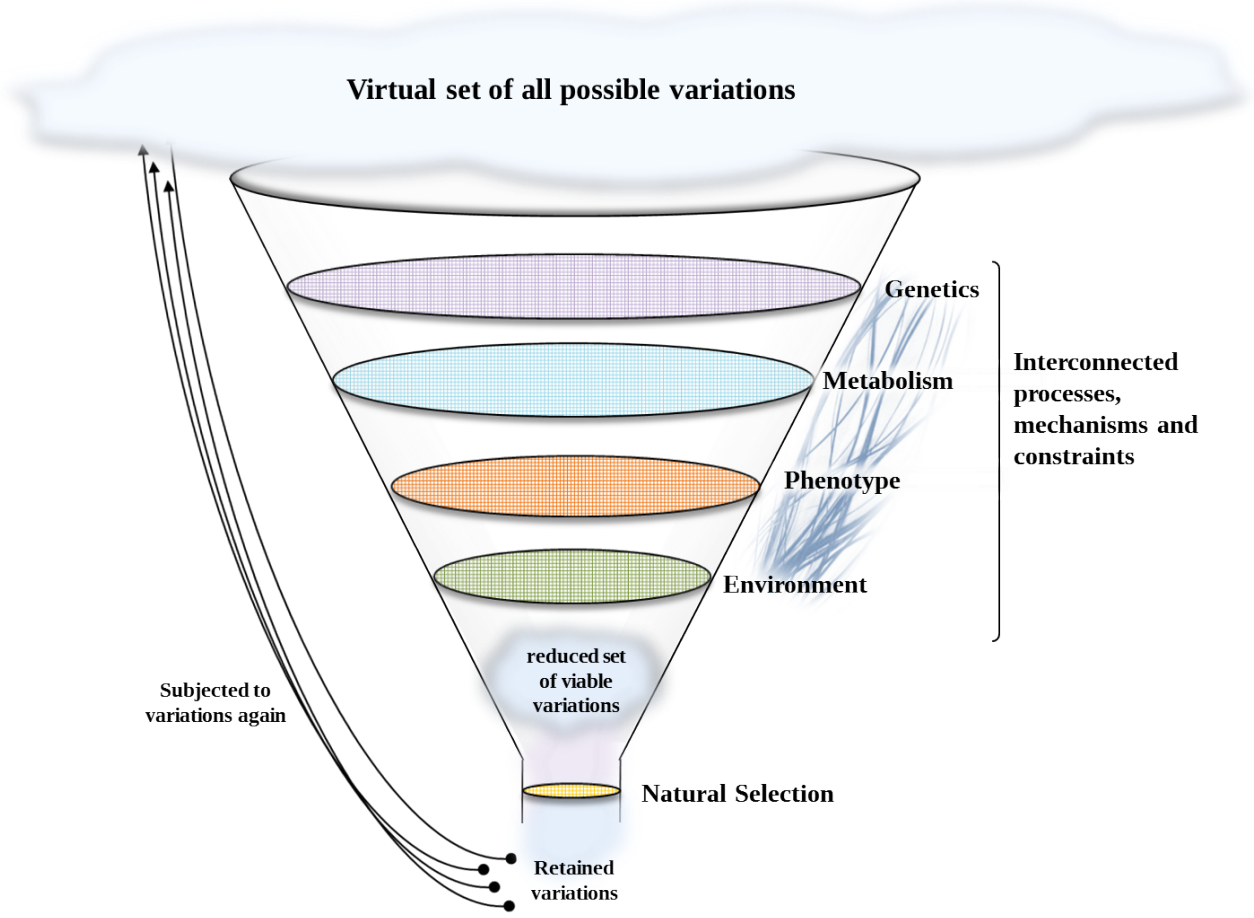


Figure 2.1: **The evolutionary funnel: constraints shaping evolutionary possibilities.**
The first level of constraint is the intrinsic properties (physico-chemical) of the genetic code which enable only restricted modification of the genome. Genome-wise, the complex network of interacting genes restrains the modification possibilities as any modification on a gene can have cascade effects on other genes. Here, the effect of pleitropic genes is crucial as the constraints exerted on these genes is high. Then, these changes have to be viable, with core metabolic functions conserved (not carrying important modifications), while other accessory functions will be more readily modified (Lee and Marx, 2012). Then the presence of trade-offs at the phenotypic levels also shapes the potential evolutionary trajectory that can be followed. Finally, the environment, biotic and abiotic, also constrains evolution, for example by the resources available and the interacting species present, modifying the population and community dynamics. From the final possibilities existing, Natural Selection (i.e. adaptation to current environment) will drive the conservation of a few solutions over others.

2.4 A Metabolic approach to Predict evolution in a specialization context

Despite the stochasticity attributed to evolution (random mutation, genetic drift, chaotic system), the repeated emergence of similar gene modifications and of identical phenotypes under similar environmental constraints suggests that evolution is at least partly shaped by constraints at the different biological levels (constraints on the functioning of the genome, functional trade-offs, biotic and abiotic environments) which narrow the set of alternative evolutionary trajectories that can be followed by organisms to fewer optimal solutions.

2.4.1 Metabolic Constraints

A level of constraints which has however been overlooked in most of the studies is the *metabolic* level.

The metabolism is often considered as the interface between the environment and the organism's adaptation. Genes expression encoding for enzymes catalyzing every chemical reactions are at the basis of the metabolic network. Metabolic functions expressed by an organism are sometimes used to characterize its phenotype, classified for example as phenotypes able to produce typical molecules, such as melanin, toxin, anti-toxins, siderophores, etc.

The metabolism can influence interactions (production of chemo-repellent or attractant) and can even shape co-evolution between related or distant species. It holds for endosymbiotic relationships, as in the relationship between aphids and *Buchnera aphidicola*, but also for free-living organisms, as in the evolution of dependencies based on the production of a common good (eg. Black Queen Hypothesis, Morris et al, 2012, loss of common good production gene(s)).

A metabolic approach is thus expected to improve the characterization of constraints shaping evolutionary trajectories, both at the phenotypic and ecological level. It could also permit to directly link environmental constraints to potential genetic modifications.

2.4.2 Reductive environment and Specialization

As defined by the "evolutionary funnel" described in section 2.3.3, organisms are subjected to multi-factorial (and dynamic) constraints, which are both intrinsic, and environmental. The organisms develop adaptive compromises to balance each of these constraints (trade-offs). In the current context it is not possible to

predict precisely the overall evolution of an organism, as many of its functions or traits could be modified simultaneously to entail a better fitness.

One can hypothesize that it is more suitable to study evolution in a range of constraints to which the organism is known to respond, rather than to explore a new set of constraints previously inexperienced by the organism. In other words, there is a higher potential for prediction of evolutionary trajectories in an theoretically 'reduced' environment (with either less variables than the initial environment, or variables with a smaller range of variation) than in an 'expanded' environment with new types of variables. For the sake of the example let us consider only one metabolic function of an organism (such as uptake of nutrients), and that the organism as the capacity to 'display' this function in 10 variations (it can uptake 10 different sort of nutrients). If the mechanisms behind the uptake of these 10 nutrients are known, it appears simpler to predict the optimization of the uptake-function toward the uptake of one single nutrient (specialization toward this particular nutrient), rather than to predict an adaptative innovation enabling the uptake of an eleventh nutrient.

In the following, the word specialization is used to encompass such a "reductive" evolution of functions (loss of function) of an organism compared to its basal capacities (Figure 2.2).

Such reductive specialization of organisms to their (biotic and abiotic) environments is not rare. It has been intensely studied for endosymbionts (Lai et al., 1994) and more recently shown to be existing also for free living organisms (e.g. Giovannoni (2005); Boscaro et al. (2013); Swan et al. (2013). As long as the environmental constraints are strong and stable enough over evolutionary times, it can give rise to specialization.

2.4.3 Forecasting metabolism and evolution during Specialization

In stable environments, organisms are expected to specialize, as specialization confers fitness benefits. The specialization to particular constraints of the environment is regularly associated to genome reduction in micro-orgnaisms (e.g. Lai et al. 1994; Dufresne et al. 2005; Morris et al. 2012a). In short, such a reduction can happen either through selection for reduction (superfluous and costly function *have* to be deleted) or because functions that are superfluous in a given environment undergo a lift of selection thus the genes implied in these func-

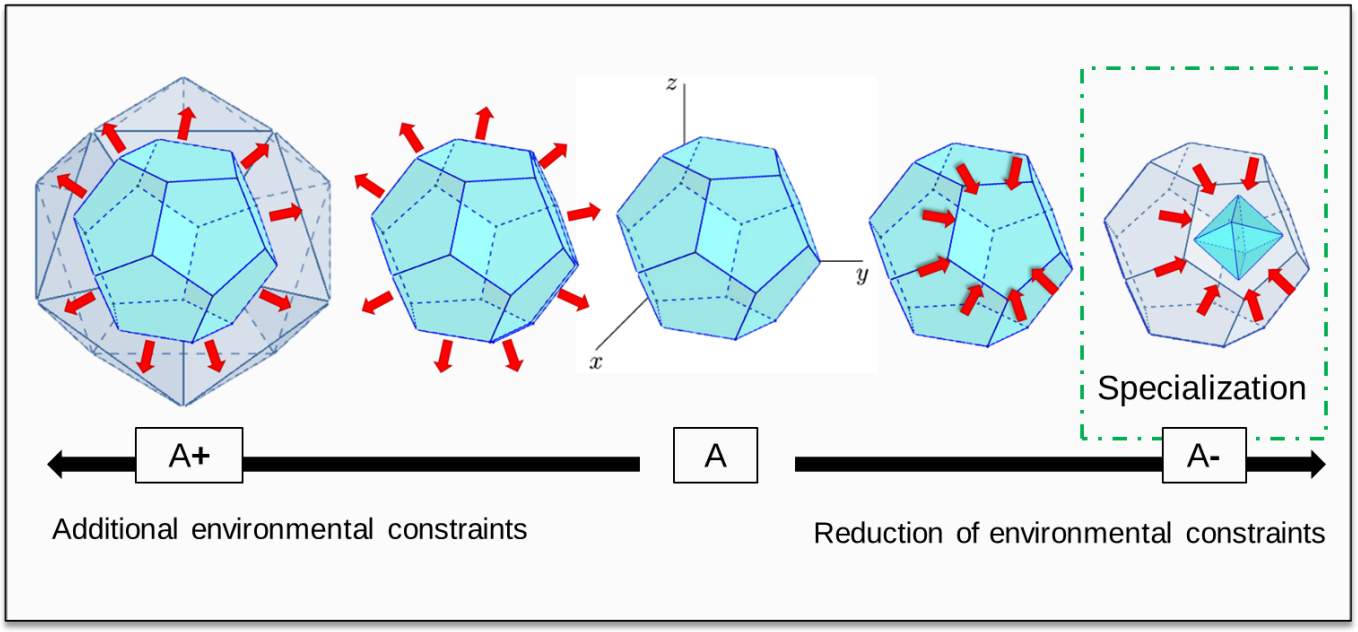


Figure 2.2: **Evolution in a "reduced" environment to entail specialization.** **(A)** The central hyper-volume represents in a simplified way the phenotype of an organism; each ridge of the volume is a feature of the organism which can display variable values (for example, a ridge is "color of the organism" which can take the values beige, grey, brown, black etc. **(A+)** The expanding space of possibles on the left (grey space of the volume) suggests that if changing to a new environments with previously inexperienced parameters, the organism might have to develop new features, or new variations within existing features that are not part of the current phenotype (which could happen through gene duplication or gene acquisition via horizontal transfer). Prediction in this context is difficult, as one would have to determine all the possible innovations and evolutionary trajectories. **(A-)** The reduced volume schematize the specialization of an organism when its environment is reduced, more constrained regarding the range of parameters existing. In this case predicting evolutionary trajectories is in the field of possibles as we expect already existing features to be modified to optimize their activity. For example, features permitting a response to a constraint that is conserved should be enhanced, while features that permitted a response to constraints that were removed could decay.

tions are fated to decay (by accumulation of neutral or slightly deleterious mutations) (Lahti et al., 2009). It is possible that the two mechanisms act in parallel. These notions of 'selection for' or 'lift of selection' at the genetic level can be explained by the fact that at the metabolic level, adaptive trade-offs occurs. In a temporally stable and spatially homogeneous environment the activity of some metabolic functions will be favored over others, leading to a differential expression of the genes implied in these functions. Changes in metabolism is one of the first response of organisms to adapt to new or changing environments. If these changes continue over time, these metabolic optimization are expected to be progressively integrated in the constitutive genome during specialization as the organisms will benefit from specializing to the environment.

It is expressly in this context of evolution toward specialization (concordant with a "reduced" environment compared to the initial environment) that we believe it is possible to predict, at least to some extend, the evolutionary trajectories that will be followed by the organism at the genetic level, but also eventually at the population or community levels.

These metabolic optimization are expected to be progressively integrated in the constitutive genome during specialization. If a process of genetic accommodation happens (i.e. a phenotype originally produced in response to environmental conditions through differential gene expression is later being stabilized by a genetic modification, West-Eberhard, 2005; Schlichting and Wund, 2014), information about early stages of the genome expression might be important information to consider to predict evolutionary events.

Prediction of genome modifications according to Metabolic constraints

Through the latest DNA improvements, it is possible to characterize the genes implied in various metabolic reactions, whose activity is determined by the presence/absence of metabolites in the environment.

Additionally, modelling such as Flux Analysis permit to define the activity of metabolic functions, depending on the environmental variables input in the model. Thus the metabolism of a specialized individual can be determined (in comparison with a more generalist model), and the metabolic pathways preferentially activated (respectively inactivated) can be retrieved. Genes implied in these pathways can be characterized and assumed to express enhanced or reduced expression. Therefore evolutionary predictions under the frame of specialization could be assessed by Flux Analysis. In a context of specialization to a

reduced environment, we are making the assumption that using this approach, the genes under strong selection and those which are not under selection could be predicted to show particular genetic modifications (e.g. stop codon mutation within unused genes).

In parallel, genetic modifications can be characterized from experimental evolution experiments matching model conditions to test for the accuracy of this approach.

2.5 What a (R)evolution!

The exploration of Evolution is century old, yet it is still developing, with an always more fine-tuned understanding of the mechanisms underlying its functioning. The addition of high throughput sequencing data permitted to revisit the approach of evolution exploration and brought great genetic comprehension. In parallel various modeling approaches enable to test for evolutionary trajectories and dynamics simultaneously, while experimental approaches permit to answer precise questions. The cross-disciplinary lines of evolutionary studies empower an integrative view of evolutionary mechanisms in detail. With this deep understanding being reached, repeatable and convergent configurations are being unveiled, traducing potentially predictable patterns of evolution at different level of integration.

As Gould (1998) proposed, if we were to rewind the clock of evolution , it would probably give rise to something completely different (Gould, 1989). This outcome is usually associated to the randomness of evolution, while it could also be due to the complexity of biological systems where the components are multidimensional and interact together, which results in dynamic emergent properties. Even if deterministic, any infinitesimal change could change the whole face of evolution, making it an apparent chaotic system.

On the other side of the spectrum, Conway Morris (2010) stated that "Evolution: like any other science it is predictable" (Conway Morris, 2010). Indeed, as determined above there is without doubts some patterns and mechanisms underlying evolution which can lead to predictions. Yet, evolution is not like any other science. A recent discussion with a physicist inspired me: their community is excitingly becoming aware of, and integrating, the "active matter" to their research. After getting around the arrangement of static 'basic' matter, they are now being interested in how active matter reacts: take a system where you put supra-molecular bits together, well, the bits will organize in a way that they all fit

in the space as they can, and have room to move a little, without being interested in whats going on a few 'space' away. Such a simplified system, where there is no interactive or retroactive effects of the units on the others or on the environment, nor emergent properties at the system scale, is not too difficult to predict: once you understand the forces at stake (physics) and the initial conditions, it is possible to predict how the system will be put together. But living matter? It is an entirely different thing! Life presents an extraordinary organization, with interactions, from all kind, changing dynamics, synergy, and antagonistic relations, even willed choices or altruism. We are here in very intricate and multifactorial systems, which is both dynamical and paved with eco-evolutionary feedback. Is thus much less easily predictable.

Under the hypothesis that evolution (of an individual, a species, a system) could be predicted, only an omniscient intelligence aware of the dynamic, the influence, the interactions, the retro-effects between each parameters and able to integrate them all could predict some evolutionary trajectories. Until then, the tremendous efforts made in the research on evolution have permitted to affirm that some mechanisms underlying evolution were similar, either because they are constrained, or because they present the best (parsimonious, higher positive effects) path possible. Thus, constraints (environmental, metabolic, genetic) and the interactions between these constraints, with selection acting on the phenotype, are current factors known to determine and shape evolution the way we perceive it today. This system is being better understood every day and may be completed by constraints undetected so far.

It is interesting to remark that, to some extent, this concept meets with the laws of usage and non usage underlying Lamarck's adaptative theory, but at the molecular and metabolic levels (vs the phenotype level) this time. Lamarckism has recently rejoiced through the discovery of several mechanism conceptually close to Lamarck's ideas. These Neo-Lamarckist mechanisms are 1) epigenetic modifications, which are modifications acquired during the individual's life, and can be transmissible to the next generation, and 2) the horizontal transfer of genes, which are also acquired during the lifetime of micro-organisms and kept along generations. The environment thus straightforwardly influences the genotype (gene transfer) or indirectly (gene expression modification through epigenetics). This should push forward the need for reintroduction of Lamarck-like phenomena in evolution, as proposed in the extend evolutionary synthesize.

So far, with the knowledge gathered, evolution is not either predictable or not, it is both predictable and unpredictable, depending on the level focused on.

Evolution may be like a cursor moving on the continuum of determinism and randomness, and maybe one day we will be able to quantify the degree of each.

Chapter 3

Thesis Context and Objectives

Are we able to predict evolution? Is evolution partially deterministic or totally stochastic?
For most evolutionary ecologists these questions are not to be addressed anymore, given the evolutionary knowledge and consensus theories indicating the stochasticity of genetic events.
Nevertheless, and strikingly, a number of example of convergent evolution (convergent features, but also parallel genetic underlying processes) have been observed and determined. The previous section (Chapter 2) provides an overview of these processes.

As developed in Chapter 2 of Part I, the expanding space of possibles (figure 2.2), including the gain of functions for a given organism, seems much harder to predict than the reduced space of possibles, as for example the loss of functions. The specialization of an organism is typically an evolutionary trajectory where a lift of selection or a selection for gene loss of superfluous genes is expected to entail a fitness increase. These changes in selection exerted on genes could be envisaged as being driven by a change in their expression activity which is defined by environmental constraints.

The evolution of loss of function and genes has recently been theorized for free living bacteria. In this theory (the Black Queen Hypothesis (BQH), Morris et al. 2012), a given microorganism is losing the ability to produce a common good, for example a detoxifying enzyme, if another co-existing microorganism is still producing the common good. This evolutionary trajectory allows to escape competition and competitive exclusion (Mas et al, 2016). This particular gene loss ends to a stable evolution of dependency. Thus the BQH framework also covers the specialization of a bacterium to another one, forming a stable interaction between these free-living microorganisms.

In the context of specialization, one objective of this Ph.D. work was to determine to which extend the evolutionary trajectories of bacteria, both at the genetic, functional and ecological interactions levels, could be forecast.

One of the strategies developed to study the evolution of specialization invokes a metabolic view of the system. Metabolism is at the interface of the environmental conditions of an organism and the functioning of this organism. It is one of the first component being impacted by a change in environment, and thus considered as a key element in adaptation and evolution. More over, through the latest description of genetic processes, it is possible to tightly link genes and metabolic functioning. The interplay between these two components

is at the source of potential predictions: the environment affects the metabolism, which can be sourced back to the functioning of genes.

The second section of the thesis, discusses the importance of interaction between species in evolution, how they can shape evolutionary patterns, and the possibility to predict the emergence of these evolutionary patterns. To this purpose, in Chapter 1, an Agent Based Modeling approach considers the fitness increased gained by an organism when it modifies its metabolism (loss of functions mutant) to depend on other organisms of the population to compensate for this precise lost functions. By analyzing *in silico* the raise of a metabolic dependency, we explored the population steady states reached under different constraints, allowing for a new understanding of the Black Queen Hypothesis.
The second chapter develops another interesting view offered by the metabolic approach : the metabolism is also the interface between organisms interactions, of the same species, or of different species. For example, different species could be led to live together because they have metabolic complementary or facilitation, and reciprocally, they could have co-evolved metabolic complementarity as a result of their long-lasting coexistence. Thus by determining the metabolic interactions existing between organisms, and still in a context of evolutionary specialization (of one organism to the other) it should be possible to both understand (and predict) the rise of tighter and obligate interactions which rely on metabolic dependencies. It would also be possible to forecast the genetic changes entailed in such an evolution of interaction.

The third part of the thesis presents a different modeling approach to predict the evolution of specialization under particular constraints at a finer grain. For this purpose, the metabolism of a model bacteria, *Pseudomonas fluorescens* was studied using a combination of Flux Balance Analyses and Flux Variability Analyses strategy in order to provide for testable predictions of evolution (Chapter 1). Following these *in silico* predictions, *in vitro* evolution experiments were realized and some results are presented (Chapter 2). From deep coverage *Pseudomonas fluorescens*' population DNA sequencing, we analyses the genetic modifications (punctual mutations) which emerged *de novo*. These detected mutations and their fitness are interpreted in the light of the metabolism functioning. In this section, information on the confrontation of the observations to the predictions are also presented. Based on working hypotheses on motility and chemotaxis, Chapter 3 of this part presents a functional analysis of the specialization which was performed by comparing the bacterial behavior of the initial *Pseudomonas fluorescens* population (i.e. pre-evolution) with the evolved population.

Finally a short general discussion summarizes the rational and results of this thesis and a perspectives section develops additional ideas in relation to the Ph.D. work presented herein. This part also discusses the bridges built between the population genomics, the genomic functioning, and the ecological functioning of living systems by combining both *in silico* and *in vitro* approaches. It also highlights the doors opened by this multidisciplinary approach to address new questions about evolution.

Part II

Forecasting Population and Community Evolution based on Metabolic Interactions: Around the Black Queen Hypothesis

Evolution is often seen as a complexifying process where organisms acquire new traits and new functions in order to stay adapted to their changing environment. This is well illustrated in the famous Red Queen Hypothesis (Van Valen, 1973), which explains the coevolution between antagonistic organisms, and the impressive array of adaptation these organisms develop in order to 'stay in the (arms) race'. For that all, a major evolutionary process has been mostly overlooked in spite of its ecological importance: Evolution by loss of trait or function (Visser et al., 2010; Ellers et al., 2012). This concept of evolution by simplification was recently theorized in a new evolutionary hypothesis: The Black Queen Hypothesis (BQH, Morris et al, 2012), according to which organisms may also adapt to their environment and congeners by undergoing adaptive loss of functions.

It is not to prove anymore that relations of the parasitism and predatory type shape the evolution of implicated populations, both in term of structure, and in term of phenotypic or functional features (co-evolution) (Connell, 1980; Schulte et al., 2010; Buckling and Rainey, 2002). Yet positive interaction can also be the source of coevolution, with a redistribution of essential functions between the two partners enabling a lighter 'functional burden' for each (Ehrlich and Raven, 1964; Brundrett, 2002; Morris et al., 2012a, 2014).

Either because they are less obvious or less ubiquituous [1], these interactions were less investigated.

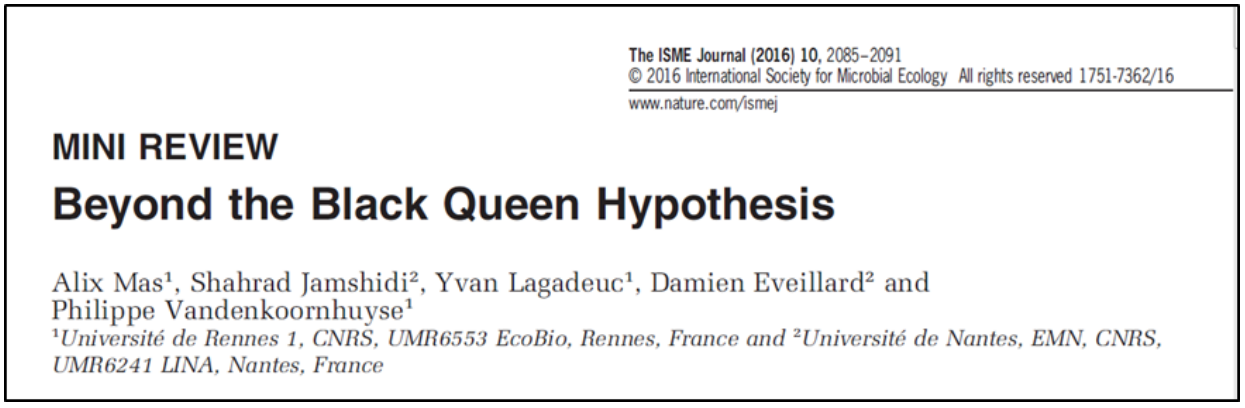
In the following, a study investigating the emergence and the consequences of such a loss of function in the frame of positive interaction is presented. This work is set in the framework of The Black Queen Hypothesis (Morris et al., 2012a). It focuses on how the evolution of dependency between two populations of organisms transforms interactions and the community. Using agent-based modeling we suggest that species specializing in the consumption of a common good escape competition and therefore favor coexistence. This evolutionary trajectory opens the way for novel long-lasting interactions. Such evolutionary events also reshape the structure and dynamics of communities, depending on the spatial heterogeneity of the common good production.

The second part of this section presents a perspective on how to potentially unveil such interactions, by focusing on metabolic contingencies existing between organisms.

[1] maybe because the selection pressure is not as strong as it is not *survival* but *better surviving* that is at stake, making the emergence of this type of coevolution less frequent or slower

Chapter 1

Beyond the Black Queen Hypothesis



The ISME Journal (2016) 10, 2085–2091

www.nature.com/ismej

MINI REVIEW

Beyond the Black Queen Hypothesis

Alix Mas[1], Shahrad Jamshidi[2], Yvan Lagadeuc[1], Damien Eveillard[2] and Philippe Vandenkoornhuyse[1]
[1]*Université de Rennes 1, CNRS, UMR6553 EcoBio, Rennes, France and* [2]*Université de Nantes, EMN, CNRS, UMR6241 LINA, Nantes, France*

This work has been published in ISME Journal as a 'mini-review'.

1.1 Introduction

Popular among theories of ecology and evolution,the Red Queen Hypothesis (Van Valen, 1973) has recently been echoed by a new hypothesis: the Black Queen Hypothesis (BQH; Morris et al., 2012), which concerns the evolution of dependency between organisms.

While the Red Queen Hypothesis sets the basis for (mostly antagonistic) co-evolution, the BQH renews and puts into perspective current understanding of the evolution of interactions between free-living organisms within microbial communities. The BQH marks a turning point in modern (micro)biology and ecology. This novel reductive evolution theory describes evolutionary mechanisms potentially at the origin of the connectedness between organisms in a community. More precisely, it provides theoretical interpretations of the evolution of dependencies through adaptive gene loss in free-living organisms.

In the BQH context, some free-living organisms "avoid" having a function in order to optimize their adaptation to the environment; such organisms are called beneficiaries. This loss of function is made possible because other organisms in their close environment (the helpers) publicly and continuously provide for the function, offering a (partially) stable environment. The mechanism underlying this particular loss of function is genome reduction through gene loss. This type of evolutionary process is of major significance in the context of long-lasting interactions, as beneficiary species develop a strong dependency on the function provided by helper species. The fact that it is impossible to grow monocultures of most bacterial species, referred to as the "uncultured microbial majority" (Giovannoni et al., 2014) , could result from such dependency-based gene loss, as the species are unable to grow when extracted from their community.

After its introduction, the BQH was echoed in papers describing the evolution of organisms through adaptive functions and gene loss (Ellers et al., 2012; Giovannoni, 2012; D'Souza et al., 2014; Giovannoni et al., 2014; Luo et al., 2014), the evolution of interactions (Estrela et al., 2012; Hussa and Goodrich-Blair, 2013), notably cooperation (Sachs and Hollowell, 2012) and evolution of the community (for example, Sachs and Hollowell, 2012; Mitri and Richard Foster, 2013; Hanson et al., 2014). It also triggered one dedicated evolutionary experiment (Morris et al., 2014) and uncovered new possibilities regarding the outcomes of other evolutionary experiments (D'Souza et al., 2014; Hosoda et al., 2014). One notable point of the BQH is the association of adaptive genome reduction with free-living organisms (that is, organisms living independently of any host), a phenomenon

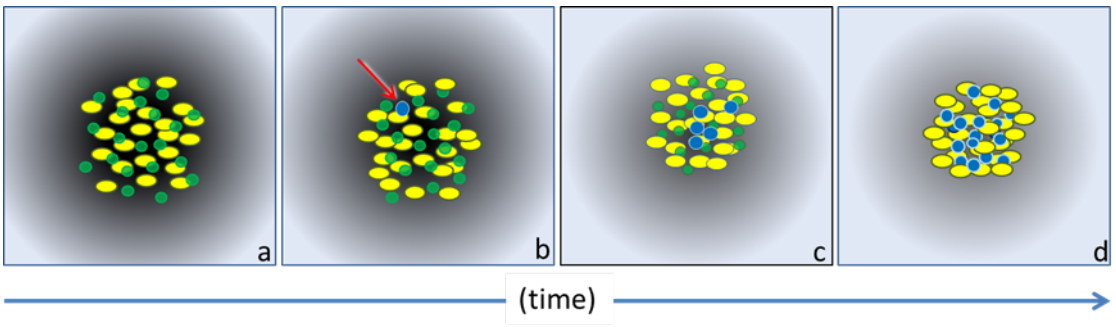


Figure 1.1: **Overview of the BQH.**
The yellow and green spots represent two types of microorganisms, namely A and B1 in the next figures, both producing a common good. A mutant that has lost the capacity to produce the common good is shown in blue (B2 in the next figures). The shade of the background (dark gray) expresses the concentration of the common good in the environment. (a) Initial state: A and B are present. (b) As the common good is extensively produced in the environment, a mutant strain (B2, blue) no longer able to produce the common good emerges; it is dependent (that is, beneficiary) on other microorganisms producing it (that is, helpers). (c) As the beneficiary's fitness is improved (no energy invested in the production), it invades the population (d) and a new equilibrium is reached in the community between the helpers (yellow) and beneficiaries (blue), which supplanted their ancestors. Figure inspired from Bjørn Østman, 2012, http://pleiotropy.fieldofscience.com/2012/05/black-queen-hypothesis.html.

that had not been apparent before. This evolutionary event of adaptive genome reduction stems from a particular type of interaction: the use of a common good (that is, a freely available element present in the environment) (Figure 1.1).

Being a recent hypothesis, the BQH has not been thoroughly tested. Nevertheless we focused on the development of new ideas related to this hypothesis. We detail herein the possible mechanisms driving the observed loss of genes, and then use a modeling approach to fathom the consequences of this adaptive genes loss on population dynamics, at the community scale and in different ecological contexts. Thus our aim was to explore the BQH beyond its original definition. A corollary of the BQH is also introduced.

1.2 Material and Methode

We tested the Black Queen Hypothesis using an Agent Based Modeling approach in different interaction contexts.

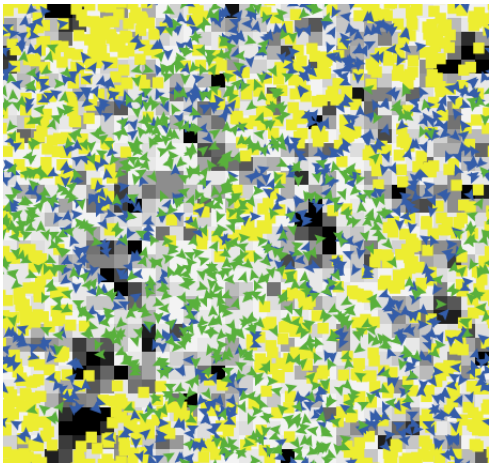


Figure 1.2: **Example of the visual output of the model.** In this output of the model, we can see species A (yellow) and species B1 (green arrow) which are potential helpers, while the blue arrows represent species B2, the beneficiary mutant. The dark squares are patches depleted of the common good, while the white squares are saturated with the common good.

1.2.1 Agent Based Modeling

Agent Based Modeling (ABM) is a type of computational modeling based on the simulation of the behavior of entities ('agents') able to perform tasks autonomously (e.g., production of an element, reproduction). This type of simulation makes it possible to observe the interactions-based evolution of a dynamic/community system over logical time.

1.2.2 The agents and their world

In our model, we consider three agents: A, B1 and B2 (resp. yellow squares, green and blue arrows in Illustration 1) that are located in a two-dimensional world. The two-dimensional world is divided into identical patches, where each patch color represents the amount of common good (M) available; The amount of common good is discretized and its value can vary between 0 (black) and 9 (white), where 9 represents the saturated concentration of M and 0 refers to an absence of M in the patch (Illustration 1). Production of the common good, M, is performed by both agents A and B1. Therefore, when either agent A or B1 is present, the amount of M in the patch increases by one per time step. If A or B1 are absent, then the M concentration for the patch decreases by one per time step. Organism B2 is a B1-mutant that does not produce the common good M. Nevertheless, B2 still requires and thus can benefit from the common good produced by surrounding As or B1s.

1.2.3 Parameters

Each agent is independent and increase its age at each logical time step. Each agent type is associated with four parameters: a density value, a latency period, a lifespan value and a minimum quantity of common good (M) required for growth and reproduction. The density value corresponds to the upper limit of an agents' abundance in a small radius, which is representative of the limited resources in a given area. The latency period corresponds to the time before an agent can reproduce, after which the agent will reproduce at every time step. Finally, lifespan is the number of time steps during which the agent can reproduce (the time after latency and before death).

1.2.4 Fitness advantage

Relative to B1, the loss of M production machinery and associated genes confers B2 with a fitness advantage. This fitness advantage is created in the model so that B2 has one out of two chances to reproduce twice instead of once per time step. That is, B1 and B2 have the same qualities but B2 reproduces more than B1. Observing the dynamics of A without the presence of B1 or B2, the steady state of each species is mainly influenced by the chosen values of density and lifespan, and only to a lesser extent by the minimum quantity of common good required for growth and reproduction and the adulthood values (see Supplementary information Figure 1.6). Because A and B1 are governed by the same rules, the dynamics of B1 on its own are the same as for A on its own.

1.2.5 Scenarios

Several scenarios are discussed. First the behaviors of respectively B1 and B2 were studied separately. The steady states reached by the two populations can be used as a proxy to study population dynamics. Then, one investigates the introduction of mutants B2 in a community where other species (A) also produce the required common good. The case of interaction between species tested here is without competition. In this case of no competition, the density of A is independent of B, and the only influence of B on A is via the metabolite M.

1.2.6 Statistics

For each set of parameters, normality was tested and non-parametric Kruskal-Wallis tests were performed using the software R.

1.3 Results and discussion

1.3.1 Size matters in the Black Queen Hypothesis.

A key element in this hypothesis is the loss of genes, which expresses a selected genome reduction. While evolution is usually associated with genome complexification (Wolf and Koonin, 2013), in the BQH the source of evolutionary opportunities is simplification through genes loss. Typically, gene loss results from two different forces: genetic drift and positive selection.

Genetic drift or positive selection for gene loss in the BQH?

Genetic drift refers to changes in allele frequencies of a population due to random sampling. The ability of drift to influence allele frequencies is inversely proportional to population size. By definition, natural selection increases fitness whereas genetic drift operates at random, and only occasionally confers patent fitness benefits. If a function becomes useless, the selection pressures on genes involved in that function are lifted. In the absence of purifying selection acting on these genes, circum-neutral mutations can accumulate (McCutcheon and Moran, 2011) and be randomly retained by genetic drift in a small population, leading to the decay and eventual loss of these genes. Given the large size of the populations encompassed by the BQH, genetic drift is excluded as the main driver of evolution (Morris et al., 2012). One example of the BQH concerns the most abundant photosynthetic organisms on Earth: Prochlorococcus (Partensky et al., 1999).

General genome reduction and removal of specific gene

Gene loss is known to be selected via two distinct operating modes: it can be favored by (i)general genome reduction, and (ii)the removal of specific targeted gene(s). The two systems most likely act together. Positive selection for genome reduction refers to genome streamlining sensu stricton (Ochman Moran, 2001; Giovannoni et al., 2005) , which is defined as the selection process that [. . .] acts to reduce genome size because of the metabolic burden of replicating DNA with no adaptive value [...](Giovannoni et al., 2005). Indeed, every function has a constitutive cost, considering both genomic content and metabolism (Giovannoni et al., 2005, Lahti et al., 2009; Kreft Bonhoeffer, 2005; Driscoll et al., 2011) but this cost is usually offset by the fitness benefits the function provides(Giovannoni et al., 2014). If a function loses its beneficial effects it will eventually be purged to

reduce energy costs. Because retaining a bigger genome is costly (maintenance, replication, regulation), positive selection for genome reduction is assumed to be the main driver of genome reduction, at least in oligotrophic environments (Giovannoni et al., 2014; Dufresne et al., 2005; Hottes et al., 2013). A reduction or elimination of redundant and useless genetic material will occur from one generation to the next. Consistently, freeliving microorganisms are reported to often experience genome reduction via a loss of paralogues for multi-copy genes (Porter Crandall, 2003). Thus in the genome streamlining theory (Ochman Moran, 2001; Giovannoni et al., 2005), selection favors gene loss, as smaller genomes provide more adaptive advantages than bigger ones. Recent experimental tests of evolutionary dynamics have shown that it may not be genome reduction itself, but particular and individual gene loss that confers the greatest advantages (D'souza et al., 2014, Cooper et al., 2001; Lee Marx 2012, Pande et al., 2013), especially if the cell's lack of metabolite production is compensated by the habitat (D'souza et al., 2014). The fitness gained from a given gene deletion is dependent on its metabolic function (e.g., which metabolite it codes for) and on the position of the deletion in the metabolic pathway. For example, the deletion of genes at the end of a bio-synthetic pathway could be more advantageous than the deletion of anterior genes (D'souza et al., 2014). Thus the energy costs linked with anabolism could be efficiently reduced by the removal of specific targeted genes.

1.3.2 Specialization towards common goods consumption

A special interaction: the circulation of a common good

Every species in a community is linked to one or several other species, thereby forming an intricate web of direct and indirect interactions, metaphorically described as a tapestry in which the weaving (i.e., interactions between species) is as important as the species themselves (Estes et al., 2013). In the BQH (Morris et al., 2012) species evolutionary dynamics are based on indirect interactions through common goods utilization. Common goods are freely available elements present in the environment for both the producing species and other species around. The producers of common goods can still have preferential access to these goods (Estrela et al., 2015), thereby avoiding the emergence of a "tragedy of the commons" situation (Hardin, 1968). Thus, the interaction between "helper" and "beneficiary" species is a case of indirect symbiosis existing through the flux of a common good, and the evolution of interactions between

the two species can be considered as a side-effect of common good consumption. Within the community, helper individuals transform their microbial vicinity into a stable and homogeneous place (by continuous production of a common good) allowing beneficiary mutant organisms to follow this adaptive path of specialization.

The rise of the mutant

To further understand the Black Queen's dynamics we used an agent-based model (ABM, supplementary information). We based our simulations on one or two species, then we introduced a mutant with higher fitness than its ancestor in order to observe (i) the temporal and spatial patterns of its invasion in the community (ii) the dynamics of the simulated organisms. These simulations showed that when a fitter mutant unable to synthesize the common good emerges in a population (i.e. a loss-of-function mutant) this new strain will supplant its ancestors. But if the ancestors are the only helpers around, the beneficiary mutant population never excludes its original population due to its vital dependence on the production of the common good (Figure 1.3a). Thus the BQH also fits for a single species model. Nevertheless, if another species in the community produces enough of the common good, the mutants (because they are fitter) will replace the ancestors (Figure 1.3b). In our simulations, the helpers' population is always sustained and their density (Supplementary Information) reaches a steady state (Figure 1.5a). The dependence of the beneficiaries on the helpers forces both populations to be in equilibrium. This confirms the BQH prediction (Morris et al., 2012) whereby a loss-of-function mutant will be able to expand within its ancestral population if the function loss gives the mutant a growth advantage over its ancestors but the mutant retains a need for the common good.

Helper or Beneficiary?

What, in a community, determines which species will evolve into a beneficiary or into a helper? The current proposition by Morris et al. (2012) is that beneficiaries are species that evolve the most rapidly. We propose a new idea. Our simulation reveals that when the minimum value of a common good needed for the mutants to multiply increases, it lowers the density of the mutants at steady state, while it increases the density of the helper species(Figure 1.4, b). This is tied to the fact that increasing the need for a common good increases the dependency of

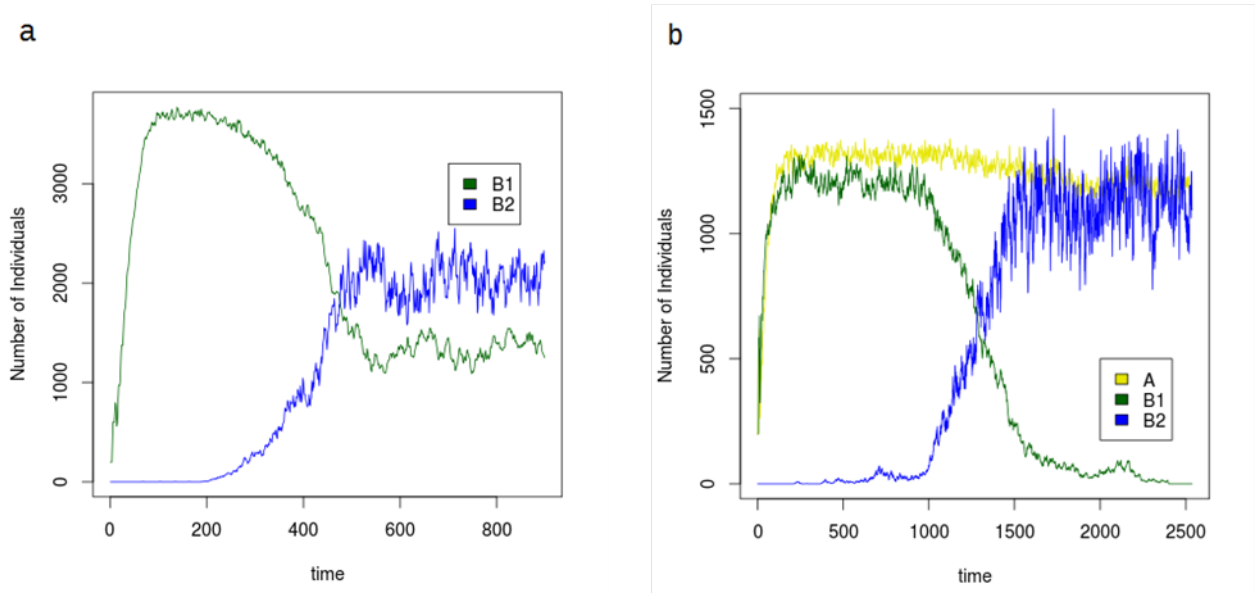


Figure 1.3: **Trajectories for two species populations (A and B1) when a loss of function mutant (B2) arises in one population.** Trajectories for two species populations (A and B1) when a loss-of-function mutant (B2) arises in one population. The trajectories shown represent a single simulation that measures the population of each species over time as defined by the species' rules given in the Supplementary Information. **(a)** As the fitter mutants B2 (blue) emerge within their original population B1 (green), they will increase to the detriment of this original population. However, because the mutants still depend on the common good produced by the helpers (B1s), they can only spread if there are enough B1s present to produce the common good, leading to an equilibrium state of helper and mutant populations. **(b)** When the fitter mutants arise from their original population B1, if another species (A, in yellow) also produces the common good, then the B2 population will not be exclusively dependent on the B1s, and (B2s) will entirely supplant the B1s, if the B1s provide enough of the common good to sustain the B2s.

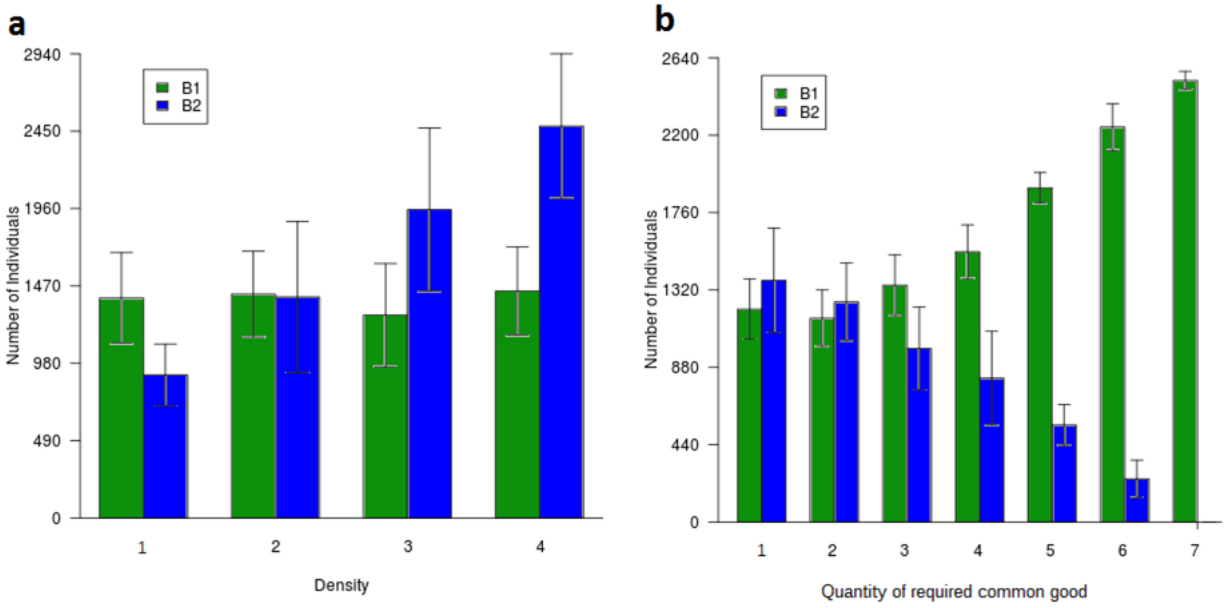


Figure 1.4: **Steady state populations of the original population B1 (helper) and of the fitter mutant B2 (beneficiary), with changing parameter values of density and of quantity of common good required.** The population of B1s in green (the original population, which became a helper population) and B2s in blue (the beneficial mutants) is shown after they have reached equilibrium. Fifty simulation replicates were performed for each modality, and data were collected after 500 time-steps (Supplementary Information). The default values for the fixed parameters were in (a,b) reproduction latency=3, lifespan=3; (a) minimum of common good required=5 and (b) density=3. Both (a) the density value (in arbitrary units, corresponding to the upper limit of individuals living in a small radius) and (b) the quantity of required common good by beneficiaries (arbitrary units) are altered to see how the equilibrium changes. Whiskers show confidence intervals (alpha=0.5) of the means. **(a)** The density is indirectly representative of the nutrient resource available: if more resource is available, more organisms can live together on a given surface unit. Because the beneficiaries B2 are fitter, they will supplant the B1s when more nutrients are available (Kruskal−Wallis test, $P< 2.10^{-16}$) while B1 density hardly changes (Kruskal−Wallis test, P=0.0191). However, because the B2s are dependent on the common good, they can only spread if there is enough common good produced, thus the helpers' population (B1) cannot be excluded. **(b)** When the minimum quantity of common good required by beneficiaries B2s is higher (i.e. the mutants have a greater need in common good) the B2s will be more dependent on the helpers B1. As the need in common good for the B2s increases, their population density at steady state is lowered (Kruskal−Wallis test, $P< 2.10^{-16}$) and conversely for the B1 (Kruskal−Wallis test, $P< 2.10^{-16}$).

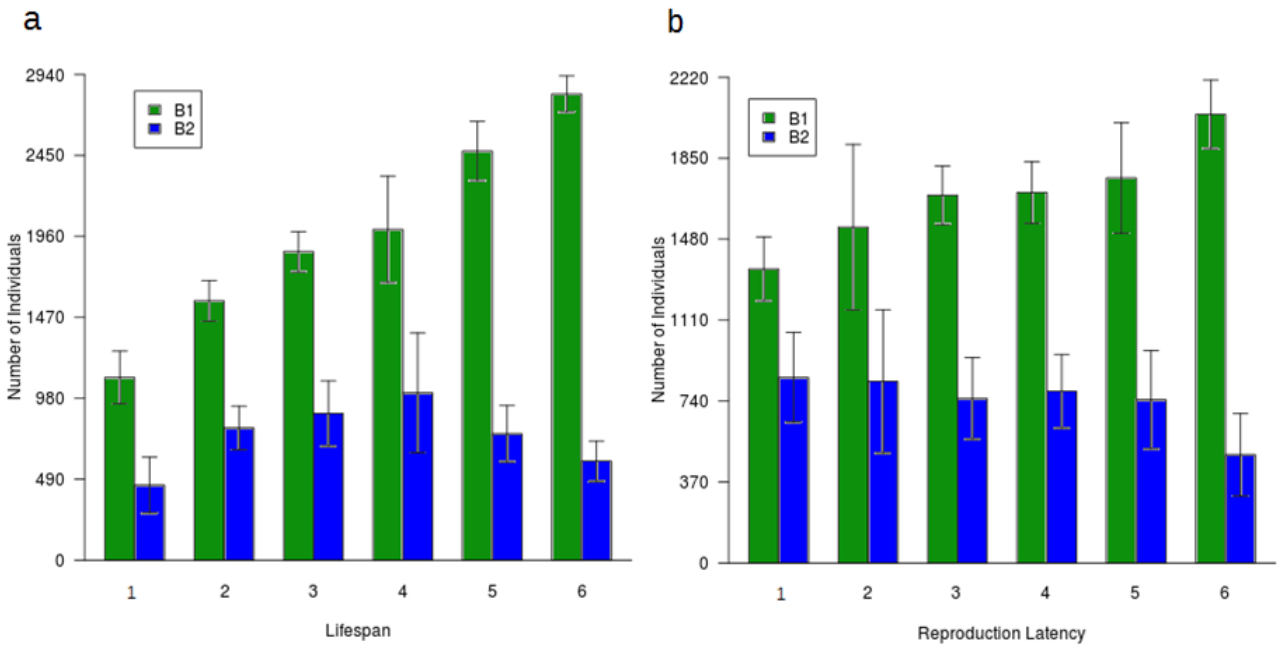


Figure 1.5: **Steady-state populations of the original population B1 (helper) and of the fitter mutant B2 (beneficiary) with changing values of lifespan and reproduction latency parameters, a and b, respectively.** The population of helpers (B1s) and beneficiaries (B2s) is shown after they have reached equilibrium. The lifespan (corresponding to the time available for reproduction, in arbitrary units) and the reproduction latency (the time before individuals can reproduce, in arbitrary units) were altered to see how the equilibrium changed. The populations of helpers (B1) and beneficiaries (B2) were sampled at 50 different time points after introducing the mutant (B2) and when an equilibrium state (500 time steps) was reached for each set of parameters. Each bar shows the mean of the 50 time points and the whiskers represent the 95% confidence interval. The default values of the fixed parameters were as follows: lifespan=3, reproduction latency=3, density=3, and the minimum quantity of common good required by beneficiaries= 5. **(a)** With increasing lifespan (Kruskal–Wallis test, P= 2.10^{-16}) a B1 individual is guaranteed to have enough common good in the final time steps of its life. A B2 individual, however, needs to coexist with a B1 on the same patch for there to be enough common good. That is, the chances of a beneficiary individual reproducing decrease because of the greater dependency on helpers. **(b)** Similar reasoning as for lifespan holds for the reproduction latency.

the mutants on helper species producing this common good. As dependency is constraining, we suggest that the fittest of the loss-of-function mutants will be the one on which a rise in dependency will have the least effect, i.e. for which the loss of function is less enslaving. Nevertheless, if the function is accessory, it could be entirely lost by the community, and the hypotheses would no longer stand, suggesting that the BQH might only be valid for functions vital to the organisms. We suggest that a potential loss-of-function mutant may be a species with a 'silently advantageous' trait, which will only fully express its benefits after a concomitant mutation. Such a trait could be a minor need for the common good or a faster consumption rate for example. Thus, in species adopting a beneficiary trajectory, the level of dependency on the helper might be low or potentially facilitated.

1.3.3 Effects of the Black Queen trajectory : transformation of interactions and of the community

We propose that the BQH offers new perspectives as to how evolution can modulate community life. Notably: the emergence of a mutation within a population can (i) transform its interactions with other species and (ii) deepen community life and modify the global dynamics of the community.

The Black Queen as a way to elude competition

Niche partitioning (through environmental filtering or through species interactions sorting) and neutral processes (Hubbell, 2001) are classically acknowledged to drive community assemblies and explain diversity patterns. When competition occurs in spatially and temporally homogeneous environments, coexistence is mainly assumed to result from complementarity in resource use (Webb et al., 2002). In addition to these classical explanations of assembly rules we suggest that the transition from competition to dependency relationships may also be a driver of community structure. In our simplified two-species model, the emergence of a helper-dependent mutant shifts competition toward coexistence. In this situation, helpers and beneficiaries will reach a state of equilibrium (Figure 1.4b) that bypasses exclusive competition. The fact that long-term coexistence is reached through the evolution of dependency has already been demonstrated: de Mazancourt Schwartz (2010) showed that resource trade enhances coexistence even if it decreases the abundance of one of the species (consistent with Figure 1.4b). In a similar way, Turcotte et al., (2012) showed in a context of

waste-product exploitation that dependency increases the coexistence of species. Cross-feeding interactions are also known to evolve in competitive environments (e.g. Friesen et al., 2004; Louca and Doebeli, 2015). To put it simply, in a system where several species are in competition, if a beneficial dependency emerges between two species (through resource trade, waste-product or common good use), their coexistence will be optimized. Here, natural selection will favor the establishment of tighter relationships and confer an advantage to coexistence.

One change changes it all

In simulations where two species are able to sustain themselves irrespectively of the other's presence, an emerging loss-of-function mutant can invade the population and eventually replace the ancestral strain of that population(Figure 1.3b). We observe that the population density at steady state depends on life history traits such as the quantity of common good required (Figure 1.4a), lifespan(Figure 1.5a) and reproduction latency (Figure 1.5b). Increases and decreases in population sizes are attributed to a density dependence effect (Figure 1.4a, 1.5a& b) related to the parameters used in the model. The dynamics of the loss-of-function mutant and extinction of the original population will thus be dependent on such life history traits to reach a steady state. Metabolic dependencies are potentially a major driver of species co-occurrence (Zelezniak et al., 2015). Until recently, the repercussions of such evolution at the community level were mostly overlooked, whereas it is now becoming evident that they are an integral part of community systems (Hairston et al., 2005; Johnson Stinchombe 2007; Schoener, 2011) implying that evolutionary dynamics should systematically be taken into account when characterizing communities.

Dependence and consequences on the quantity of common goods produced

We propose that the quantity of common goods produced by helpers could impact the spatial dynamic of the entire community. In a scenario where the common good is abundant, the distribution of helpers and beneficiaries is expected to be homogeneous at steady state. Conversely in our simulation where the common good is produced at a limiting concentration, the dynamic invasion of space is uneven (MOVIE S1). The spatial distribution of beneficiaries, because of their dependency, is expected to adhere closely to the distribution of the helpers, inducing a local depletion of nutrients availability. Thus, the heterogeneous spatial 'aggregation' of (micro)organisms and resulting heterogeneity

of nutrients availability will lead to temporal changes and patches displacement (MOVIE S1). Spatial and temporal heterogeneity may thus be a consequence of helper/beneficiary interaction.

1.3.4 Going further

The BQH, smoothing the way for long-lasting interactions?

Studies of co-evolution, taken in the broad sense where one organism evolves in relation to another, have tended to focus on the acquisition of traits and functions, with little attention given to the loss of functions in free-living organisms. Consequently, the mechanisms behind the emergence of a compensated loss of function remain unknown. It has been assumed that trait loss [...] is only expected to evolve in long-term, stable co-evolutionary physiological relationships[...] (Visser et al., 2010). However, we consider that the BQH offers a new approach: function loss, instead of resulting from long-lasting relationships, could actually be the cause of such long-lasting co-evolutionary relationships. Indeed, in the BQH, a 'passive' interaction (production and consumption of a common good) is at the basis of the emergence of tighter dependencies between two species. We perceived that initially a mere coexistence of species within a community happens to be fortuitously beneficial for one of the species (via the redundancy in production of the common good). Then if the beneficial conditions remain sufficiently stable over time, the (future) beneficiary species will tend to lose the compensated function. More than the trait loss itself (Ellers et al., 2012), we believe that the 'point of no return' in the shift from a facultative interaction to an obligatory dependency is the loss of gene(s) underlying the loss of function. In this case, genome reduction sets up the first steps for stable long-lasting metabolic interactions (Pande et al., 2013). The loss of essential genes is widespread in free-living bacteria and most certainly at the root of inter-organisms networks (D'souza et al., 2014). In line with this, some vitamin exchanges are known to cement species connectedness within the community (Giovannoni, 2012) and it has also been inferred from the BQH model that cooperative interactions could rise "automatically" in this context (Sachs and Hollewell, 2014).

Corollary of the Black Queen Hypothesis

A species, by being dependent on another one, puts itself in a weakened position. If no helper species is around, the species cannot survive. What is more, such

species can be considered as "accessory" for the community since they do not provide for the essential function. Admittedly, accessory species could be wiped out more easily than essential species since their loss will not have an immediate effect on the community. On the other hand, being able to handle essential functions, even if more costly, guarantees some degree of 'security'. First it allows for resistance against environmental changes or perturbation. Secondly, in a dependency context, helpers are indispensable, which will partly prevent them from being replaced by competing species (Nadell et al., 2009). It is therefore of particular interest to keep, or even to acquire, the status of helper. In their paper, Morris et al (2012) suggest that helpers are keystone species in the community. We propose a possible corollary of the BQH: species could benefit from ensuring the status of helpers by having mutations that enhance the production of the common good, or more extremely, by acquiring genes producing the common good. The occurrence of mutation which enhances the common good production could also explain why some species turn into helpers while others become beneficiaries. Horizontal gene transfer (HGT) is a common phenomenon in (micro)organisms (Mc Ginty et al., 2010) , often leading to the acquisition of functions and suggested to be an underlying mechanism of microbial cooperation (Smith, 2001). If the enhanced or acquired function leads to the development of a dependency interaction, then (micro)organisms possessing it may be better off, thanks to the key status acquired within the community. This substantiates the suggested BQH corollary, i.e., genome expansion for helpers who tend to embrace a generalist ecological status. Thus the BQH and its corollary invite a new interpretation of the networks of interactions and ecological status (i.e., generalist/specialist) of co-occurring organisms and their evolution.

1.4 Supplementary Material

1.4.1 Figure S1

| Steady state reached by a population when the values of reproduction latency, density, lifespan, and threshold are altered. The steady states represented show how the equilibrium of the population changes when values (in arbitrary units) of reproduction latency (the time before individuals can reproduce), density (upper limit of individuals in a small radius), lifespan (corresponding to the time after latency and before the individual dies, i.e., the time available for reproduction) and the minimum quantity of common good required for growth and reproduction, are modified. a.| With longer reproduction latency, each individual ensures its ability to reproduce due to a sufficient production of common good produced by themselves during reproduction latency. That is, each individual is more independent of others with increasing reproduction latency and can reproduce more, which is why we see the slight increasing trend. b| More individuals are allowed per small neighbourhood so the same space can be more densely packed with individuals. The parameter of density can be seen as the nutrient resource, where if more food is available then more organisms can live in a small area. c| For longer lifespans, individuals are dying later and reproducing longer. That is, the probability of replacing a dead individual is higher than for shorter lifespans because every neighbouring individual is ready to give birth to take its place. (This is analogous to limited parking spots in a big city. As soon as a parking spot opens up it is immediately filled because there are so many people ready to park). Therefore, each small neighbourhood holds the maximum number of individuals as dictated by the density. d| Increasing the minimum quantity of common good required by beneficiaries does not have a significant influence on the population. Only for the very high quantities of common good required by beneficiaries can we see a slight decreasing trend because in the case of a higher quantity of common good required by beneficiaries the individuals are more dependent on the common good.

Caption Figure S1 : The population of B1s in green (the original population which became a helper population) and B2s in blue (the beneficial mutants) is shown after they have reached equilibrium. Fifty simulation replicates were performed for each modality and data were collected after 500 time-steps (see supplementary information). The default values for the fixed parameters were in (a) and (b) reproduction latency=3, lifespan=3, in (a) only, minimum of common good required=5 and in (b)

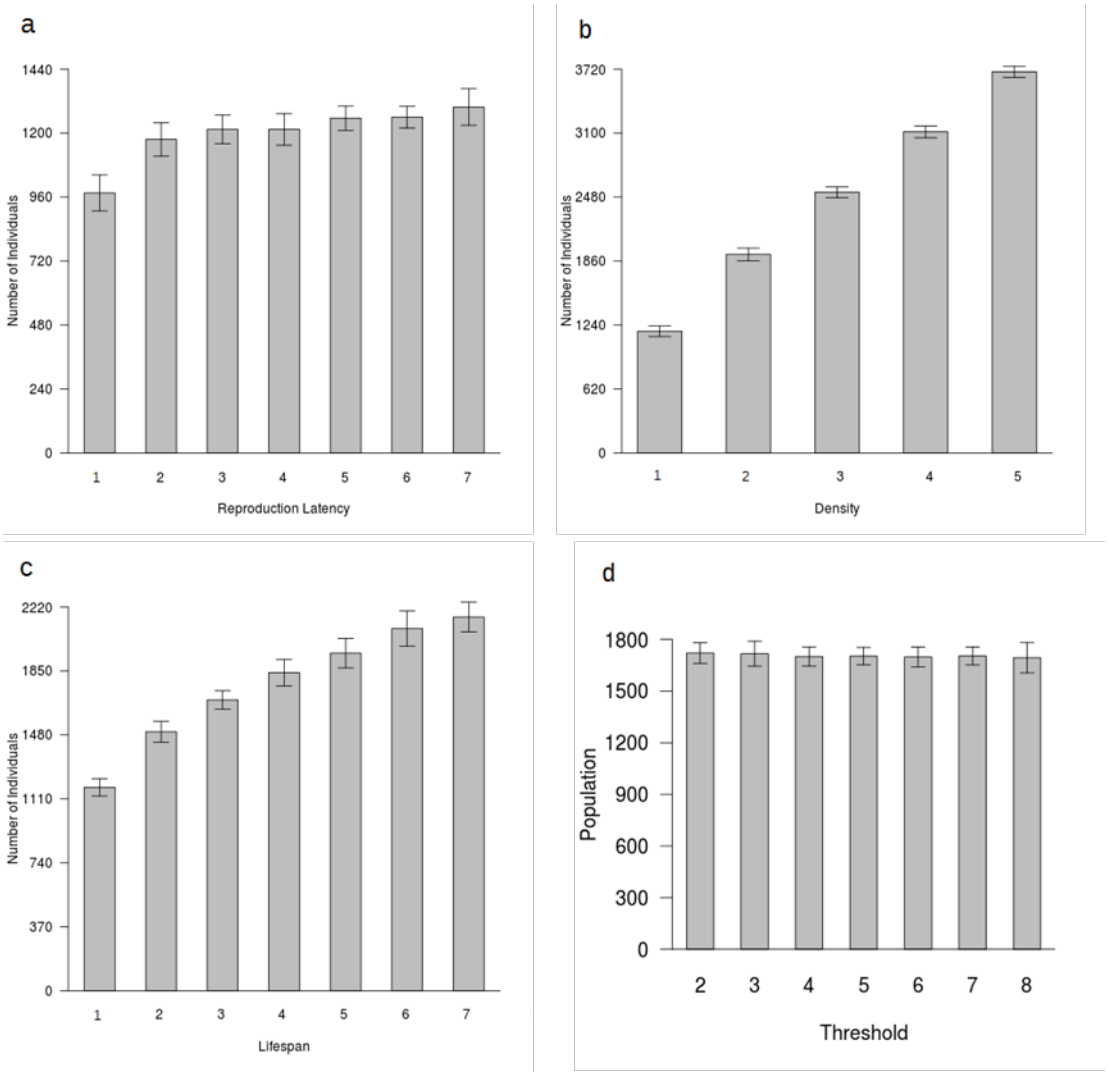


Figure 1.6: **Steady state reached by a population when the values of reproduction latency, density, lifespan, and threshold are altered.**

only density=3. Both (a) the density value (in arbitrary units, corresponding to the upper limit of individuals living in a small radius) and (b) the quantity of required common good by beneficiaries (arbitrary units) are altered to see how the equilibrium changes. Whiskers show confidence intervals (alpha =0.5) of the means. **(a)** The density is indirectly representative of the nutrient resource available: if more resource is available, more organisms can live together on a given surface unit. Because the beneficiaries B2 are fitter, they will supplant the B1s when more nutrients are available (Kruskal-Wallis test, P<2.10-16) while B1 density hardly changes (Kruskal-Wallis test, P= 0.0191). However, because the B2s are dependent on the common good, they can only spread if there are enough helpers (B1) present to produce the common good, thus the helpers' population cannot be excluded. **(b)** When the minimum quantity of common good required by beneficiaries B2s is higher (i.e. the mutants have a greater need in common good) the B2s will be more dependent on the helpers B1. As the need in common good for the B2s increases, their population density at steady state is lowered (Kruskal-Wallis test, P<2.10^{-}16) and conversely for the B1 (Kruskal-Wallis test, P<2.10^{-}16).

1.4.2 The Movies

Video S1 | Temporal dynamics of the invasion of two populations by a beneficiary mutant. The dynamic(s) of three populations (two potential helpers and a beneficiary mutant) are represented in this video. Each patch colour (9 shades of grey from white to black) represents the amount of common good present in the environment: white = patch saturated in common good; dark = patch depleted in common good. The yellow squares represent a population of species A, green arrows represent the population of a species B1 and blue arrows (appearing at 0.02 seconds) represent the beneficiary mutants B2 emerged from the population B1. Both populations of As and B1s are common good producers. The mutant does not produce the common good but does consume it. After the appearance of the beneficiary mutants (in blue, t=0.01sec) we can see how they gradually replace their original population (green) until they totally supplant it (t=20sec). It is apparent that the distribution of both common good and species is not homogeneous, creating dark patches in the visual output of the model.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)

Chapter 2

Perspective & Conclusion

2.1 How to unveil and predict potential co-evolution and interactions

The papers presented above (Morris et al, 2012, Mas et al, 2016) pinpoints that shared functions existing between organisms in a population can lead to the evolution of tight long-lasting interactions determined by genetic modifications, and fitness increase. As for the egg or the chicken, in coevolution the question of what happened first stands: Is it a fortuitous genetic modification carried by one (or both) organisms interacting that entails a fitness advantage in the coevolution, rendering the interaction specific and obligatory? Or is it a progressive process in which the organisms first adapt their behaviors, their functions, their metabolisms to each others, leading to advantageous fitness and finally being genetically integrated? This question is not to be answered here, but both options can be explored to understand the evolution of communities.

Staying in the concept of evolutionary modification among co-occurring microorganisms, and based on the hypothesis of gene losses and evolution of dependencies, a prospective idea is developed below using a metabolism-based approach of biological systems, to detect obligatory interactions, and to predict the evolution of communities.

A prospect for the understanding of microbiota complexity and organization

Usually, the evolution of gene loss is detected through phylogenetic analyses: if there is a conspicuous lack of genes in one of several closely related species, it is assumed that the gene was lost. In some cases, this loss of genes can be associated to loss of functions, or modification of traits coherent with the environment (Cooper et al., 2001; Kettler et al., 2007; Pande et al., 2013).

A preliminary study compared the metabolic functions performed by a microbial community (microbiota), with the set of metabolic functions that could actually be performed by individuals only. By reconstructing the metabolism of a community, it is possible to predict the functions performed at the community scale. Nevertheless, while looking at each singular genomes, it appeared clearly that some of the functions cannot be realized by single genomes, i.e by single organisms. Such functions, that are not supported in their entirety by any single organism, strongly suggest that the realization of particular functions in the community is depending on the interaction and the association of several organisms' metabolism. What is more, preliminary data analyses on the community

strongly support that these complementary functions at the community level are paired with strong co-occurrences of particular microorganisms. Individuals' genome analyses of these particular co-occurring bacteria indicated metabolic completions for particular functions.

Thus, it is most likely that synergies between organism will permit the realization of more reactions altogether than the reactions of each organism taken separately. It recalls how important it is to consider, as much as possible, the systems in their whole, as they reveal emergent properties not conceivable when focusing on sub-part only. There is potentially a metabolic complementarity, that could be depicted as a factory 'chain work', between organisms of a same community. Nevertheless, a crucial point here, is the fact that microorganisms are physically distinct entities and thus this metabolic complementarity is only possible if the co-occurring species are able to share metabolites by diffusion processes, excretion/uptake and associated transporters.

Predicting the emergence of co-evolution

Focusing on such shared metabolisms, and based on all the genomic knowledge acquired and stored, it is possible to determine the metabolic genes necessary to perform a given function. Thus, if we observe (in a exaggeratedly simplified way) that some organisms carry half of the genes needed to perform a community-level function, and other organisms carry the other half, we can assume that these organisms act in interaction (even if passively) in the production of this community-level function. The metabolic approach considered could permit to focus on targeted organisms, to see if this metabolic complementarity is obligatory or not, and beneficial or not. Recent work (Bordron et al., 2016) proposed an equivalent approach to determine the functional roles of organisms within a consortium of five bacteria.

From another angle, it is also possible to predict that in a community where some metabolic pathways are redundant, the evolution of a redistribution of the metabolic 'tasks' between sub-populations could emerge if they permanently interact.

To get back to the BQH, as was shown in the first part of this chapter, the rise of dependency of an organism to another one through the loss of a function, and more precisely through the loss of genes, can explain to a certain degree the assemblage and co-occurence of species in a community (Mas et al., 2016). Thus, one important prospect would be to look at the metabolisms of co-occuring organisms (e.g. micobiota), to infer the genes at stake, to predict possible emer-

gence of metabolic complementarities impacting the of one or several actors of the interactions, we could eventually predict the emergence of BQH types of coevolution through the loss of genes implied in 'abandoned' metabolic function, and thus predict the modification of their respective population and of the community which could reach new equilibrium.

Therefore, here we suggest a conceptual approach, based on metabolic inferences, which could both unveil the existence of BQH types of interactions (i.e interactions based on metabolic complementary and inducing some dependence via genome modifications) and permit to predict the evolution of a synthetic community through metabolic optimization at the community scale. Let us just remember here that evolution, as well as individuals, populations and communities are all dynamics systems, and therefore there is not a unique answer, no unique state of equilibrium : everything is continuously changing and evolving.

2.2 Conclusion on the chapter

This section of the thesis showed that focusing on metabolic insights such as the rise of dependency on a common good Morris et al. (2012b), we could predict the dynamic of a population undergoing such evolutionary trajectory (Mas et al, 2016). With yet another approach, we also demonstrated that still through metabolic contingency, we could potentially explain (and thus partially predict) the co-occurrence of species within a community. These are thus two complementary approaches, both relying on metabolic resolution that can help predict dynamics of evolution at the population and community scales.

Part III

Strategy to Study Evolution and Prediction

Chapter 1

Introduction of the Section

It is commonly known that organisms subjected to particular environmental constraints (which are not perturbed) will follow an evolutionary path of specialization in which only a subset of their fundamental functional capacities will be conserved to optimize responses to these constraints (Tripp et al., 2008; Giovannoni et al., 2014). For example, several population of organisms (notably fishes) were found to have evolved in caves, and phylogenetic comparisons with out-of-cave-living closely related organisms showed that the switch to a cave mode of life entailed the loss of functions such as the capacity to see or body pigmentation (Gompel and Prud'homme, 2009; Conway Morris, 2010). It can be assumed that such populations of organisms, living in pitch dark environments would not be hindered by the loss of vision, since they already cannot see in caves. In this case, the environmental constraints maintaining vision were lifted, and the system of vision was lost without burdening the individuals and potentially providing them with the benefits of resources and energy reallocation to other functions (*i.e.* ecological trade-off (Bono et al., 2017)). More generally this adjustment to ecological constraints can be expected to be observable for any non-vital function that cannot, or does not need to be expressed in a long lasting stable environment. Such streamlining can be assimilated to functional or metabolic specialization.

To some extent, we considered the BQH rise of dependency as the consequence of a specialization event to the common good constent production in the environment (see Morris *et.al*, 2012; Mas *et.al*,(2016) and Section 2 of this manuscript). Indeed, as seen through our *in silico* approach of the BQH in the second part of the thesis, if we consider the optimization of the metabolism of organisms as a driving force of evolution, we could potentially explain (and thus partially predict) the co-evolution of species within a community. This co-evolution is characterized by a switch in metabolism that can be assimilated to specialization.

In this third part of the thesis, the objective is also to investigate evolutionary trajectories of specialization induced by potential metabolic modifications. While in the precedent part the exploration of evolutionary trajectories was more theoretical and wrapped around the importance of interaction and the rise of metabolic dependency, this time we focus on experiments to seize genetic and functional insights.

Consistently with the frame of the core assumptions of evolution, especially the dogma that natural selection tends to increase fitness of organisms and the facts that evolution and metabolism are constrained by the existence of trade-

offs, we hypothesized that when switching environment, some functions expressed by an organisms, and most specifically the metabolic pathways implied in these functions, would become more important while others would be lessen to a facultative use. These changes could be reflected at the gene, metabolic and functional levels. As for fishes in caves, we also hypothesized that in a long lasting stable environment in which some functions are superfluous, a generalist organism would improve its fitness by disengaging from these superfluous functions to specialize to the actual environment (see introduction section).

This section hence focuses on the driving idea of predicting evolutionary trajectory of specialization.

In the following, we first aim at forecasting which functions will be subjected to changes and what will be the fate of these functions, before experimentally testing the predictions. Indeed, focusing our work on the specialization of the (quasi-)model organism *P. fluorescens* Pf0-1, the first chapter of this section presents a metabolic modeling approach (Flux Analysis) which purpose was to determine which functions in Pf0-1 could potentially be altered or enhanced if the bacteria was to specialize to a constrained stable and homogeneous environment. In the second and third chapters, the metabolic assumptions were experimentally tested through two distinct approaches to retrieve information at different levels: 1) in the first case, the *in silico* approach of metabolic modeling is combined to a set of *in vitro* evolutionary experiments and genomic analysis in order to verify if predictions made through metabolic inferences are actually expressed at the genetic level; 2) in the second approach, observation of functional traits of initial and specialized population were performed to test for functional modifications potentially linked to specialization and mutation effects.

Chapter 2

Metabolic Modeling: Targeting genes and pathways subjected to Evolution

Using Flux Balance Analyses to predict evolution in a specialization context

[1]Alix Mas, [2]Marko Budinich, [2]Damien Eveillard, [1]Yvan Lagadeuc, [1]Philippe Vandenkoornhuyse.

1 Université de Rennes 1, ECOBIO UMR 6553

2 Université de Nantes, LS2N

2.1 Rationale of the Metabolic Modelling approach

The metabolism of an organism is here considered as the interface between the environment and the functioning of this organism. Depending on the presence or absence of nutrients in the environment (and on other environmental conditions), some functional pathways will be activated or shut off, thereby modifying the activity of the organisms (such as its nutrient consumption, its motility, its growth). For example, when iron is scarce or in inaccessible forms in the environment of *Pseudomonas fluorescens*, the metabolism producing siderophores such as pyoverdine (molecule able to chelate iron particles and enhance its retrieval) will be activated (which also make the bacteria fluoresces) to release pyoverdine in the environment (Albesa et al., 1985).

As presented in "Metabolic Modeling in Brief" below, it is now possible to infer from a genome's sequence of an organism the metabolic functions this organism is potentially able to perform or not. Further modeling approaches permit to implement conditional and stoechiometric information in the models such as the presence-absence of particular resources in the environment.

Therefore in the context of specialization leading to a niche reduction for the evolved population (they specialized toward the exploitation of particular environmental resources, to the detriment of others), it is possible to infer which metabolic pathways are essentials and which ones are superfluous. We hypothesize that as generations of organisms go along certain metabolic pathways will be completely modified until they are disabled (for superfluous functions), and others will be clearly enhanced (for essential functions). On the basis of the current knowledge gathered on metabolism it is subsequently possible to determine which genes are concerned by these reactions, and thus hypotheses which could be conserved or could be discarded. Based on those predictions, genes can be classified by their relevance for fitness functions, i.e genes not relevant to fitness function are expected to decay in contrast to those who are critical for fitness. For example, in the context where an organism is able to metabolize many carbon sources, but is subjected to a single of these carbon sources, we could expect the organism to specialize for the consumption of this resource, with the superfluous pathways implied in other carbon sources use progressively decaying (Figure 2.1).

To put it in simple words, the modeling approaches we encompass permits to describe, for given conditions, which metabolic pathways will be activated and to which extend. To test this in a specialization context, metabolic models under

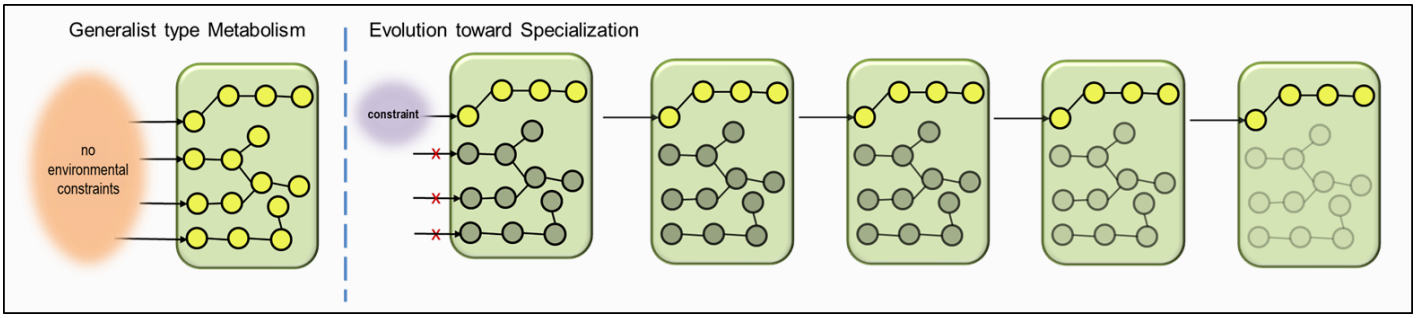


Figure 2.1: **Simplified representation of the metabolic optimization hypothesis.** This schema represents in a simplified way the "metabolic streamlining" hypothesis (Giovannoni, 2005; Tripp et al., 2010). Green rectangle is the organism, circles and lines represent the metabolic network of the organism, arrows represent the uptake of nutrients. Yellow circles are 'activated' metabolism, green circles are inactivated. In an environment without constraint (left) organisms can exploit every nutrients present (if they have the capacity to) and the corresponding metabolic pathways will thus be activated. On the contrary, in an environment where the nutrient resources are constrained, say, to one carbon source, only the metabolic pathway activated will be essential. The other unused pathway could, after some evolutionary time and according to streamlining hypotheses, decay, leading to the specialization of the metabolism of the organisms.

various constraints conditions were compared, and functions and genes implied or discarded were characterized.

2.2 Metabolic Modeling in brief

The combination of the latest advances in molecular techniques such as the elucidation of DNA sequences and genomes (Kim et al., 2012), and more generally the 'omic' revolution" (as presented in the compilation of articles gathered under the name "Genome editing, the 'omic' revolution and genetic technologies" of Trends in Biotechnologies and Trends in Genomics), associated to the development of bioinformatics (data analysis, storage, and representation, Isea 2015; Johnson and Hertig 2014) made it possible to infer from genome sequences of organisms, the proteins potentially produced when they function. Enough genetic data was accumulated and cross validated by the expertise gained from experiments or through previous chemical and physiological knowledge (SEED database on genomes annotation, Overbeek 2005) to be able to define genes and to associate a nucleotidic sequence (DNA) to an amino-acid sequence (proteins), and to characterize the reactions these proteins are involved in (NCBI database) as depicted in Figure 2.2. Proteins of particular interests are *enzymes*, as they

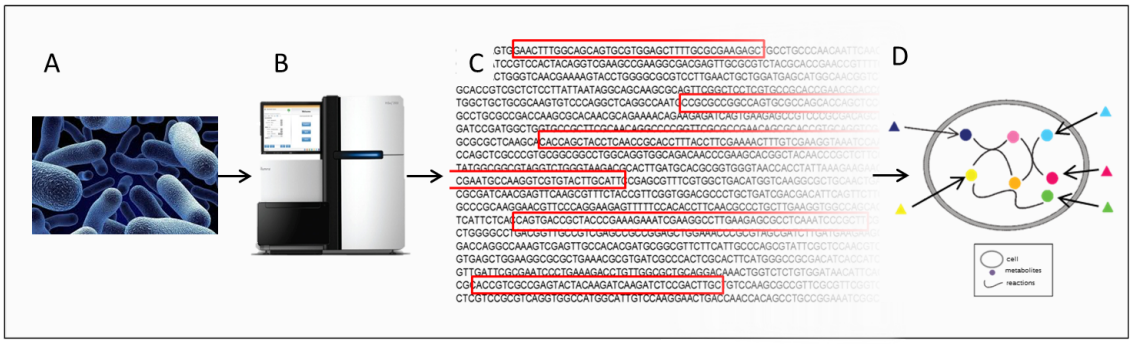


Figure 2.2: **Schematic representation of how a metabolic model is produced.** **A**: Cells, or an entire organism is sampled **B**: DNA of the cell or organism is sequenced to retrieve nucleotidic sequences of entire genome. **C**: Automatic annotation (enhanced with experts cross validations) will detect and characterize DNA sequences (red rectangles) corresponding to genes encoding enzymes. **D**: A network of detected enzymes is created, producing a complete metabolic network, proxy of the functioning of the cell or organism.

are essential to catalyze almost any chemical reactions. Therefore within an organism, enzymes will be key elements for either anabolic or catabolic reactions, *i.e* the whole metabolism at play. It is thus possible to use enzymes potentially produced as a proxy of reactions taking place within an organism. As products of some reactions are used as substrates of others, the interactions between reactions will define the metabolic network of the organism (Schomburg et al., 2013). The metabolic network is thus based on the reactions potentially existing in an organism, which are determined by the presence/absence of enzymes essential to the reactions. The existence of these enzymes are themselves determined by the recognition of the genes encoding them.
In short, the ensemble of genes encoding for enzymes present in a cell will be determinant of the metabolic pathways that exist in this cell.

Once such a metabolic network is defined it is possible to infer, through complementary modeling approaches, the metabolism (**fluxes** of metabolites) of cells (Varma and Palsson, 1994). As metabolism constitutes the first and most direct layer interacting with the media it is a particularly informative system to focus on in Ecology. The set of biochemical reactions encoded by a whole genome is usually called a Genome Scale Model (GEM) (Radrich et al., 2010). GEMs can be used to develop realistic *metabolic models* of a given organism where the whole set of biosynthesis processes of an organism will be abstracted into a single pseudo-reaction, called 'biomass function' which represents the growth of the organism. Classically in metabolic models, growth rate is considered as the main fitness measure. The interpretation of specific growth rate as a fitness

measure enables connections between models and other concepts in ecology.

The model provided is thus a set of biochemical reactions linked together within a network, and represents the metabolic capacities of a given microorganism (Thiele and Palsson, 2010). The biomass function is a linear combination of the different flux in the network. The maximization of such a function constitute the kernel of Flux (Balance and Variability) Analysis hypothesis (Orth et al., 2010).

Metabolic Modeling

A **metabolic model** can be considered as the reconstruction of the molecular physiology of an organism based on the analysis of its genome. It is systematically built, taking into account enzymes present in genomes annotation, any available "omic" datasets and legacy knowledge.

From the **SEED database** (Overbeek et al., 2005), three types of models can be obtained:
a network of metabolic reactions, a model of the associations between genes, proteins and reactions, and a whole microbial metabolism model from the biomass composition reaction.

Flux Balance Analysis (FBA) and **Flux Variability Analysis** (FVA) analysis are mathematical techniques used to simulate metabolism, through genome scale reconstructions of metabolic network, especially for reactions that are essential for the production of biomass. Each reaction in the network can be categorized as either "essential" if, when the reaction is removed, the flux through the biomass function is significantly reduced, or as "non-essential" if the flux through the biomass function is only marginally reduced or unaffected when the reaction is removed. Such simulations can calculate metabolic fluxes at steady state for models over thousands reactions.

Box 2.1: short definitions for metabolic modeling.

As explained above, it is possible to retrieve, from the genome sequences of an organism the span of its metabolic capacities, i.e, the constraints to which the organism will be able to respond to, in order to entail its growth and reproduction. This knowledge of the set of functional capacities of an organism can be assimilated to its fundamental niche (i.e the set of potential resources an organism is capable of using). It is also possible to submit this type of models to environmental constraints (qualitative and quantitative) and determine the new metabolic capacities of the organism when subjected to these constraints, which corresponds to the reduces niche of a specialized organism. The work presented hereafter aims at addressing the issue of the predictability of evolution in the particular context of specialization by a two-sided metabolic modelling, with the

first stage being the determination of optimal growth condition and media and the second stage being the production of hypotheses regarding which functions and genes may be altered during evolutionary trajectories of specialization.

2.3 Material and methods

2.3.1 *Pseudomonas fluorescens* Pf0-1 Metabolic Model

The *Pseudomonas fluorescens* Pf0-1 metabolic model was obtained from SEED Database (http://theseed.org/). It consists in 1645 reactions and 1642 metabolites, separated in 2 compartments (external and cytosol, respectively). One hundred and twenty seven (127) of the reactions are exchange reactions, *i.e.* they perform the intake/export of metabolites from/to environment.

2.3.2 Using Flux Balance Analysis to define an *in silico* medium

As defined by Orth et al. (2010), Flux Balance Analysis simulates the functioning of a cell. This approach which modelize metabolism is based on constrained stoechiometric relations between chemical species. Flux Balance Analysis gives access to a structural analysis of the metabolic network, in the sense that it gives information on both the nature of the interactions happening between the different chemical species, and the quantitative aspects of these interactions.

Usually the optimal solution to the flux-balance problem is rarely unique and present many possible, and equally optimal solutions (Gudmundsson and Thiele, 2010). On the other hand, Flux Variance Analysis returns the boundaries for the fluxes through each reaction that can, paired with the right combination of other fluxes, produce the optimal solution (Gudmundsson and Thiele, 2010).

First, analysis were run using COBRA Toolbox under MATLAB environment (Schellenberger et al., 2011) to define essential metabolites necessary for the *in silico* medium. Then, series of Flux Balance Analysis (FBA) were used to check model responses to different carbon sources (sucrose, lactate, arabinose, D-fructose, D-glucose, D-xylose, succinate or fumarate), in order to determine which culture conditions could sustain growth of *P. fluorescens* Pf0-1.

2.3.3 Using Flux Variability Analysis to establish flux span of common reactions in various models

In the flux variability analysis (FVA), environmental parameters are taken into account as quantitative metabolites being present in the vicinity of the organisms. These metabolites can be taken up by bacteria either passively or through active transportation (membrane transporters).
The information obtained from FVA is the activity expressed by each reactions for the models tested. As reactions depend on enzymes encoded by genes, a proxy of genes *activity* was made based on the activity of the reactions they encode for. Subsequently it was determined which genes could be affected when constrained to particular environments.

Determining genes statuses

For each reaction of interest, it is possible to go back up the stream of information from different databases to determine which genes, functions or pathways are implied by the reaction. Briefly the ID number of the reaction from the metabolic model can be associated with their contributing EC numbers (enzymes identifier) in the SEED database, which will enable the determination of which CDS (coding gene) are implied in the reaction (NCBI database), then from proGenome database and NCBI feature tables, we can find the gene functions and pathway associated to these genes.

As it is common that a single gene is implied in various reactions with different statuses, rules to attribute a unique status to a gene had to be established:

- If a gene is present in an essential reaction, then the gene is essential.
- Else, if the gene is not essential and is in an alternative reaction, then the gene is alternative.
- Else, if the gene is not alternative and is in an excluded reaction, then the gene is excluded.
- Else, if the gene is not excluded and is in an blocked reaction, then the gene is blocked.

Defining the models

To address the question of specialization, three FVA were performed on *P.fluorescens* Pf0-1 model with different carbon-source constraints (see Figure 2.3):

- a first FVA was done, for which the environment was considered as being non-limiting: i.e. the model includes the panel of carbon sources the organism is able to uptake. It was called the null model M0. This null model evokes an organism able to consume diverse carbon sources and can be assumed to express a generalist behavior.
- in the second FVA, the environment is constrained: the unique carbon source in the environment is glucose. Inherently, the metabolic pathways activated in this model are the ones specific to glucose utilization (in addition to the general pathways used for the cell functioning), this metabolism can thus be perceived as the metabolism of an organism specialized to glucose. It is referred to as the Glucose model, or M-Glu.
- finally the third FVA was done with glycerol as a unique source of carbon in the environment. It is referred to as the Glycerol model or M-gly.
  The three Flux Variability Analysis were performed with constraints matching the *in vitro* evolutionary experiments as closely as possible.

Glucose was chosen as a unique carbon source following the previous FBA determination of growth yield. Glycerol was chosen as an alternative constraint to test if the addition of lipases in the medium (see chapter 3) could constitute an additional factor of specialization.

Activity Statuses for reactions and genes

From FVA outputs, depending on the intensity of their activity, each reaction of the models is classified into three statuses:

- "Excluded" when the flux carried by the reaction has to be equal to 0 in the maximal biomass scenario; *i.e.* to increase biomass production the reaction *has* to be shut off (2.4).
- " Essential" when the flux carried by the reaction has to always be effective (i.e, the flux will never equal 0 in maximal biomass scenario, 2.4)
- "Alternative" when the reaction can either carry a flux or be replaced by another reaction (an alternative path enabling the production of a same metabolite exist) in the maximal biomass scenario (2.4).

Once the reactions are characterized into statuses, the identity of the genes implied in these reactions are retrieved, as explained in part 2.3.3 : Defining genes status.

Once the statuses of reactions and genes are retrieved, further analysis are performed.

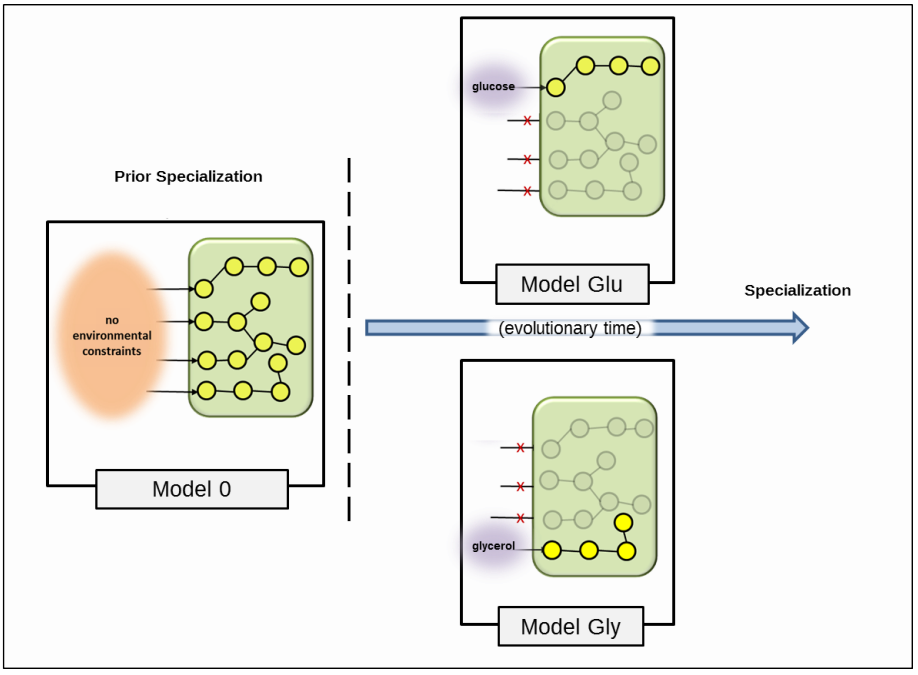


Figure 2.3: **Figuration of the Metabolic Models.** Linked to our metabolic hypothesis, this figure represents the three models used in this section : Green rectangle is the organism, circles and lines represent the metabolic network of the organism, arrows the uptake of nutrients. Yellow circles are 'activated' metabolisms, green circles are not. The first model M0 correspond to the metabolic model of a biomass function without environmental constraints, the two other models (M-Glu and M-Gly) correspond to the same function but with glucose or glycerol as unique carbon source, which we assimilate to metabolic specialization.

First, the statuses of genes themselves are considered in each model *separately*, with the underlying idea that genes implied in reactions with a blocked or excluded status should undergo modification reducing their expression, while the expression of essential genes should be enhanced. Indeed, under the flux analyses modelling assumption, the biomass functions are constrained to be close to their theoretical maxima, simulating an organism maximizing its fitness. Under that assumption, it was considered that reactions not carrying flux (excluded) had to be **not** used to increase fitness, and therefore we hypothesized that these functions were candidates to be lost, and that the genes encoding these functions could be targets for mutations. More precisely, our hypotheses are that 1) the expression of genes encoding enzymes implied in excluded reactions could be neutralized. If there is a metabolic advantage for the gene to be shut off (i.e. *selection for* a reduced expression of the gene), stop-codons mutations or mutations down-regulating the gene expression can emerge; if there is a *lift of selection* on the gene, then it can result in the accumulation of (circum-)neutral mutations on the gene. On the other hand, 2) we can expect the expression of genes implied in essential reactions to be enhanced, or at least not modified. Such genes are expected to not carry mutations (and especially not stop-codon mutations) or to carry mutations that will improve their expression. Genes associated to alternative reactions are expected to have no particular patterns of mutations, as they can either be used or not.

In a second analysis the focus is set on the *comparison* of the statuses attributed between the different models. This comparison permits to make predictions into which genes could potentially be more affected by the environmental constraints imposed. For example if a gene is found to switch from an "essential" status in the null model to an "excluded" status in the glucose model we can expect such a gene to carry mutations hindering its activity.

CV2s of reactions and genes

From the FVA performed it is also possible to extract information on the flux span of each reactions, i.e. the extend to which the flux "going through" a reaction can vary. It can be understood as a quantification of the activity of reactions. The information obtained is the variability expressed by each flux of the reactions in the models. Reactions which can only support a low variability of fluxes through them are constrained to express steady fluxes and likely to be of a higher importance to the organism.

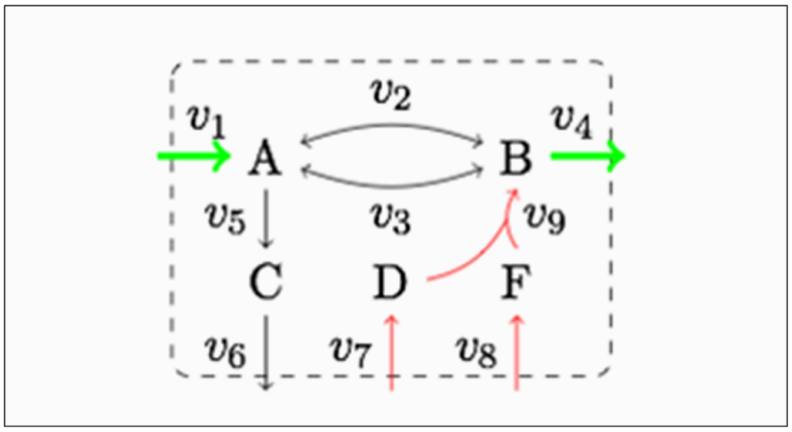


Figure 2.4: **Representation of the different types of reactions in FVA.** In this simplified figure the different kind of reactions (i.e. blocked, essential, alternative and excluded) are shown: Let's consider v4 as the biomass production reaction, we assume the production of metabolite B as a fitness proxy. If A is the only substrate in the environment, then reactions v1 and v4 are essential reactions (i.e. the reactions *have* to carry a flux to optimize biomass production), while v2 and v3 are alternative reactions (i.e. any of the two reactions can carry a flux indifferently, to optimize biomass production). v5 and v6 are blocked reactions (the biomass production would be lowered if v5 and v6 occurred). v7-v9 are excluded reaction as the necessary metabolites are not present in the environment; we expect that genes coding for v7, v8 and v9 to disappear upon evolution).

From this flux span, a metric called CV2 is calculated to characterize each reactions (Box 2.2).

CV2 Metrics

The second-order coefficient of variation (CV2) is used to characterize fluxes of metabolic reactions from a Flux Variablity Analysis. It has been chosen as an alternative to the more classical Pearson's coefficient of variation V, which was considered limited for such data analysis (Kvalseth et al., 2016) as it didn't take into account the extent of the flux variation (for example a ratio of 1/2 is similar to 50/100, yet the fluxes carried by the reactions are quite different, and the difference is informative).

formuleCV2

$$V_2 = \left(\frac{m_2}{m'_2}\right)^{1/2} = \frac{s_n}{\sqrt{m'_2}} = \left(\frac{s_n^2}{s_n^2 + \bar{x}^2}\right)^{1/2} = \left(1 - \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2/n}\right)^{1/2}, \upsilon_2 = \left(\frac{\mu_2}{\mu'_2}\right)^{1/2} = \frac{\sigma}{\sqrt{\mu'_2}} \quad (5)$$

Each CV2 of each reaction is a number between 0 (no variation) and 1 (high variation). This variation is in respect with the flux itself, and not compared to other fluxes. When a fluxes has a large span (the flux can be low or high) for a given model it is assumed that the reaction is less constrained, and thus of lesser importance than reactions of which the flux span is very constrained and thus limited (low flux variability).

Box 2.2 : CV2 metric.

As for aforementioned activity-statuses, the CV2s of the reactions were investigated at two scales:

First the CV2 of reactions were interpreted for themselves, with the assumption that genes entailing reactions with a low CV2 are supposed to be less prone to mutate (more conserved) than genes entailing reactions with a high CV2 whose expression may be more "flexible". In that sense, the CV2 metric can be used to make predictions regarding modifications of genes during specialization.

Secondly, a comparison of the CV2s from the null model to the constrained (glucose or glycerol) models was realized. This gives quantitative information on the propensity of a reaction to be modified when it is under constraints. Our hypothesis here is that reactions (and genes implied in these reactions) which fluxes differ the most between the null model and the glucose or glycerol model should display more changes. For example, if we observe a CV2 equal to 0.1 for a given reaction in the null model, and that this same reaction has a CV2 equal to 0.8 in the Glucose model, we could expect this reaction to be less conserved than reactions with a high CV2, as they offer a higher variability and are thus understood to be of a lesser importance for the metabolism.

To resume, from the flux variability analyses performed on unconstrained and constrained models, two types of information are accessible:

- 1) the constraints applied on genes, defined by their own activity statuses and CV2s, for each different model. This permits to make hypothesis as to which genes can mutate or not, and
- 2) the difference in constraints from a 'generalist' model (M0) to 'specialist' models (M-Glu and M-Gly) through the *comparison* of the statuses and CV2s of genes and reactions between models. This permits to narrow the predictions of genes that will be the most modified along specialization.

2.4 Results and Discussion

2.4.1 *P.fluorescens* Pf0-1 Metabolic model

As stated in the material, the metabolic model of *P. fluorescens* Pf0-1 is made of 1645 reactions. The modelling of the metabolism of *P. fluorescens* Pf0-1 allowed to predict fundamental reactions (categorized as essential), and on the other side of the spectrum, reactions that are offset within a given environment -such as growth medium (reaction categorized as blocked or excluded).

A thousand and sixty-one (1061) genes are implied in the model to sustain every reactions. These genes are identified as 'metabolic genes' hereafter. By extension genes encoding the core enzymes of essential reactions are called "essential genes", and respectively for genes implied in alternative and excluded reactions. The number of metabolic pathways determined for this model is 137.

The type of information obtained can be visualized in an integrative network representing all metabolic pathway potentially in action, as shown in Figure 2.5, extracted from KEGG Metabolic Pathway (http://www.genome.jp/kegg/kegg2.html).

2.4.2 FBA to define the medium

The Flux Balance Analysis of the model can be considered as a feasibility-check of the model from which one can infer metabolites necessary to the medium. In a first stage, 11 metabolites were identified as essential, they correspond mainly to trace nutrients: Mg2+, Cl– , O2, Cu2+, Co2+, SO4– , Ca2+, K, Zn2+, Mn2+ and spermidine.
From the carbon sources tested, the model was unable to produce any biomass using sucrose, lactate and arabinose, which were in consequence discarded for

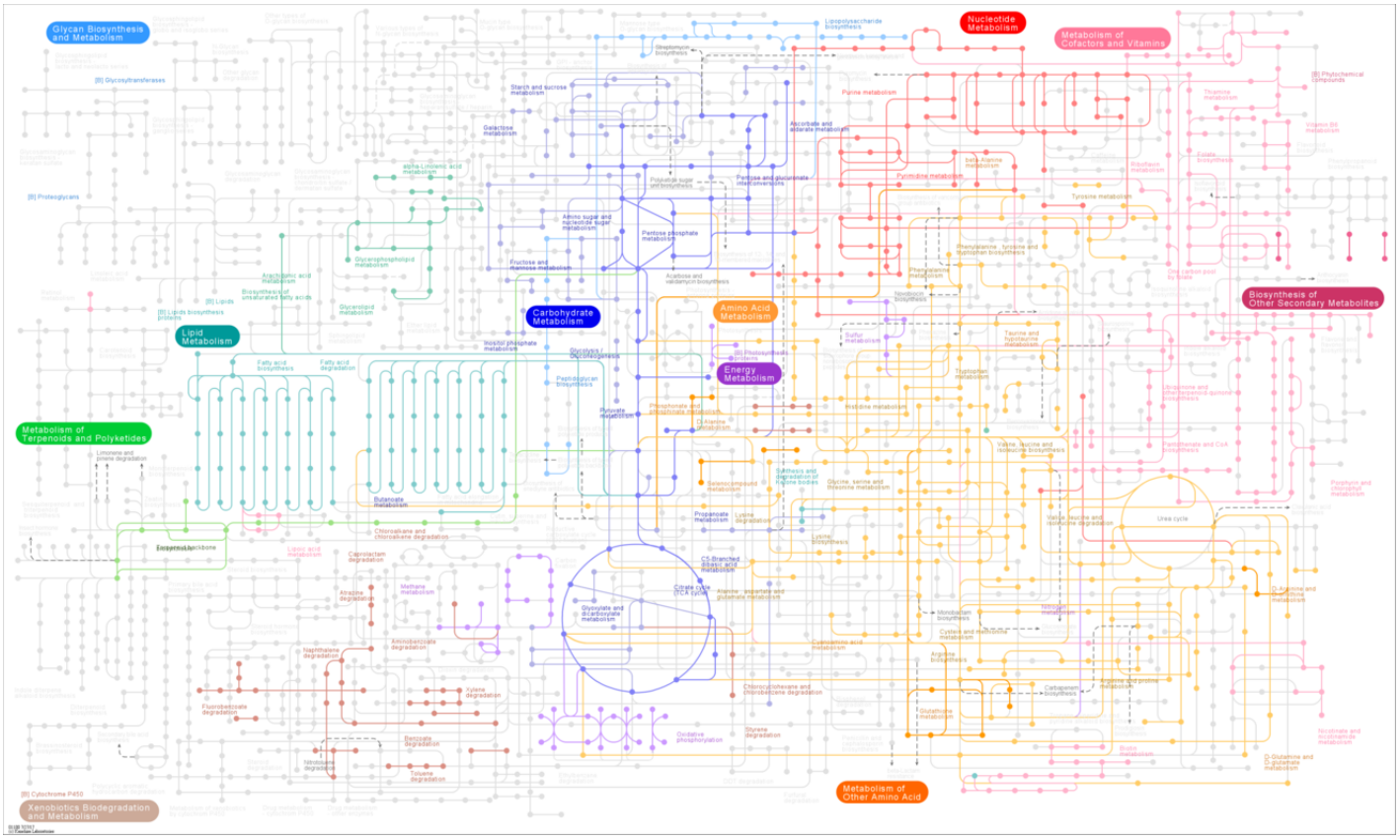


Figure 2.5: **Metabolic network of *P. fluorescens*.** Each color represent particular metabolisms sustained by *Pseudomonas fluorescens* Pf0-1 (*e.g.* Carbohydrate, Energy or Lipid Metabolism) which are constituted of several metabolic pathways (*e.g.* for Carbohydrates metabolism: Glycolysis, Citrate or Pentose pathways). Each dot is an enzyme. It is remarkable that metabolisms are all interrelated.

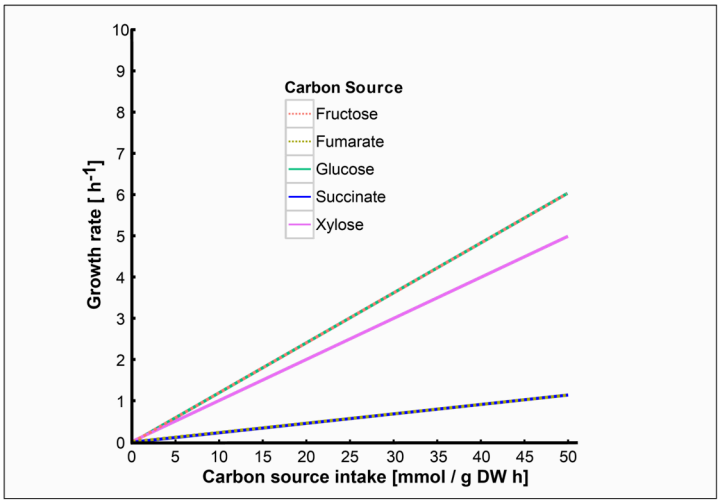


Figure 2.6: **Growth rates of *P. fluorescens* in different carbon sources.** This figure represents the growth rate (equivalent to the biomass function, cornerstone of Flux analysis) of *Pseudomonas fluorescens* Pf0-1 in various carbon sources. In some carbon sources (glucose, fructose) the bacteria will thrive while in others the growth will be slower (succinate, fumarate). When growth behavior in a given substrate are similar the curves are superposed -dashed lines.

further simulations. In fructose and glucose the growth behavior is similar (superposed curves), and show better culture yields than for succinate and fumarate (also superposed) (Figure 2.6). For this reason, Flux Variability Analysis was subsequently performed with glucose as a unique carbon source to test specialization, as presented in part 2.3.3. The choice of glycerol as an alternative specialization model was (not dependent on FBA but) dependent on the rationale according to which the addition of lipases enzymes could boost the specialization trajectory followed.

2.4.3 FVA to study metabolism under several contexts

To characterize the dynamic of reaction fluxes in Glucose and Glycerol models, Flux Variability Analysis were performed. Most of the time, the outcome of fluxes analyses for the glucose and glycerol model are analogous. For this reason and to avoid redundancy, the decision was made to extensively present the results for the glucose analyses at first, and then to complete with the glycerol analysis when needed.

Status \ count	Reactions	Genes	Number of Pathways implied
Essential	352	396	51
Alternative	449	407	45
Excluded	119	60	7
Blocked	725	198	47
Metabolic counts	1645	1061	137

Table 2.1: **Description of reactions and gene counts for FVA with Glucose constraint.** The statuses (Essential, Alternative, Excluded and Blocked) correspond to statuses attributes by FVA analysis. The line "Metabolic count" refers to the total existing number of reactions, metabolic genes and pathways known for *Pseudomonas fluorescens* pf0-1.

Blocked reactions in Glucose and Glycerol

As summarized in Tables 2.1 and 2.4, out of the 1645 reactions defined in the model, 725 are considered as blocked at all time, and these blocked reactions are common to FVA outputs in both glucose and glycerol. This result signifies that to maximize biomass production, 44 % of the reactions the bacteria can usually perform should be shut off. It implies that whatever the resources present in the environment, it is best for the bacteria to not perform these reactions if the main purpose of its functioning is to produce maximal biomass (i.e. fitness proxy). This information can seem peculiar at first, but can be explained by the fact that we assume a steady state hypothesis for the model: metabolites are not allowed to accumulate, neither in the external nor in the cytosolic compartment. Exchange reactions which usually take in and out metabolites from external compartment from/to the environment sustain equilibrium. If a reaction produces a metabolite in the cytosol which is not subsequently consumed for biomass production or taken out by any exchange reaction, the metabolite will accumulates in the cytosol. This type of reactions is considered as sub-optimal for the model and is therefore characterized as 'blocked'.

The blocked reactions are encoded by a set of 452 genes. Nevertheless it appears that 254 (56 %) of these genes are also involved in other non-blocked reactions. According to the rules defined in section 2.3.3 these genes will not have the blocked status, and a subtotal of 198 genes only are attributed the blocked status (Table 2.1, 2.4 & Figure 2.7). Fifty-eight (58) of these blocked genes could not be linked to metabolic pathways, while 54 of them were involved in multiple pathways. The other 86 genes entail 35 metabolic pathways (out of 137).

Further analyses are focused on essential, alternative and excluded reactions.

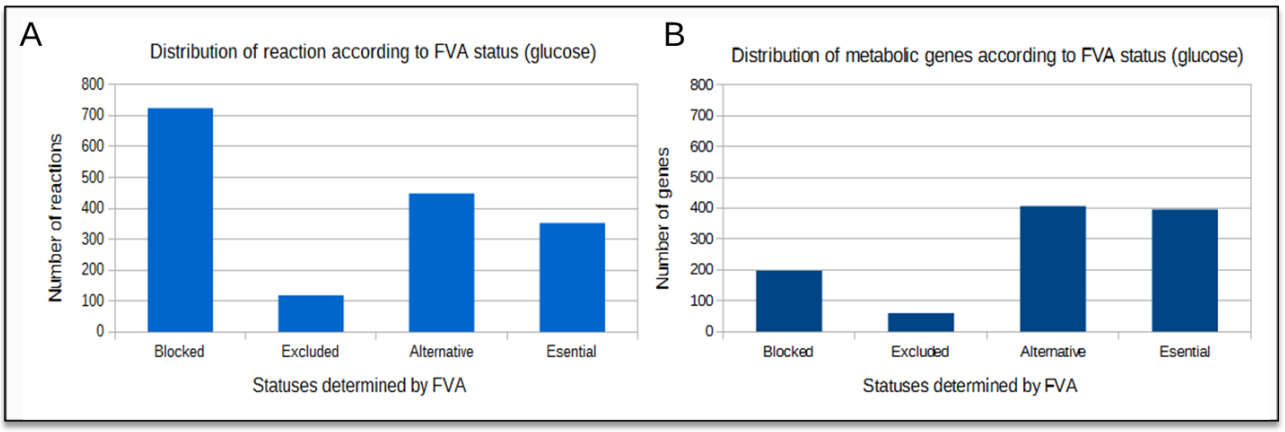


Figure 2.7: **Count of reactions and genes associated to these reactions in each FVA category.**
(A) distribution of the number of reactions per FVA category. **(B)** distribution of metabolic genes associated to the reactions presented in A. The scale was adjusted to enable direct comparison and we can see that there is a reduced number of blocked genes as the definition of statuses for genes is not straightforward (genes may be implied in several reactions defined by different FVA statuses).

Description of Statuses in the Glucose model

The 920 non-blocked reactions of the metabolic models run with glucose as the sole carbon source are described here.
Three hundred and fifty two (352) of these reactions are considered as essential, 449 are alternatives and 119 reactions are excluded (Table 2.1 & Figure 2.7.**A**). Three hundred and ninety-six (396) genes encode for the enzymes implied in reactions considered as essential for growth maximization, while 407 are considered as being alternatives. Sixty (60) of the genes are excluded (Figure 2.7.**B**), implying that in the specific case of Glucose as only carbon source, these reactions should be "shut off" to maximize growth rate (biomass production). Count of reactions for each status is given in Table 2.1.

Based on our hypothesis, it is interesting to focus on both essential and excluded reactions, as we expect particular mutation patterns on the genes involved in these reactions.

Excluded reactions (119) and associated genes (60) are a minority, and are implied in less than a dozen of metabolic pathways. As summarized in Table 2.2, out of the 60 excluded genes, 3 are explicitly included in galactose metabolism, 2 in alanine, aspartate and glutamate metabolism, 3 in taurine and hypotaurine metabolism, 2 in glycerophospholipid, 2 in nitrogen metabolism and 2 in pentose

metabolism. Interestingly, 32 out of the 60 genes are implied in trans-membrane transport of molecules (amongst which 23 are ABC transporters) and some of these transporters are also implied in metabolisms cited above such as taurine, nitrogen and amino-acids metabolisms. This means that several exchange reactions from the environment to the cytosol are precluded, most probably preventing the activation of the suite of the pathways.

From the Table 2.2, it is also noticeable that a few genes implied in xylose, succinate and galactose metabolism are considered as excluded, which is in line with our supposition that genes involved in alternative carbon source metabolism should be offset. Moreover, when studying genes associated to single pathways only, it appears that the galactose metabolism, the taurine metabolism and the pentose and glucuronate interconversions metabolism imply genes with the excluded status only (Table 2.3).

On the other hand, genes implied in reaction tagged as essential are more numerous (396), and are implied in 89 metabolic pathways, amongst which we find again all the pathways (except for taurine) also entailed by excluded genes. This means that reactions and genes implied in pathways can be of different statuses, and some reactions of a pathway may be essential while others will be excluded. Similarly, when studying the 407 alternative genes it appears that they are mostly implied in pathways also entailed by genes of other statuses. Nevertheless for some pathways, predictions are homogeneous as every genes retrieved have a similar status (Table 2.3). These pathways with unique tendencies (essential, excluded) potentially traduces stronger constraints on their functioning.

In a singular pathway, it is possible to encounter reactions that are excluded, essential or alternative. This traduces that pathways are seemingly not to be entirely conserved, or entirely excluded. Most probably modifications of particular reactions within the pathways only are sufficient to entail the changes necessary for growth optimization. This limit reached in the pathway scale lays in the fact that the notion of metabolic pathways is, to some extent, arbitrary, and the boundaries of a pathway are subjective. Indeed, not only enzymes can be implied in diverse set of reactions, and thus in several metabolic pathways, but also metabolic pathways themselves frequently overlap, implying common metabolites and enzymes. For instance, gene encoding for Triose-phosphate isomerase can be mobilized in glycolysis / gluconeogenesis metabolism, in fructose and mannose metabolism, in inositol phosphate metabolism and in carbon fixation (Patric pathway database ; Figure 2.8). Here, we are not confronted to genes

Galactose
2-dehydro-3-deoxy-6-phosphogalactonate_aldolase
2-dehydro-3-deoxygalactonokinase
galactonate_dehydratase
Pentose
xylulokinase
membrane-bound_PQQ-dependent_dehydrogenase
Alanine, aspartate and glutamate metabolism
aspartate_aminotransferase_family_protein
succinate_dehydrogenase
Nitrogen
nitrite_reductase_(NAD(P)H)_small_subunit
nitrite_reductase_large_subunit
nitrate_ABC_transporter_substrate-binding_protein
nitrate_ABC_transporter_substrate-binding_protein
Taurine
taurine_dioxygenase
taurine_dioxygenase
dioxygenase
taurine_import_ATP-binding_protein_TauB
taurine_ABC_transporter_permease
taurine_ABC_transporter_substrate-binding_protein
Glycerophospholipid
ethanolamine_ammonia_lyase_large_subunit
ethanolamine_ammonia-lyase_light_chain
ethanolamine_permease
ethanolamine_permease

Xylose
D-xylose_ABC_transporter_substrate-binding_protein
xylose_ABC_transporter_permease
Malonate
malonate_decarboxylase_ACP
malonate_decarboxylase_subunit_alpha
malonate_transporter_subunit_MadL
biotin-independent_malonate_decarboxylase_subunit_beta
biotin-independent_malonate_decarboxylase_subunit_gamma
Amino Acids Transporters
amino_acid_ABC_transporter_ATP-binding_protein
arginine-ornithine_antiporter
arginine-ornithine_antiporter
histidine/lysine/arginine/ornithine_ABC_transporter_permease_HisM
histidine/lysine/arginine/ornithine_ABC_transporter_permease_HisQ
methionine_ABC_transporter_substrate-binding_protein
methionine_ABC_transporter_substrate-binding_protein
methionine_ABC_transporter_substrate-binding_protein
Uncharacterized transporters
ABC_transporter_permease
ABC_transporter_permease
ABC_transporter_substrate-binding_protein
ABC_transporter_substrate-binding_protein
ABC_transporter_substrate-binding_protein
metal_ABC_transporter_permease
metal_ABC_transporter_permease
metal_ABC_transporter_permease
metal_ABC_transporter_substrate-binding_protein
MFS_transporter
MFS_transporter
sulfonate_ABC_transporter_ATP-binding_lipoprotein
sulfonate_ABC_transporter_ATP-binding_protein
sulfonate_ABC_transporter_substrate-binding_protein

Table 2.2: **Excluded genes distributed according to their metabolic pathways.** Excluded genes are included is several pathways such as the galactose, the nitrogen, the taurine or the pentose metabolisms, and almost 40% of the excluded genes are transporters.

Pathways unique to Alternatives	Pathways unique to Essentials	Pathways unique to Excluded
Ascorbate and aldarate metabolism	Aminoacyl-tRNA biosynthesis	Galactose metabolism
Betalain biosynthesis	Benzoate degradation via hydroxylation	Pentose and glucuronate interconversions
Biotin metabolism	Biosynthesis of siderophore group nonribosomal peptides	Taurine and hypotaurine metabolism
Carbon fixation in photosynthetic organisms	D-Alanine metabolism	
Fructose and mannose metabolism	D-Glutamine and D-glutamate metabolism	
Glutathione metabolism	Fatty acid metabolism	
Glycerolipid metabolism	Lipopolysaccharide biosynthesis	
Glyoxylate and dicarboxylate metabolism	Peptidoglycan biosynthesis	
Methane metabolism	Photosynthesis	
One carbon pool by folate	Riboflavin metabolism	
Oxidative phosphorylation	Terpenoid backbone biosynthesis	
Propanoate metabolism	Valine, leucine and isoleucine biosynthesis	
Thiamine metabolism		

Table 2.3: **Distribution of pathways entailed by genes corresponding to unique pathways.** This table summarize the pathways retrieved as associated to one status of metabolic reactions only. Pathways presenting genes or reactions of two or more status are not represented here.

having different effects according to the metabolic reaction they entail, but we are facing metabolic pathway implying various genes, some of which will be essentials, while others will be excluded.

This constitutes a limit to our prediction approach: FVA enables relatively precise predictions at the reactions level, but reactions of different statuses can belong to a same pathway, making it intricate to predict modification at the pathway scale. Moreover the complexity of the metabolic system, where single genes can be implied in different pathways hinders the precise prediction at the genetic level. The relations between genes, reactions and pathways are not linear, and the inference of prediction at one level cannot be determined by the sole analysis at another level.

Description of CV2s in the Glucose model

In addition to statuses, reactions are characterized by their CV2s. Recall that reactions with a low CV2 (i.e, low flux variability) are expected to be more constrained and thus of a higher importance for the metabolism than reactions with a high CV2 (reactions with more variable fluxes).

The 920 non-blocked reactions have CV2 spanning from zero to one. Flux having a minimum and maximum flux equal to zero have no variation (*i.e.* flux span and CV2 $= 0$) and correspond to excluded reactions. Interestingly, no other reaction than the excluded one (i.e. with non zero-based flux), has a CV2 equal to zero, meaning that every reaction has at least the possibility of some variability. Reactions with the lowest CV2 all have the Essential status, but reciprocally all essential reactions status do not have a low CV2 which can span from 0,01 to

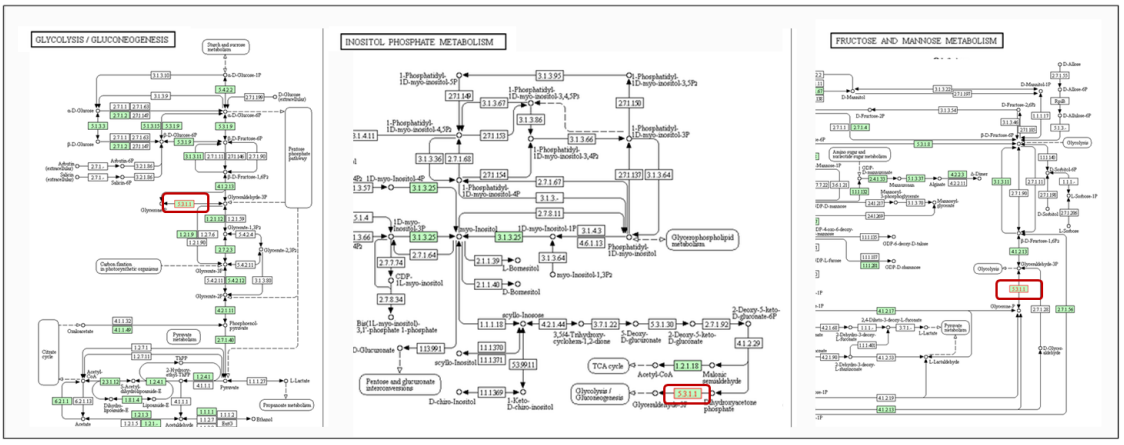


Figure 2.8: **Representation of three pathways sharing a common enzyme encoded by a single gene.** In this figure, three distinct metabolism are shown to share the 'use' of a single gene. This underlines that there are overlaps in metabolic pathway due to their interconnection

.

0,8. Reactions having the highest CV2 (>0,8) are alternative reactions. Indeed these reactions can be either shut off, or active, thus the fluxes may vary a lot. Below (Figure 2.9) is the histogram representing the distribution of the number of reactions according to their CV2.

Complements for the Glycerol model

Description of Statuses For FVA in glycerol, the trend of distribution of reactions (and respectively genes) into the four status is similar as for Glucose. For excluded statuses, the number of reactions (119) and genes (60) are exactly the same if constrained to glycerol or glucose. Slight differences lay in the essential and alternative categories. Four (4) reactions considered as essential under glucose constraints become alternative under glycerol constraints. Regarding genes involved, the pattern is the same: the count is almost exactly similar as for Glucose except that one gene switches from the essential category to the alternative category when under glycerol constraints (Table 2.4).

Nevertheless, when actually looking at the genes distributed inside the different categories, it can be noticed that 13 genes had switch from the essential to the alternative categories, while 14 had switch from the alternative to the essential categories (hence the count difference of one detected). The list of these genes and their respective function are presented in Table 2.5. To notice, there is no switch of status for genes of the essential/alternative categories to the excluded category (or inversely) from a model to another. The pathways im-

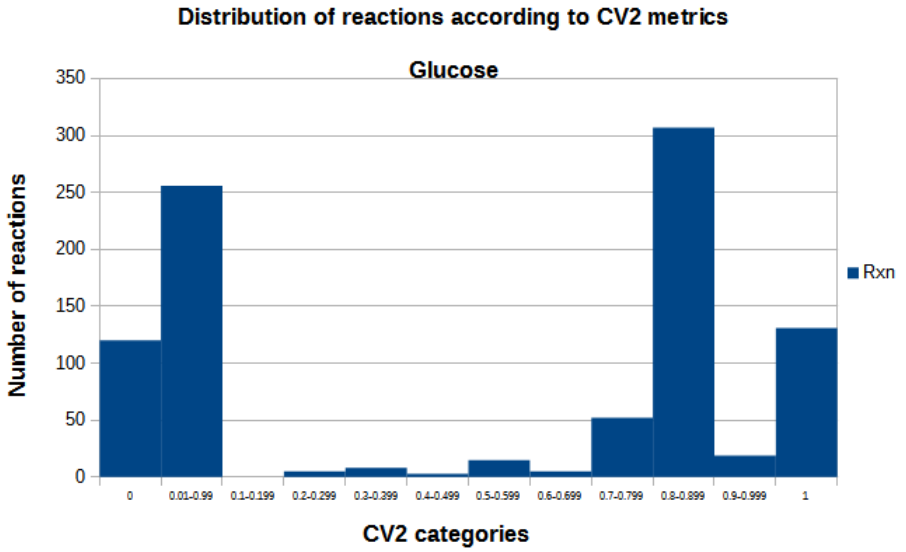


Figure 2.9: **Predicted distribution of reactions of *P. fluorescens* pf0-1 according to CV2 metrics, in glucose minimal medium.** The CV2 metrics helps considering the constraints existing on reactions. Reactions with a low CV2 (left of the histogram) are more constrained than reactions with a higher CV2 (right) because the fluxes of these reactions are not variable and should always be sustained at the same level to maximize biomass production. A total of 350 reactions have a relatively small CV2 (conserved fluxes), while more than 450 have a high CV2 (variable fluxes).

Status \ count	Reactions	Genes	Number of Pathways implied
Essential	348	397	50
Alternative	453	406	45
Excluded	119	60	7
Blocked	725	198	47
Metabolic counts	1645	1061	137

Table 2.4: **Description of reactions and gene counts for FVA with Glycerol constraint.** The statuses (essential, alternative, excluded and blocked) were attributed by FVA analysis for *P. fluorescens* Pf0-1 in glycerol minimal medium. The line "Metabolic count" refers to the total existing number of reactions metabolic genes and pathways known for *Pseudomonas fluorescens* Pf0-1.

Genes	Pathways	Functions	Status in Glucose	Status in Glycerol
PFL01_RS13040	Pentose phosphate pathway	6-phosphogluconate_dehydrogenase	essential	alternative
PFL01_RS06375	NA	anion_permease	essential	alternative
PFL01_RS21725	NA	anion_permease	essential	alternative
PFL01_RS05980	Photosynthesis	ferredoxin--NADP_reductase	essential	alternative
PFL01_RS24605	Photosynthesis	ferredoxin--NADP_reductase	essential	alternative
PFL01_RS22920	NA	FMN-binding_glutamate_synthase	essential	alternative
PFL01_RS21935	Pentose phosphate pathway	ketohydroxyglutarate_aldolase	essential	alternative
PFL01_RS22005	Pentose phosphate pathway	phosphogluconate_dehydratase	essential	alternative
PFL01_RS27095	Pentose phosphate pathway	ribose-5-phosphate_isomerase	essential	alternative
PFL01_RS25790	Pentose phosphate pathway	ribulose-phosphate_3-epimerase	essential	alternative
PFL01_RS13705	Pentose phosphate pathway	transketolase	essential	alternative
PFL01_RS26515	Pentose phosphate pathway	transketolase	essential	alternative
PFL01_RS13710	Pentose phosphate pathway	transketolase	essential	alternative
PFL01_RS28815	Oxidative phosphorylation	ATP_synthase_epsilon_chain	alternative	essential
PFL01_RS28830	Oxidative phosphorylation	ATP_synthase_subunit_alpha	alternative	essential
PFL01_RS28820	Oxidative phosphorylation	ATP_synthase_subunit_beta	alternative	essential
PFL01_RS28835	Oxidative phosphorylation	ATP_synthase_subunit_delta	alternative	essential
PFL01_RS28825	Oxidative phosphorylation	ATP_synthase_subunit_gamma	alternative	essential
PFL01_RS04430	Purine metabolism	bifunctional_sulfate_adenylyltransferase_subunit_1/adenylylsulfate_kinase	alternative	essential
PFL01_RS14445	Valine, leucine and isoleucine degradation	crotonase	alternative	essential
PFL01_RS05650	Pyrimidine metabolism	CTP_synthetase	alternative	essential
PFL01_RS06810	Fatty acid elongation in mitochondria	enoyl-CoA_hydratase	alternative	essential
PFL01_RS28855	NA	F0F1_ATP_synthase_subunit_I	alternative	essential
PFL01_RS05400	Pyruvate metabolism	phosphoenolpyruvate_carboxylase	alternative	essential
PFL01_RS05065	Glycolysis / Gluconeogenesis	phosphoenolpyruvate--protein_phosphotransferase	alternative	essential
PFL01_RS04015	NA	phosphoenolpyruvate--protein_phosphotransferase	alternative	essential
PFL01_RS04425	Purine metabolism	sulfate_adenylyltransferase_subunit_2	alternative	essential

Table 2.5: **Description of the 27 genes for which the statuses switch categories when in Glucose, or in Glycerol.** This table focuses on the subset of genes (implied in reactions) that change status depending on if they are retrieved from the Glucose or Glycerol FVA results. This means these reactions are not submitted to the same constraints whether the media provides Glucose or Glycerol.

plied by these status switch are mostly the pentose phosphate pathway and the oxidative phosphorylation pathway. Interestingly, the pentose phosphate pathway is a metabolic pathway parallel to glycolysis which involves the oxidation of glucose. It generates NADPH and 5-carbon sugars (pentoses) as well as ribose 5-phosphate, precursor for the synthesis of nucleotides (Kruger and von Schaewen, 2003). The loss of the essential status when in a media with glycerol only is thus totally appropriate and in line with our hypotheses. Also, as mentioned the oxidation of glucose releases NADPH. NADPH is the principal electron source of biosynthetic reactions in the cell. Thus a stop in the oxydation of glucose will stop the production of NADPH. We can suppose that this diminution in NADPH is compensated by the increased production of ATP through the oxydative phosphorylation pathway which switches to the essential status under glycerol constraints.

Analyses of CV2s of reactions in glycerol models Here again, the distribution of reactions within CV2 categories is very similar for Glycerol and Glucose and reflects the observations made when studying statuses (Figure 2.10).

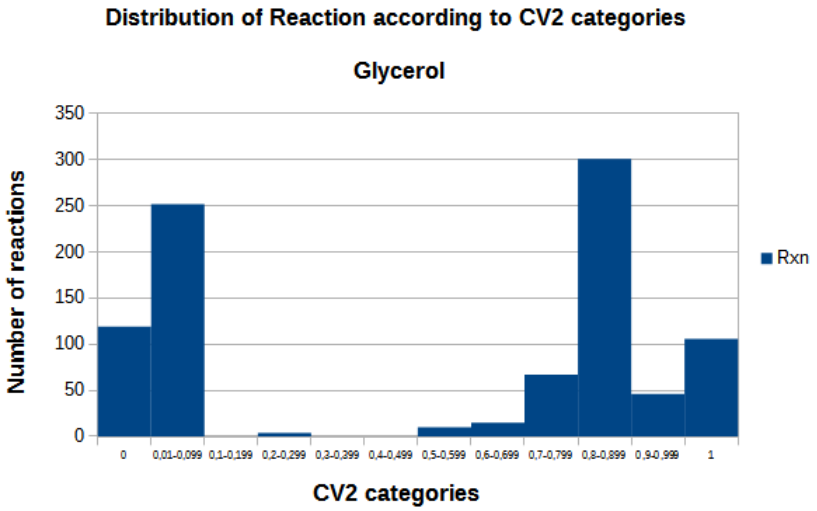


Figure 2.10: **Distribution of reactions of *P. fluorescens*. pf0-1 according to CV2 metrics, in glycerol minimal medium** Most reactions are either highly constrained (bars on the left, low CV2) or not too contraisned (right, high CV2).

2.4.4 Comparison of the three models

The three models compared are the unconstrained null model, and the constrained Glucose and Glycerol models. As mentioned above, the Glucose and Glycerol models have very close FVA outputs. Thus only the comparison of the glucose model with the null model is presented in detail herein, knowing that the conclusions also hold for the glycerol model . The particular differences observed in genes statuses and CV2s between Glucose and Glycerol are still explained.

Comparison of Statuses

If the vast majority of the metabolic genes (>800) determined for the glucose model share similar statuses with the null model, 180 genes still present different statuses than for the null model. Of these 180 genes,

- 99 genes have an alternative status in the null model and switch to essential in the constrained models
- 60 genes have an alternative status in the null model and switch to excluded in the constrained models
- and 21 genes have an essential status in the null model and switch to alternative.

In the frame of our prediction hypotheses, the most interesting pattern would have been genes switching from an essential to an excluded category, but there

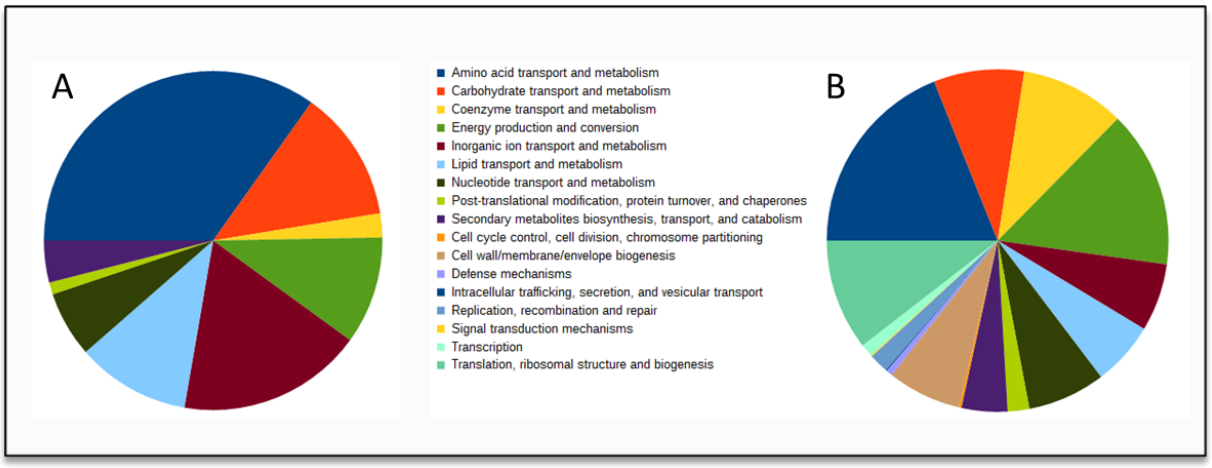


Figure 2.11: **Comparison of the distribution of genes in COG categories according to their status changes from the null model to the glucose model.** **A** Distribution of genes whose status is different in the null and in the glucose model. It is noticeable that the proportion of genes implied in amino acid, carbohydrate and inorganic ion transports & metabolism categories is more important for genes changing status. For comparison, the distribution within COG categories of genes which have the same status in the two models is presented in **B**.

is none.

The list of the genes which changed statuses will be checked for mutations in the third chapter and their distribution within COG categories is represented in Figure 2.11.

Additional difference for the Glycerol model : The difference in genes statuses between the Glycerol model and the null model is very close to what is observed between the Glucose and null models, except that instead of 180 genes that differ status, only 157 are different (and they are included in the 180 of the glucose's), thus status-wise, the glycerol model is closer to the null model.

Comparison of CV2s

The comparison of CV2s between models (i.e, between unconstrained, and Glucose or Glycerol constrained models) was also done. It permits to see if the constraint imposed by the limiting carbon resource changes the dynamic of some reactions.

The histogram of Figure 2.12 present the counts of each reactions per CV2 categories. Essentially, as stated above, it is noticeable that through modelling the flux variability of reactions are very close in the two models where carbon sources are constrained (glucose=orange and glycerol=blue), but they can differ

substantially from the unconstrained model (green).

Precisely, the comparison of CV2s of reactions in Glucose model related to null model shows that 63% of the CV2s are similar between the two models. The difference between the models lays in that 45 reactions have a higher CV2 in the Glucose than in the null model, and 292 reactions have a smaller CV2 in Glucose than in the null model. This is coherent with the fact that the Glucose model is more constrained than the null model, and thus reactions have less flexibility in their expression span (hence more genes with a lower CV2).

The most noticeable difference is for reactions with CV2=0: the null model has no such reactions, while the glucose model has around 120 reactions with a CV2=0. This reactions with CV2=0 correspond to reactions classified as Excluded. Hence this consolidates the prediction in which Excluded genes are potential targets for mutations in a specialization context.

Another quite important difference is the one observed for CV2 comprised between 0,7 and 0,8, where the model without constraints present almost 150 reactions, while the glucose model has 50 of them, which became more constrained with a lower CV2.

In order to quantify the difference between the unconstrained and constrained models, the difference between their CV2 was calculated (CV2 of constrained model minus CV2 of unconstrained model) (Table 2.6). A high CV2 difference (e.g. $>$0.5) traduces an important change in constraints.

From the CV2 changes observed between the unconstrained model and the constrained models, when considering reactions for which the fluxes changed the most, the predictions are of relatively low accuracy at the genetic level as 84% of the reactions imply several genes, themselves implied in multiple pathways. Still, interestingly the characterized genes are related to amino acid, carbohydrate transport, energy production and inorganic ion transport and metabolisms (COG-categories).

The patterns of CV2 changes observed from the null model to either Glucose or Glycerol models are very similar, indicating that the constraints imposed by the change of a unique carbon sources are of the same amplitude, and have comparable effects at the metabolic level. Accordingly, when contrasting Glucose and Glycerol models themselves, most of the reaction (499) have the same CV2. Fifty (50) reaction have a higher CV2 for Glycerol than Glucose, and inversely, 98 reactions have a lower CV2 in Glycerol than Glucose models. The detail of genes implied in the most different reactions (with higher and lower CV2 differences) for each comparison is given in Table 2.6.

Reactions	Extent of the CV2 difference	genes implied	COG categories	mmetabolic Pathways
rxn10131_c0	1,000	multiple_genes	#N/A	#N/A
rxn10121_c0	1,000	PFL01_RS08975	Energy production and conversion	Nitrogen metabolism
rxn05625_c0	1,000	PFL01_RS08985	Inorganic ion transport and metabolism	#N/A
rxn05627_c0	1,000	PFL01_RS08985	Inorganic ion transport and metabolism	#N/A
rxn05937_c0	0,989	multiple_genes	#N/A	#N/A
rxn12822_c0	0,989	PFL01_RS22920	Amino acid transport and metabolism	#N/A
rxn03239_c0	0,972	multiple_genes	#N/A	#N/A
rxn02804_c0	0,972	multiple_genes	#N/A	#N/A
rxn03240_c0	0,972	multiple_genes	#N/A	#N/A
rxn10162_c0	0,972	multiple_genes	#N/A	#N/A
rxn10163_c0	0,971	multiple_genes	#N/A	#N/A
rxn00693_c0	0,963	PFL01_RS15620	Amino acid transport and metabolism	Multiple
rxn00904_c0	0,852	PFL01_RS24160	Amino acid transport and metabolism	Valine, leucine and isoleucine biosynthesis
rxn00912_c0	0,807	multiple_genes	#N/A	#N/A
rxn00539_c0	0,800	multiple_genes	#N/A	#N/A
60 reactions	[...]	multiple_genes		
rxn01204_c0	0,800	PFL01_RS00965	Amino acid transport and metabolism	Multiple
rxn10160_c0	0,800	PFL01_RS20690	Carbohydrate transport and metabolism	#N/A
rxn02173_c0	0,800	PFL01_RS20695	Carbohydrate transport and metabolism	Galactose metabolism
rxn02429_c0	0,800	PFL01_RS20705	Carbohydrate transport and metabolism	Galactose metabolism
rxn01199_c0	0,800	PFL01_RS13280	Carbohydrate transport and metabolism	Pentose and glucuronate interconversions
rxn03974_c0	0,800	PFL01_RS19665	Coenzyme transport and metabolism	#N/A
rxn00508_c0	0,800	PFL01_RS16450	Energy production and conversion	Multiple
rxn00509_c0	0,800	PFL01_RS16450	Energy production and conversion	Multiple
rxn00905_c0	0,800	PFL01_RS03095	Nucleotide transport and metabolism	Multiple
rxn00623_c0	0,789	multiple_genes	#N/A	#N/A
rxn01014_c0	0,764	multiple_genes	#N/A	#N/A
rxn00903_c0	0,666	PFL01_RS17385	Amino acid transport and metabolism	Multiple
rxn11268_c0	0,513	multiple_genes	#N/A	#N/A

Table 2.6: **List of reactions for which the CV2 varies the most between the null and the Glucose model.** The extent of CV2 difference corresponds to the absolute value of the CV2 of the null model minus the CV2 of the glucose model. These genes are expected to be the most prone to be modified if the organisms were to specialize to either Glucose or Glycerol. Are listed genes for which the CV2 variation was >0,5 and for which the genes were recognized.

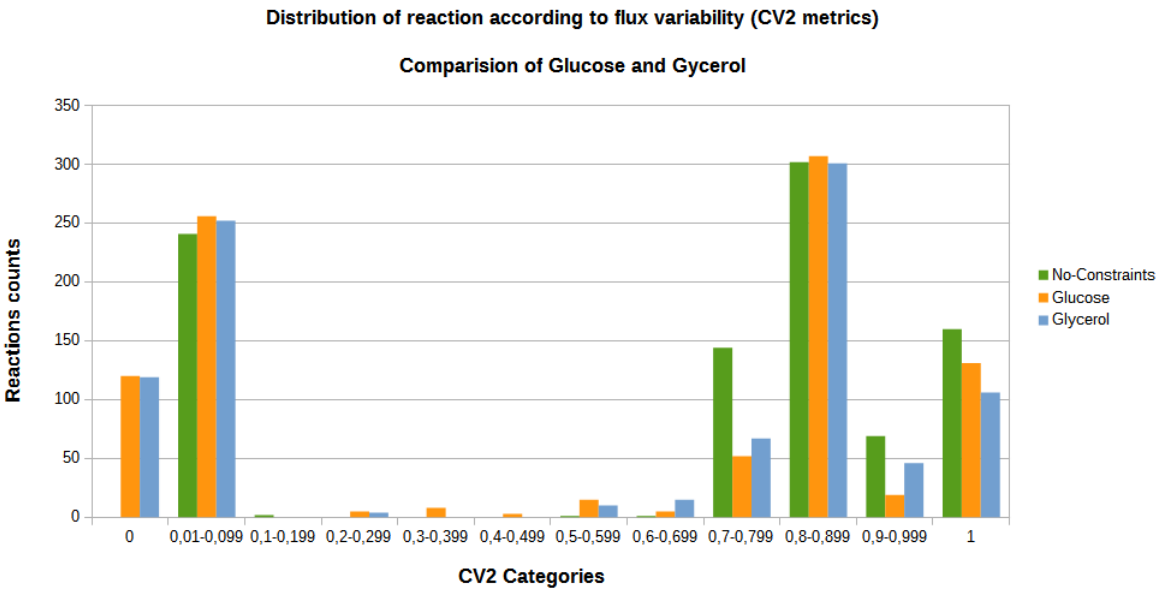


Figure 2.12: **Comparison of the distribution of reactions between the model without constraints, and Glucose and Glycerol models according to their CV2 metrics.** The bigger differences are perceived for CV2 categories of 0, [0,7-0,8] and [0,9-1]. This is interesting as it may mean that some unconstrained reactions in the null model have become constrained in the other models, and respectively, that some reaction's flux are constrained to no flux at all (most of the CV2=0).

2.5 Conclusion on the chapter

The metabolic modeling approaches was used to encompass the extent to which Flux Balance and variability Analyses could be used to forecast the apparition of mutations on a gene. Through FVA we could infer :

- the statuses of genes, which indicates the importance of their activity under different environmental constraints.
- the flux characterization (CV2) of each metabolic reactions according to different environmental constraints.

The direct status and CV2 metrics give information by themselves: we globally expect that excluded-status genes can accumulate more mutations, and similarly for high-CV2 genes. The *comparison of* statuses and CV2 of the null model with the constrained model gives further information into which genes may be modified during a specialization event, considered as an evolution from the null model.

Excluded reactions entail genes whose expression should be shut off, and the CV2 of the excluded reaction is zero, which means that the constraint on these genes to not be expressed is strong. We should thus expect to see the genes

implied in Excluded metabolic reactions to harbor stop mutations, or mutations down-regulating them. It is interesting to observe that many of the predicted excluded genes were involved in the nitrogen, galactose, taurin and pentose metabolisms. More particularly, the three pathways mentioned last harbor only genes with the excluded status, which may traduce of stronger constraints on these pathways. We can thus expect these entire metabolism to be affected by the specialization. COG categories that emerge as supporting the most genes predicted to change are the amino acid, carbohydrate transport, energy production and inorganic ion transport and metabolisms categories.

The FVA modelling has a great potential to characterize the activity of reactions in constrained model, reflecting actual environmental conditions. Yet, several limits can be emphasized. We exposed that even if the activity of metabolic reactions could be forecast, the inference of the fate of genes entailing these reactions was not straightforward. Indeed the relationship between genes and reactions is not linear, as one gene can be implied in several reactions (pleiotropy), and one metabolic reaction can depend on several genes. Additionally, at the larger scale of metabolic pathways, it appeared that reactions often overlap in several pathways (since the metabolic network is an interconnection of somewhat arbitrarily defined pathways). The linearity of our rationale is too distant from the reality of the metabolism complexity to predict with high accuracy which genes will mutate.

The conceptual framework of this study, i.e. the niche reduction resulting from a stable environment inducing the specialization, has been investigated in depth by different complementary metabolic modeling strategies, and despite the known limits of the approach, the work presented herein allowed to forecast how the generalist microorganism studied would turn to a specialist trajectory.

There are two possibilities to test the hypotheses generated by the predictions: By genetically engineering micro-organisms, it is possible to knock-out targeted genes to verify the effects of pathways shutdown. Alternatively it is possible to test the actual evolutionary specialization through *in vitro* experiments and to compare the outcome of the experiments (e.g. *de novo* genetic modifications) with the predictions.

In the second chapter of this section, from an experimental *in vitro* evolutionary experiment realized on a bacterial population of *P. fluorescens* Pf0-1, we detected and analyzed the dynamics of *de novo* mutations in the population which was grown under controlled conditions.

Chapter 3

In vitro evolutionary experiments and mutations analysis

3.1 Preamble of the Chapter

In the Introduction section of this PhD thesis the knowledge gathered on the predictability of evolution was presented, and a new conceptualization on how to enable further prediction was introduced. By mobilizing knowledge and tools from the field of system biology, we demonstrated in the first chapter of this section that it is possible to modelize the specialization of a generalist model organism toward specialization within a stable environment. In spite of the extensive number of studies comparing generalist *vs* specialist behaviors, so far the evolutionary transition from being generalist to being specialist has been poorly investigated experimentally. A second relevant phase of our driving conceptual framework for the predictability of evolution is thus to demonstrate *in vivo* the possibility for a given organism to specialize. The ambition of the following section is to address this specialization process by performing experiments under controlled conditions to retrieve synchronic and diachronic informations on the dynamics of the genomic changes of a model generalist bacteria, *P. fluorescens* Pf01.

The chapter below will eventually be divided into two articles. The first article will focus on the results of the mutation patterns observed during the evolutionary trajectory followed by *Pseudomonas fluorescens* Pf0-1 during an *in vitro* experiment. In this article the dynamic of specific mutations and their potential effects will be presented.

Evolutionary trajectories of specialization in *P. fluorescens* Pf01.
Alix Mas, Philippe Vandenkoornhuyse, Pierre Peterlongo, Wesley Delage & Yvan Lagadeuc

The second article will present a proof of concept strategy to predict evolutionary trajectories of bacteria, by comparing predictions made via metabolic modeling (chapter 1 of this section) and the modifications of the genome observed after an *in vitro* evolutionary experiment entailing specialization.

Predicting the evolution of *P. fluorescens* in a specialization context: a proof of concept strategy based on Metabolic modeling.
Alix Mas, Yvan Lagadeuc & Philippe Vandenkoornhuyse

At the present time, the information of these two articles are merged into one chapter to prevent overlapping information.

3.2 Introduction

The width of an organism's niche traduces its panel of adaptations possibilities. A specialist organism will be adapted to a narrower range of environmental conditions, while a generalist species will be able to thrive in a wide panel of environments and use a variety of resources (Townsned et al., 2003). A major difference between specialist and generalist organisms is their efficiency in exploiting resources: it is widely acknowledged that generalist species are less efficient to exploit a given resource than specialists would be (Futuyama 1988; Sultan 1992). This results in that, for a particular and constant environment, the specialist species will have more efficient behaviors than generalist species, pointing to the advantage of being specialists in stable environments. It is thus interesting to look into the potential switch of niche type (generalist vs specialist) in the context of constant environments. Specialization at the metabolic level can have two different sources (See section 2.4.3 part I): either the specialization stems from adaptive trade-offs, where the activity of some metabolic functions will be favored over others or the specialization ca be due to a high mutation rate resulting in the accumulation of neutral mutation that would be deleterious in another environment (Leiby and Marx, 2014).

To investigate the specialization evolutionary trajectory of a bacteria in different medium, two *in vitro* evolutionary experiments were realized. The aim of the work presented herein was twofold: 1) to investigate the potential of a generalist bacterium, *Pseudomonas fluorescens*, to follow a specialization evolutionary trajectory through mutations selection and fitness increasing within a stable environment -by analyzing the mutation dynamic during a specialization event when minimizing genetic drift on a short term experiment, and 2) to determine if the genes targeted through metabolic modeling were showing off expected patterns of mutations. The rationale behind the experiments can be integrated in the running schema of metabolic modelling (Figure 3.1).

In the first experiment, a bacterial population is subjected to evolution in a medium with glucose as the unique source of carbon, to detect evidences of a rapid specialization trajectory in a constrained environment. For example we expect genes implied in the metabolism of alternative carbon-sources to be modified along the adaptation of the bacteria to their environment.

In the second experiment, the same initial population is subjected to a medium where glycerol is the unique source of carbon. Lipases were added to the medium to ensure the total degradation of glycerides present. As glycerol is

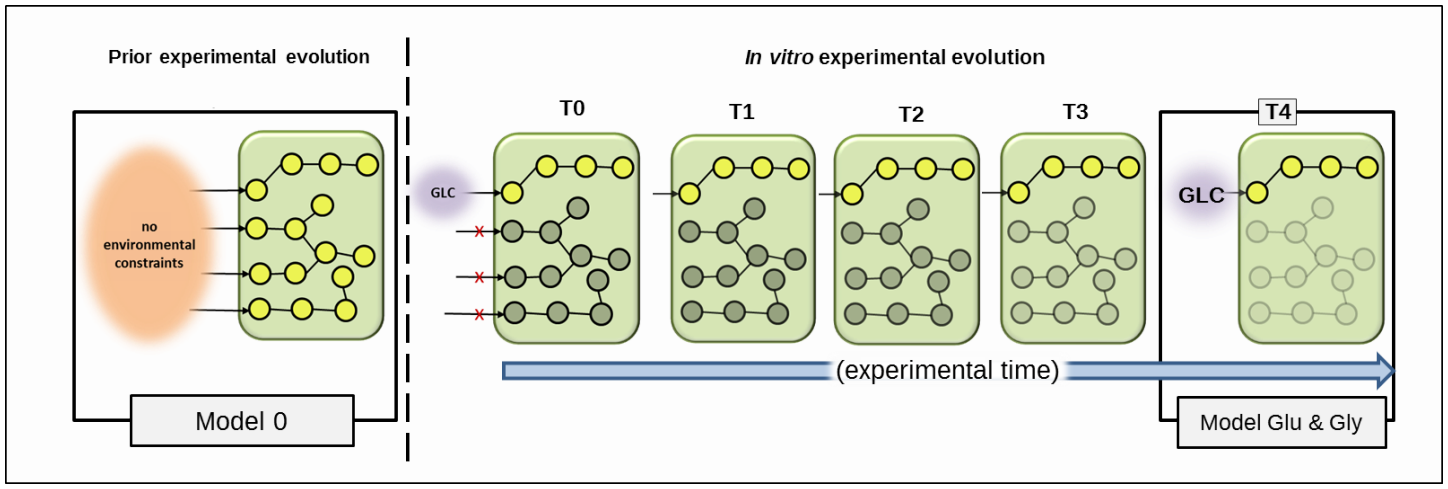


Figure 3.1: **Experimental evolution rationale, in the frame of metabolic predictions.** Green rectangle is the organism, circles and lines represent the metabolic network of the organism, arrows the uptake of nutrients. Yellow circles are 'activated' metabolism, green circles are inactivated. T0-T4 are the sampling times along the experiment. The goal of the experimentation is to constraint bacterial population to evolve in environments where the medium allows access to a single carbon source (glucose for the first experiment, glycerol for the second one) assuming that the bacterial populations will specialize to the carbon sources present. GLC is the constraint and refers to either glucose or glycerol. A reference to the metabolic models presented in the previous chapter contextualizes the comparison that will eventually be made between the model and experimental data. The null model is unconstrained carbon-source wise, entailing a generalist metabolism, while model 2 and 3 are constrained with respectively glucose and glycerol as unique carbon sources, entailing specialized metabolisms.

a product of glyceride degradation by lipases, in a medium were glycerol is already present and lipases are artificially provided, we expect the production of lipases by bacteria to be redundant (with the lipase 'function' provided by the environment), and thus submitted to a lift of selection.

When studying genes here the focused is set on three levels of information:

- the function of the genes, usually corresponding to the protein they encode, which is retrieved in the NCBI database (genes features, https://www.ncbi.nlm.nih.gov).
- the metabolic pathways of these proteins, which permits to make direct link with the modelling work, and
- COG (Clusters of Orthologous Groups) categorization of the genes which enable a functional classification of the genes (Box 3.1).

COG categories

 Nucleic Acids Research, 2000, Vol. 28, No. 1 33–36

The COG database: a tool for genome-scale analysis of protein functions and evolution

Roman L. Tatusov, Michael Y. Galperin, Darren A. Natale and Eugene V. Koonin*

The database of Clusters of Orthologous Groups of proteins (COGs, Tatusov et al., 2000) has been incepted as a phylogenetic classification of proteins from complete genomes. Each COG includes proteins that are thought to be orthologous, i.e. connected through vertical evolutionary descent. The purpose of the COGs database is to serve as a platform for functional annotation of newly sequenced genomes and for studies on genome evolution. To facilitate functional studies, the COGs have been classified into 17 broad functional categories, including a class for which only a general functional prediction, usually that of biochemical activity, was feasible and a class of uncharacterized COGs. Additionally, some of the COGs with known functions are organized to represent specific cellular systems and biochemical pathways.

Box 3.1 Clusters of Orthologous Groups

3.3 Material and methods

3.3.1 Evolutionary experiment

Bacterial isolates for experiments

The bacterial strain used in this study is *P. fluorescens* spp. Pf0-1 (from the collection of Prof. Levy at the University of Boston in 2014, see Figure 3.2).

P. fluorescens is a physiologically diverse species of opportunistic bacteria (gammaproteobacteria) found throughout terrestrial habitats and often used for

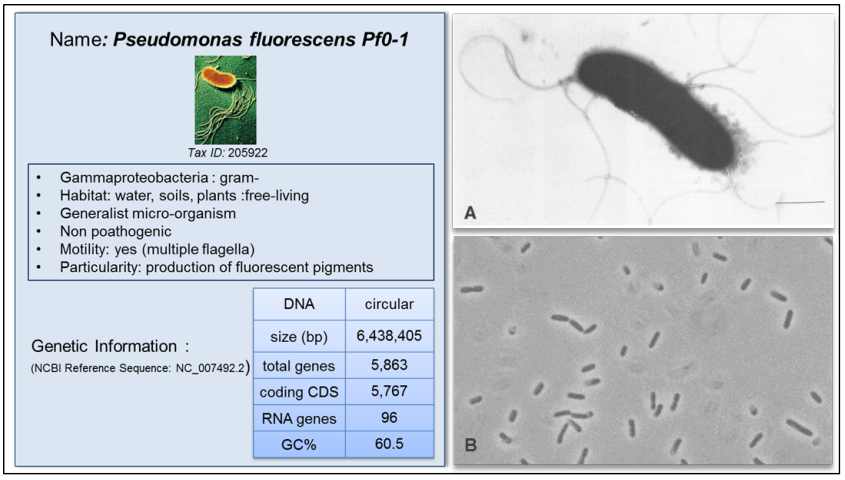


Figure 3.2: ***Pseudomonas fluorescens* Pf0-1 Identification Card**: On the left are basic information for *Pseudomonas fluorescens* Pf0-1, **(A)** single individual of Pf0-1, note the flagella, **(B)** A population of *P. fluorescens* Pf01.

experimental work. The species contributes greatly to the turnover of organic matter and, while present in soil, it is also abundant on the surfaces of plant roots and leaves. The strain is particularly interesting in that its whole genome is completely sequenced (Silby et al., 2009) and a well curated recent annotation (December 2015) is available on the SEED database (http://www.theseed.org). *P. fluorescens* Pf0-1 also seemed convenient for growth in chemostat because of its basic resource requirements and its little promptness to biofilm formation. Moreover this bacterium was chosen for several of its characteristics (such as its motility, flagellated body, ability to consume several carbon sources, fluorescens and the thorough genomic knowledge existing) that make it coherent for our evolutionary study, as we expect modifications in Carbon use, and that motility will be modified by the experiments, as presented in the hypothesis above. Additionally, *P.fluorescens* is a quite common bacteria which has the specificity to evolve rapidly under new environmental conditions and can promptly generate mutants (Rainey and Travisano, 1998). Beside the fact that *P. fluorescens* is a model organism another reason for this choice was the possibility to make targeted genetic manipulations, which could validate the evolutionary predictions in a near future.

A suspension of *P. fluorescens* spp. Pf0-1 has been prepared and grown in controlled conditions, as described below.

Preparation of the initial microbial suspension

The microbial culture was prepared starting from the frozen stock received. It was inoculated into 3ml of lysogeny broth medium (LB, Luria-Bertany) and incubated at 29°C and 60 rpm overnight to re-initiate growth. The experiments were conducted in Pseudomonas Minimal Media (PMM, recipe in Box 3.2). The bacteria were switched from the nutrient rich medium (LB) to the minimal medium used for the experiment (PMM-Glucose) gradually to avoid physiological shock. Samples of the cultures were successively re-suspended in LB diluted with PMM supplemented with Glucose (PMM-Glucose) as follow: PMM25%-LB75%, then PMM50%-LB50%, PMM75%-LB25%, and finaly PMM100%. The fresh culture was then plated on solid LB in petri-dishes and incubated at 29°C. After 24 hours a colony coming from a single *P. fluorescens* cell was retrieved and diluted in 1 mL of 100% PMM-Glucose. The culture obtained was used to inoculate the bioreactor, and once in exponential phase, the volume of the culture was stabilized to 350 mL.

The PMM-Glucose composition per liter is the following:

- 1.2ml of 1M MgSO4
- 25ml of 0,67M Glucose
- 100ml of 10XPPM Salts and 1ml of trace element solution
  PMM Salts solution contained per liter:
- 79.9g of 0.35M K2HPO4Â·3H20
- 29.9g of 0.22M KH2PO4
- 10.6g of 0.08M (NH4)2SO4
  Trace element solution was prepared as follow:
- FeSO4 10g
- ZnSO4.7H2O 2.5
- CuSO4.5H2O 1g
- MnSO4.5H2O 1g
- CoCl2.6H2O 1g
- Na2MoO4.2H2O 1g
- CaCl2.2H2O 5g
- Na2B4O7.10H2O 0.2g dissolved in 1 liter of 5 M HCl

Box 3.2: Pseudomonas Minimal Medium (PMM) composition.

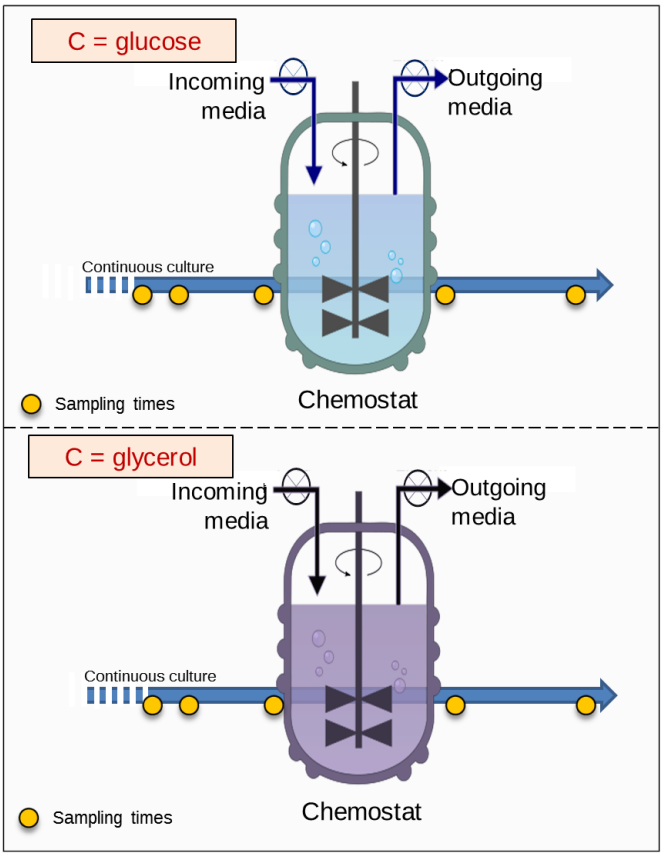


Figure 3.3: **Experimental evolution setup.** The two experiments were realized in chemostat to enable continuity and high culture volumes to limit the genetic drift (high population size). Samples analyzed (yellow dots) are distributed along time of the experiments to ensure a dynamical study of evolution. Except for carbon sources supplemented, the two experiments were run in similar conditions.

The same protocol was implemented for the experiment in PMM-Glycerol. Except for a few differences : First, the carbon source was switched from glucose to glycerol, with a conservation of carbon concentration (1,2g of carbon per liter of PMM). Additionally to each bottle of fresh media was added 0.8 mg/L of lipases to ensure the degradation of any glycerides present in the media. Also, cobalt was removed from the trace element solution as it hinders lipases activity.

In vitro evolutionary conditions

In order to force an evolutionary trajectory of specialization in *P. fluorescens* Pf0-1, constant and constrained environmental conditions were imposed during the culture to favor a strong evolutionary force that would excess genetic drift. For these reasons, chemostats are more than adequate to study adaptive evolution

Figure 3.4: **Experimental setup of the running experiment**.

(Dykhuizen, 1993). Therefore the *in vitro* experiments were realized in a bioreactor (MiniBio, Applikon®) which allows for precisely controlled and continuous culture. The fact that the culture is continuous is an important factor, as it permits to avoid bottlenecks usually happenings with more classical studies of evolution in batch cultures. This bioreactor can hold up to 500 mL of culture, it has a central stirrer with two level propellers, and sealed entries for input and output of solutions and oxygen (also permitting sampling at will). The pH, oxygen and temperature are measured continuously with multiple sensors. Each parameter is set at the beginning of the experiments and automatically balanced afterwards. Having a continuously renewed medium with a high volume (350ml) of culture entails a large population size and thus excludes genetic drift as being the main evolutionary force in such experiments (Figure 3.3).

Therefore, we designed an evolutionary experiment that consists in monitoring the growth of the prepared microbial suspension under controlled conditions. Constant and homogeneous exposure to the same carbon source, Glucose dissolved in minimal medium, was done in a closed volume of V = 350 mL where local conditions (temperature, oxygen, pH, agitation) were accurately controlled and monitored: pH 6.8 (via the automatic titration with 1M NaOH), temperature of 29°C - known to be optimal for the growth of *P. fluorescens* Pf0-1; agitation was fixed at 250 rpm to maximize homogenization while minimizing foam formation. An homogeneous and stable medium is essential in the evolution of specialization of a population. If the parameters vary temporally, then the effect of mutations acquired may change when the environment changes, and if the parameters vary locally, sub-populations may follow different evolutionary path. Once the microbial suspension was inoculated within this device, the solution of PMM-Glucose (composition described below) was injected at the constant flow of Q = 0.48 mL/min while, at the same time, the same amount of liquid

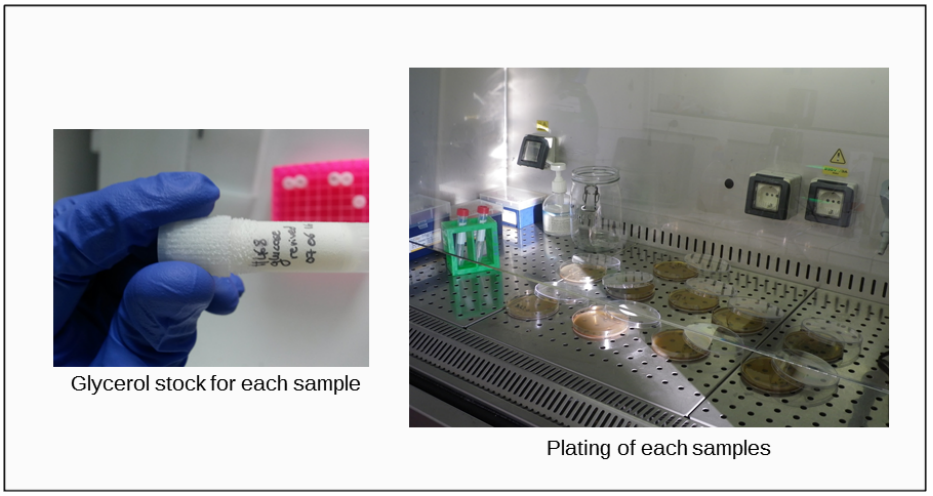


Figure 3.5: **Pictures of a cryotube for Glycerol Stock and agar plates under sterile conditions for plating daily samples**.

was withdrawn from the bio-reactor during the whole experiment. The internal volume was continuously stirred, assuring well-mixed conditions. Under these conditions, we assume that bacteria continuously reproduced.

Experimental Procedures

The experimental population was sampled daily to complete the following tasks:

- 1) Preparation of samples for the nucleic acid extractions : Two falcon tubes containing 50 mL of cell culture were centrifuged (5000 rpm for 20 min), 50 mL of culture proved to produce an acceptable amount of biological material to perform nucleic acid extractions. The centrifugation results in the separation of the liquid phases of the initial suspensions and the microbial populations that are collected at the bottom of the tube as pellets. The latter were immediately stored at -80°C for further DNA and RNA extractions (see sequencing procedure).
- 2) Estimation of bacterial abundance : diluted samples were plated onto solid PMM-Glucose (respectively PMM-Glycerol for the alternative experiment). Several dilution were tested and incubation was realized at 29°C for 48 h before counting the number of colonies formed.
- 3) Archiving of daily samples: For each sample of bacteria retrieved, 500 microL of the culture was mixed with 500 microL of 70% glycerol (>99%), for molecular biology (Sigma), and stocked in cryotubes at -80°C. These archives were used to restart the culture for subsequent experiments.

3.3.2 Sequencing collected samples

Procedure

In this study, both for Glucose and Glycerol experiments, five sampling points were selected along generation time for sequencing. A first sample at the beginning at the experiment, three sample at intermediary times and one at the end of the experiment. The number of generations elapsed per sequenced sample is equivalent between Glucose and Glycerol experiments.

Extraction of DNA

Both for Glucose and Glycerol experiment, DNA from the 5 samples was extracted and purified with the DNeasy Plant Tissue kit (Qiagen) according to the manufacturer's instruction, with the exception that it was optimized for bacterial DNA with no tissue lysis phase. Prior to sequencing, quality checks of extractions and purifications were performed (Nanodrop2000 (ThermoScientific) and BioAnalyzer (Agilent Technologies)). DNA concentrations were measured through PicoGreen spectrofluorometry on a SpectraMax (Gemini). DNA was then fragmented at 200 bp (Covaris technology), without posterior purification, and the fragmentation quality was checked (Caliper LabChip).

Sequencing

DNA libraries were constructed with the NebNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs) following the manufacturers instructions. Initial samples contained 500ng of concentrated (with SpeedVac) fragmented DNA. NebNext adaptors were provided in the NEBNext Multiplex Oligos kit for Illumina (Neb). Post PCR and post purification librairies quality were checked on a high sensitivity chip for Bioanalyzer (Agilent Technology).

The samples were sequenced at the Biogenouest Genomics center. Isolates were subjected to RapidRun paired-end whole genome sequencing on the Illumina HiSeq2000 platform (5 samples, paired-end data of 2x100pb for each experiment) (Illumina). The flow cell used had 2 lanes, each containing a pool of the 5 samples.

It is important to recall here that for each sample, it is a population which is sequenced, and not a single genome. After sequencing, sequences were grouped by sample using the the specific indexes used and filtered by their sequence quality value.

SNP

A Single-Nucleotide Polymorphism, commonly referred to as a SNP (read 'snip'), corresponds to the variation of a unique nucleotide — A, T, C or G — between two DNA sequences.
In the short following sequences: the 5th nucleotide is a SNP.

SNPs are traces of evolutionary events (mutations). In our study we focus on SNPs with the underlying idea that for genetic regions that may have under-gone relaxation of selection pressure, mutations could accumulate without being deleterious, thus SNPs would be more frequent.

Figure 3.6: **Single Nucleotide Polymorphisms.**

3.3.3 Genetic Analysis

For each experiment (Glucose-Glycerol), in order to uncover patterns and dynamics of mutations specific to specialization, the mutation content of the five time-samples was analyzed and compared. The mutations detected were also analyzed in regard to the FVA modeling predictions (see Metabolic Modeling).

Detection and characterization of mutations

To detect mutations in the populations, we used the DiscoSNP++ software (Uricaru et al., 2015) which is a recent informatic tool designed for discovering Single Nucleotide Polymorphism (SNP, see Box 3.6) from raw sets of reads obtained with Next Generation Sequencers, with no need for genome reconstruction. DiscoSNP++ has the important advantage of enabling SNP detection independently of a reference genome. The principle on which DiscoSNP++ software is based to detect mutations, is the recognition of similar sequences of small sizes (Kmers) and comparison of downstream read sequences. When a nucleotide is substituted, a "bubble" (on the de Bruijn graph 3.7) is formed between the two identical Kmers. These "bubbles" are potential mutations, which are subsequently trimmed and ranked by the tool.

The non requirement of a reference genome also allows to efficiently ana-

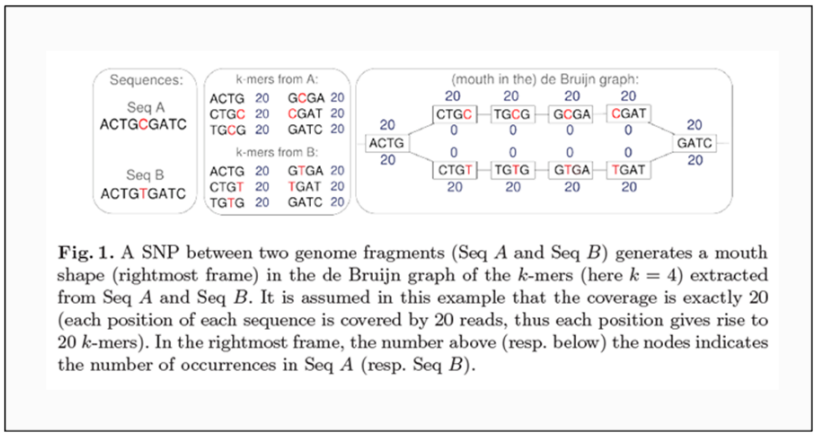


Fig. 1. A SNP between two genome fragments (Seq A and Seq B) generates a mouth shape (rightmost frame) in the de Bruijn graph of the k-mers (here $k = 4$) extracted from Seq A and Seq B. It is assumed in this example that the coverage is exactly 20 (each position of each sequence is covered by 20 reads, thus each position gives rise to 20 k-mers). In the rightmost frame, the number above (resp. below) the nodes indicates the number of occurrences in Seq A (resp. Seq B).

Figure 3.7: **de Bruijn's Graph, extracted from uricaru et al, 2015**

lyze huge data quantity, with low memory requirement (Uricaru et al., 2015). In comparison to other tools available, DiscoSNP++ provides at least similar precision and recall values, and also provides the advantage of ranked predictions of mutation that are less likely to be false positives. DiscoSNP++ was particularly interesting in the frame of our analysis where sequenced sample corresponded to population and not single organisms. Additionally, the number of input read sets is not constrained, the reconstruction of a genome was thus of no interest for us. The software is composed by two modules. The first module, kissnp2, detects mutations from read sets. To facilitate downstream genotyping analyses, a second module, kissreads2, enhance the kissnp2 results by giving ranks predictions based on outputs quality of reads generating polymorphism, and on mean reads coverage per read set and for each variant found.

The parameters used to run DiscoSNP are the following:

- k=31 (kmer size)
- c=1 (solidity threshold(s))
- b=1 (branching strategy (can be 0,1 or 2))
- D=1 (max size of searched indels)
- t=True (extend bubbles to unitigs or contigs)

A SNP was considered as a true SNP if it was present in at least two samples (i.e. two different time points). The presence of a SNP in a single sample can not be differentiated from sequencing errors. Thus, from this stringent filtering, the deleterious mutations or mutations having a negative effect on fitness are not detected (it is unlikely to find the same negative mutation within sequences at 2 different time-points). Reciprocally, only mutations having a circum-neutral or positive effect on fitness are detected. Once mutations are detected through DiscoSnip, the output information has the subsequent format:

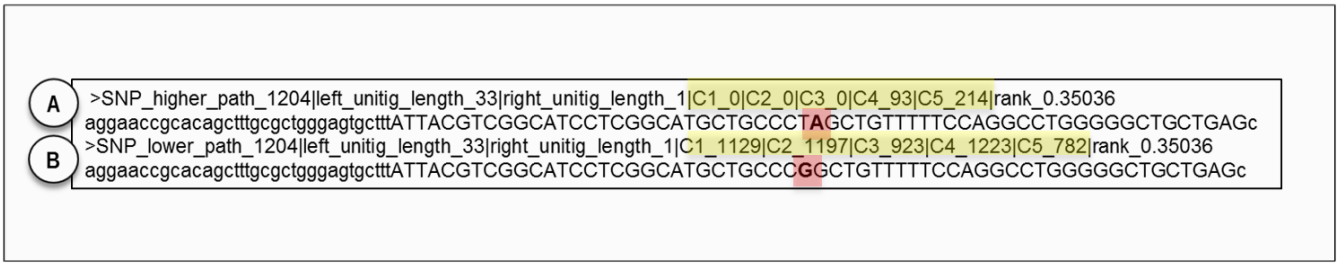


Figure 3.8: **Example of DiscoSNP outputs for a couple of variants (A and B).** The two first lines refer to the variant 'A' of the couple, while the two last lines correspond to the variant 'B'. SNP higher/lower path are DiscoSNP identification tags, and left and right unitig information indicate how many nucleotides are added on each size of the classic 61-nucleotides sequences. Highlighted in yellow is the coverage count of each variant found per sample (C1_0 means 0 of the variant was found in sample one, and respectively for C2-C5); highlighted in red is the location of the SNP detected.

After the DiscoSNP++ variants detection, subsequent steps integrated 1) mutations characterization, 2) description of their frequency patterns in the population, and 3) the interpretation of their impact on the genome and metabolism.

Frequencies of variants in the population were determined as follow: For each couple of variants detected, the coverage of 'original' and 'mutant' sequences is given for each sample. The ratio of mutant on original was calculated for each couple of variants per sample, giving the proportion of mutant variant over its original version. The comparison of these proportions from a sample to another offers the opportunity to follow the frequency dynamics of the variants along time. We can thus infer if, and at which rate the mutant is taking over on its original variant, or if it is bound to stay at low frequencies or even disappear.

For this characterization purpose, a new dedicated tool called 'GetoPe' was developed (using Python 2.7). It allows to overcome identified limits of mutations detection. The aims of this tool were both to homogenize and automatize analyses from the DiscoSNP++ outputs, and to ease the characterization of mutations and determine the impact of these mutations on functional activities.
Precisely, GetoPe (1) identifies and locates the SNPs on the genome, (2) defines the loci and genes implied, (3) determines the impact of the mutations on the genes and protein, and (4) identifies pathways that are affected by these mutations. It also (5) indicates which genes accumulate mutations and which mutations invade the population, based on the coverage values of each sequence given in discoSNP outputs.

This tool was developed in collaboration with Wesley Delage (masters student in 2016) in the frame of my thesis. For a comprehensive and detailed ex-

planation of the *GetoPe* workflow (databases and data used, precise explanation step by step, outcome data) please refer to the Supplementary Information section. Below is a short summary of the steps completed.

From the DiscoSNP output files (Figure 3.8), it possible to retrieve a couple of nucleotide sequences, and the coverage (number of time the sequence is detected) of each sequence in the couple. One sequence in the couple is a particular variant, and the other is an alternative variant. These two sequences differ by just one nucleotide, positioned centrally in the sequence. Usually, one of the variant will correspond to the NCBI published reference genome of *P. fluorescens* Pf0-1, but not always.

(1) To localize the variants on the genome, a comparison of the query sequences against a reference database (reconstituted on purpose to characterize CDS and intergenic regions of *P. fluorescens* Pf0-1 only) is performed through BLAST++. This allow to recognize *where* precisely on the genome (start and stop position of the query sequence) the mutation is located.

Because mutations can be located in repeated regions or duplicated genes, the results are filtered in order to remove each query having more than one position matching with the database. Indeed, with these queries corresponding to several locations on the genome, we cannot determine where exactly the mutation has appeared, or even if the SNP discovered is a false positive (as it would actually exist under different variant forms naturally in the genome, but at different location in the genome).

(2) This information is then crossed with other databases (NCBI, SEED, KEGG, COG, ProGenome) to determine if the position inferred correspond to a gene or an intergenic sequence on the genome. Once this is done, if the mutation is on a gene, the mutated sequence is integrated to the classic gene sequence to reconstruct a "mutant gene". This is necessary to determine the protein encoded by the gene, and the modifications they potentially undergo when carrying mutations.

(3) For each substitution we identified the impact of the mutation, compared to the reference genes, if the mutation encode for a similar amino-acid, we consider the mutation as synonymous. On the other hand if the amino-acid is changed in the mutant gene, the reconstructed protein may be nonfunctional. If a stop-codon appears due to the mutation, the gene and protein will be considered as nonfunctional (i.e. lost). Physico-chemical properties of the newly formed gene inform on the plausible structural changes of the protein induced

by the mutation.
Also, mutation location on the gene (early in the coding sequence, or at the end of the sequence) help to know if a mutation can have an important impact on protein functionality.
(4) Finally a parameter we are also interested in is to know if the mutations we observe were invading the population (i.e. fitness gain). For this purpose we used the coverage of the mutated vs non-mutated sequences detected by DiscoSNP as a proxy of the number of individuals carrying the mutations in the population. This way if the coverage of the mutant was greater in the last time-sample than in the first, we consider that this mutant was invading the population as its frequency was increasing (i.e positive effect on fitness). The faster a mutation invades the population, the stronger is the beneficial effect of the mutation.

Back to modeling

Once mutations were detected and characterized, they were then compared to the predictions made via FVA (see chapter 1 of the section). Predicted 'Excluded' genes were expected to be lost (e.g. stop codon, modification of reading frames by insertion or deletion, modification of promoter regions) or to accumulate mutations because of a lift of selection, while 'Essential' genes were predicted to be selected and thus would support no mutation, mutation enhancing their activity or silent mutations. To this aim, the list of targeted genes was compared to the list of mutations found.

Circos and Krona for visual representation

Circos is a software package for visualizing data and information (Krzywinski et al, 2009). It visualizes data in a circular layout which makes it ideal for representing the circular genome of a bacteria. In our case, we used it to represent the distribution of mutations along genome compared to predicted targets, and we included the effects of a subset of chosen mutations.

Krona allows hierarchical data to be explored with zooming, multi-layered pie charts. It was used to represent the distribution of genes into diverse functional categories of metabolic pathways.

Statistical Analyses

To determine whether genes size (number of base pairs) induced mutation distribution, an ANOVA analysis was performed using the linear model procedure in R 3.1.358 (R Core Team, 2016). The number of mutation per genes were compared between genes sizes to control for the possibility that gene extents is determinant of the number of mutations they carry.

To determine if the fact that genes are 'metabolic' or not may influence on their propensity to be mutated, a Chi2 test was performed using the Chi-squared procedure in R 3.1.358. The number of mutated genes were compared between metabolic and non-metabolic genes to control for the possibility that gene function is determinant of the frequency of carrying mutations.

3.4 Results and Discussion

In order to avoid cumbersome repetitions, for each results presented a higher focus is made on Glucose analysis. When results are similar for glycerol, they are mentioned. If results for Glycerol differ from results for Glucose, they are presented into details.

3.4.1 *In vitro* evolutionary experiments outcome

Glucose

The growth of the microbial population in the bioreactor lasted 575 hours (24 days) with an estimated 173 generations. Each day a sample for DNA/RNA extraction and glycerol-stock were produced and conserved.

Glycerol

The growth of the microbial population in the bioreactor lasted 623 hours (26 days) with an estimated 190 generations. As for Glucose a sample for DNA/RNA extraction and glycerol-stock were produced and conserved each day.

Evolutionary experiments with similar generation times (around 140 generations) are recognized to be insightful for molecular evolution analyses (e.g : Blank et al. 2014; Toll-Riera et al. 2016).

The Illumina sequencing outcome for Glucose and Glycerol experiments are presented in the table below.

	Sample ID	Cultivation time (hours)	Number of Generations	Number of sequences F+R	Coverage (X)
GLUCOSE	0	0	25	68186076	1070
	1	24	33	72349184	1135
	2	111	51	62399034	979
	3	303	99	79459594	1246
	4	468	139	72540634	1138
	Total			354934522	

	Sample ID	Cultivation time (hours)	Number of Generations	Number of sequences F+R	Coverage (X)
GLYCEROL	0	22	7	132408436	2077
	1	118	36	111398684	1747
	2	263	79	126304190	1981
	3	478	143	105410804	1653
	4	621	186	128197240	2011
	Total			603719354	

Table 3.1: **Sequencing information of samples for Glucose and Glycerol analyses from 2 runs of Illumina HiSeq2000.**

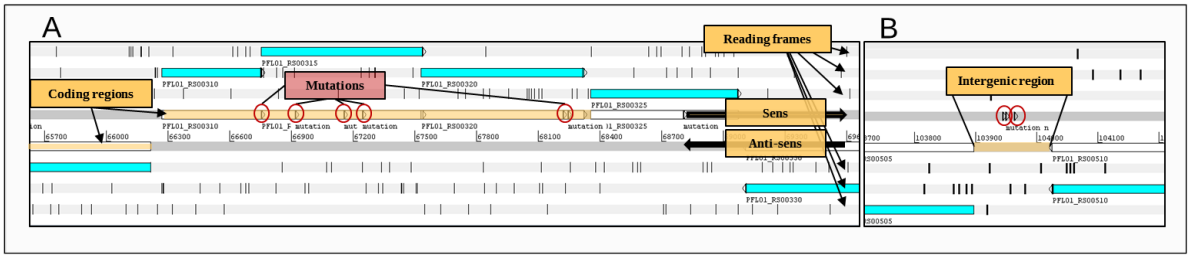


Figure 3.9: **Representation of mutation localization on *P. fluorescens* Pf0-1 genome sequence.** In blue are the CDS (coding sequences) distributed on their reading frames. The upper parts corresponds to CDS read forward and the lower half of the picture correspond to CDS read reversely. These blue CDS are projected in white, on the grey lines in the center of the figure (on the upper line for CDS read forward, and on the lower line for CDS read reversely). **A** Highlighted in yellow are CDS projections on the sequence and we can see that several mutations are found on a single CDS. **B** Highlighted in yellow is an intergenic region (non-coding sequences located in-between two CDS), as seen here intergenic can also carry one or several mutation, despite their usually small size.

3.4.2 Information retrieved through Genetic Analysis

Mutations Analysis for Glucose

Mutations count and location

A total of 3767 mutations were conserved after the pipeline DiscoSNP/GetoPe and associated filters. The 3767 mutations conserved are supported on 2178 locus. Two-hundred and eleven (211) SNPs are located on 171 intergenic regions, and 3533 SNPs are positioned on 2007 coding regions. An example of mutation location is presented in Figure 3.9.

Mutations distribution As remarkable on the green outer circle of Figure 3.10, mutation are spread over all the genome. There are 1881 portions (> to 3 bp) of the genome which do not carry mutation, and some of them are as big as 20687 bp. It thus appears that these regions are particularly conserved, and prone to be kept as they are.

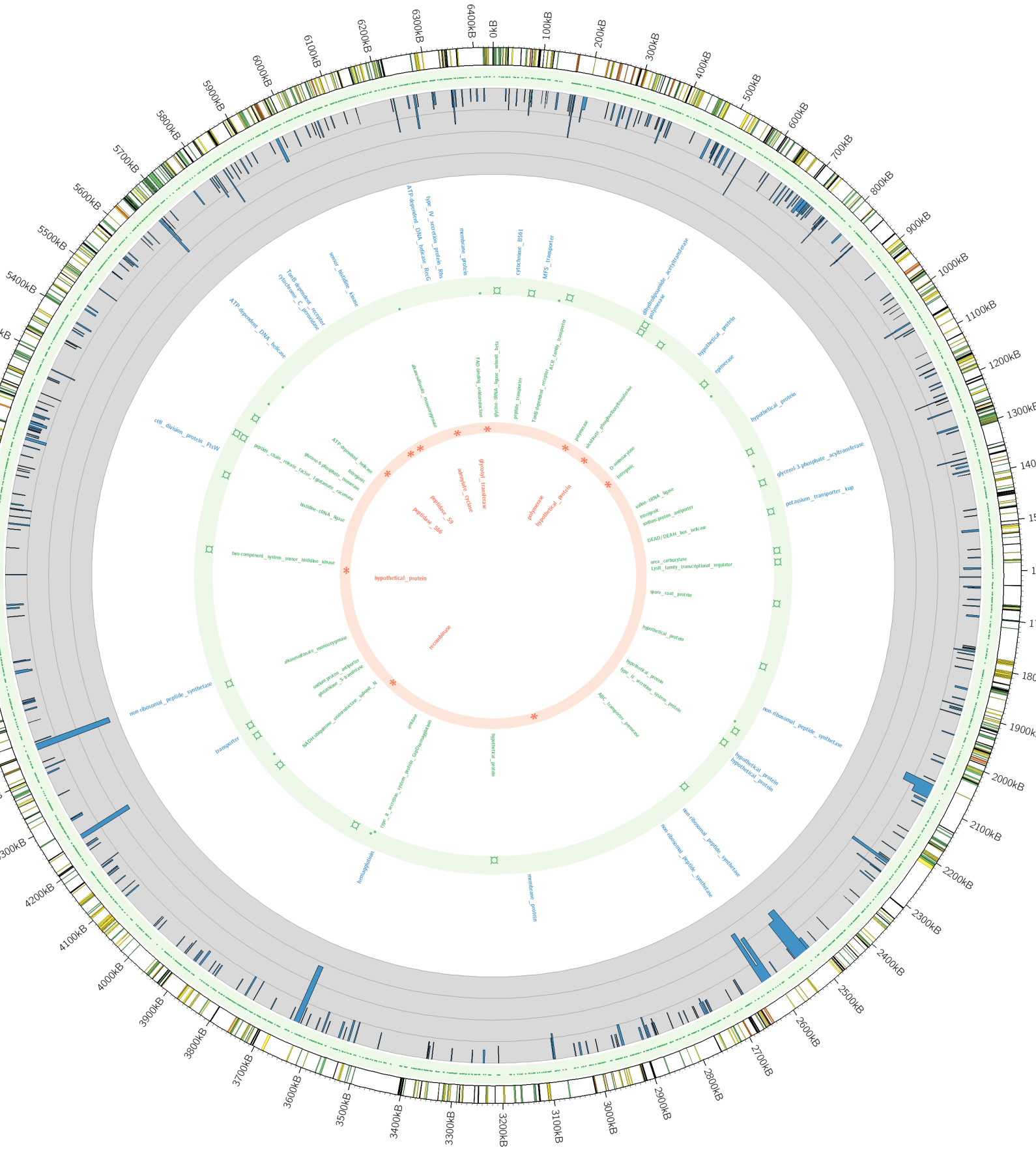


Figure 3.10: **Synthetic description of mutations found on the genome of *P. fluorescens* Pf0-1 during *in vitro* evolutionary experiments in Glucose media.**

[Legend Circos Glucose] Synthetic description of mutations found on the genome of *P. fluorescens* Pf0-1 during *in vitro* evolutionary experiments in Glucose media. The outermost circle represent the statuses of genes according to FVA modelling [Black=Blocked; Red=excluded; Green=Essential, yellow=Alternative]; The histogram represent the number of mutations accumulated by population's genes, and their function is written in blue. The green inner circle and text represent respectively mutations increasing in frequency over experimental time, and their function. When the mutations are synonymous they are represented by a sot only. Finally the inner red circle represent the stop-codon mutations, and their function.

The genome of *P. fluorescens* Pf0-1 is made up of 6438405 bp, and genes sizes span from 76 bp to 16965 bp. As 3767 mutations (single nucleotide modifications) were detected, if mutations were totally random they should be found roughly every 1709 bases. Thus smaller genes would have low probabilities of carrying mutations, while the biggest genes (16965 bp) would be expected to carry on average 10 mutations. If only random, the size of genes (number of base pairs) should therefore be determinant of the number of mutations they carry.

Actually, 83% of the genes modified, whatever their size, support 1-2 mutations, while 11 genes are carrying more than 10 mutations and up to 18. More specifically 1761 genes (out of 5208) smaller than 1709 bp (size of a portion which should bear one mutation if distributed stochastically) carry at least one mutation, and a few of them carry between more than 10 mutations. On the other hand the biggest genes are more rarely altered, and only occasionally support more than 1-2 mutations. Yet, the ANOVA statistic test, run on the linear model to assess the effect of genes size on the probability that they will carry mutations, states that the size of genes significantly affects the number of mutations accumulated on genes (p-value$<$2e-16, F-value =199.27), explaining 35.7 % of the variance (Adjusted R-squared: 0.3571). If the distribution of mutation was purely stochastic, the number of mutation per genes would be highly related to the genes' size. Here, more than 60 % of the variation in mutation distribution has to be explained by other variables or constraints than gene size, suggesting that the distribution of mutation on genes is not utterly random.

It is also important to recall here that each sample represent a population, and not a single genome. We are thus confronted to a collection of individuals, a population of genomes, with the incapacity, from this analysis to determine 1) if for example, the 18 mutations supported by a gene are actually observed on a

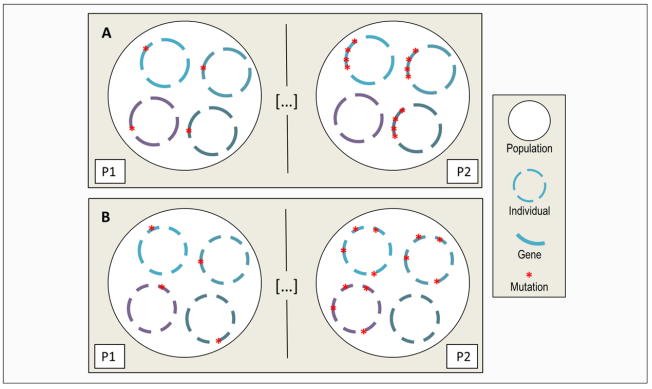


Figure 3.11: **Schematic representation of possibilities for mutation distribution on genes (A) and genomes (B).** The population is schematically composed of four individuals represented by their genomes (dashed circles) made of a few genes (dashed units). **[A;P1]** Describes the case where mutations are all found on a similar genes, but each mutation is carried by a different genome (i.e individuals will carry different mutation, but on the same gene). **[A;P2]** The mutations are all found on a similar genes, and also on similar genomes (i.e one individual will harbor several mutations on a single gene). These two patterns represented are the extremes of a continuum of possibilities, and intermediate situations are most likely to exist. At the whole genome scale, the problem repeats itself: it is possible that each mutation is carried by a different genome as represented in **[B;P1]** or that mutations on different genes will be carried by a single genome**[B;P2]**, and possibly intermediate situations.

single gene entity, or on similar genes of different individuals in the population, as schematized in Figure 3.11, and 2) if several mutations at different loci are carried by a single genome or by as many genomes as there are mutations. Thus, instead of talking of a 'gene' we will refer to a 'population's gene', to precise that different mutations can be scattered over different individual in the population.

Genes with frequency increasing mutations The proportion of mutant over initial variant in each sample permits to follow the dynamic of the mutations overtime. As presented in the Figure 3.12, we can note that there is quite a variability of patterns in mutation dynamics. Indeed, some mutations increase in frequency and others vary over time while globally staying at the same level. Others appear lately and increase readily, and some were present early and slowly decrease. A basic count based on the difference in frequency of the first sample and the fifth sample

Of interest are mutation that increase in frequency over time. The frequency presented is the occurrence of the mutated variant over the initial variant. Thus, if a mutated variant increases in frequency, it means that this mutant is spreading in the population by out-competing its counterpart. These frequency increasing

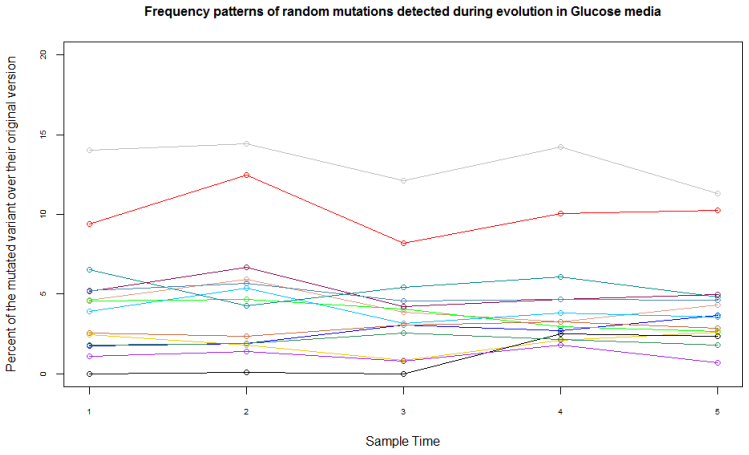


Figure 3.12: **Frequency patterns of a sample of mutations (Glucose).** From the 3767 mutations detected in Glucose experiment, the frequency patterns of a subsample of randomly selected variants is represented here in order to represent the variability existing in the frequency dynamics of mutations.

mutations can be assimilated to beneficial mutations for the individuals, and also to specialization, given our experimental design.

Nevertheless it is important to recall that the frequency observed is the frequency of a given mutant over its alternative variant, and not for a mutant against all other mutants. There could indeed be accumulated crossed effects influencing the frequencies of particular mutation (e.g, genetic hitch-hiking).

In the Glucose experiment, 40 mutations were found to continuously increase in frequency over time (Figure 3.13). All in all, a total of 1711 mutations (45%) were described as being (even slightly) at a higher frequency in the last sample than in the first, but with punctual decrease at in-between sample times.

The option chosen, to focus on strictly increasing frequency mutations, permits to focus on mutation providing the best fitness advantage.

Three (3) of these frequency increasing mutations correspond to stop-codons (presented below) and are also the ones with the highest increase (from zero to 21%). Twelve (12) of the genes supporting frequency increasing mutations are defined as metabolic genes and, through modelling, 2 were classified as Excluded, 5 as Essential, 4 as Alternative and 1 as blocked. The rest of the genes was not considered as metabolic and are categorized in various COG categories as described in Table 3.3.

Mutation Types Amongst the mutations detected after evolution in glucose, 768 were found to be silent (20.4 %), and 2999 are non silent (79.6 %). A total of 21 codon stop were discovered, hindering the associated protein production,

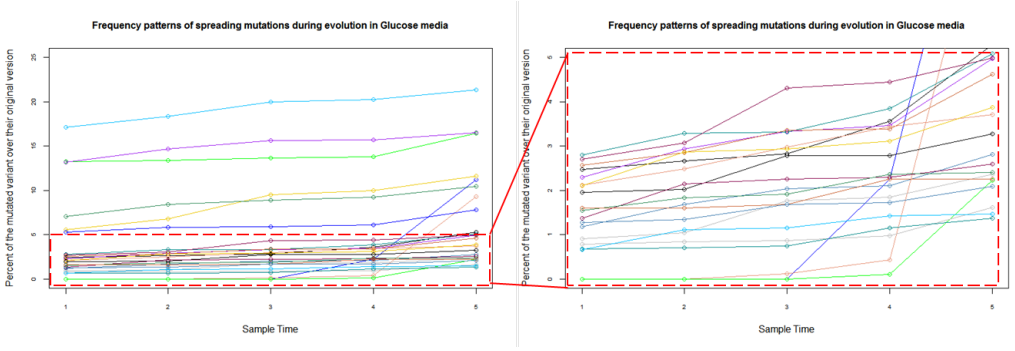


Figure 3.13: **Frequency patterns of the 40 spreading mutations (Glucose).** The frequency presented is the occurrence of the mutated variant over the initial variant.

and 40 mutations were determined to (continuously) increase in frequency in the population over time, potentially traducing of their beneficial effect on individuals carrying these mutations. By 'continuously' we refer to mutations for which the frequency in sample *n* was always superior than the frequency in sample *n-1*. There are also mutations for which the frequency is higher in the last sample than the first, but which underwent frequency variations in between.

The annotation of 271 mutated genes is described as 'hypothetical protein' giving few information on the functions implied while the other modified genes are distributed according to their metabolic and COG categories.

Metabolism wise, 388 genes carrying mutations are directly associated to metabolic function. They harbor 634 mutations. These genes are distributed in 75 different metabolic pathways.

The COG categorization (=functional distribution) of mutated genes is sustained in the Krona Figure 3.14. All functional categories harbor relatively high ratios of mutations. Yet, some categories are more conserved (with a lower rate of mutation). This is the case for functions directly associated to genetic 'mechanisms' such as transcription (30%), translation (28%) and post-translational modifications (32%). Other categories present quite higher mutation rates: it is the case for the transport of inorganic ion (49%), of secondary metabolites (42%), of lipid transport (43%) and of particular interest in the case of our study of carbohydrate transport and metabolism (45%).

Stop-Codons A total of 21 stop-codons mutations were retrieved. Astonishingly half of them (10) are located on the same gene. Two options are possible for the apparition of such a pattern: either stop-codon mutations emerged several time the population, or, a singular gene with the assortment of all these mutations is shared by a subset of the population. It is evident that intermediary

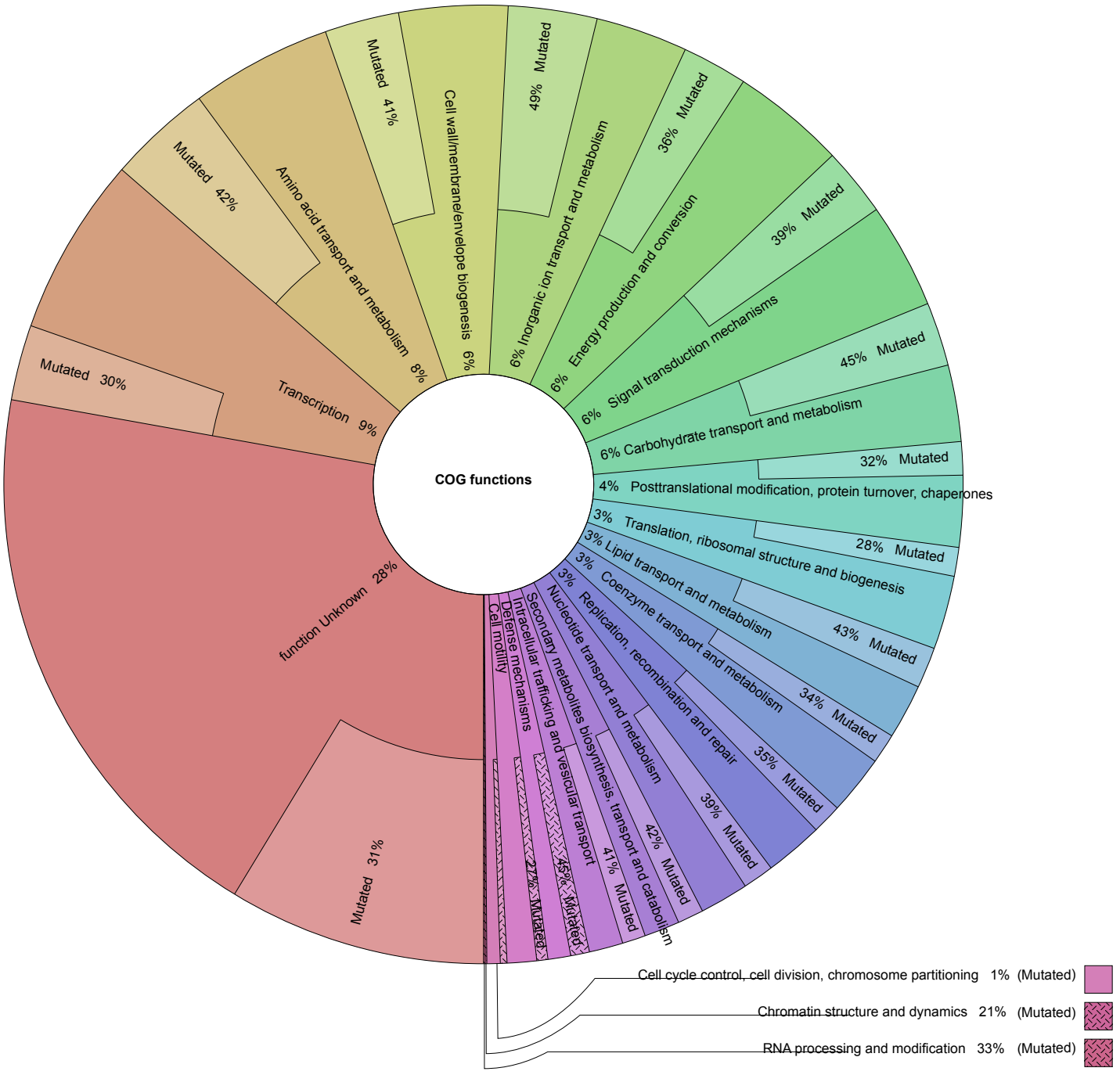


Figure 3.14: **Distribution of mutations according to COG categories.** The figure represents the COG functional categories. The size of the category correspond to the number of genes included in this category and the percentage is the portion of COG it represents. The mutated portions represent the numbers of genes that were found to carry one or more mutation after the experiment -and that could be matched to a COG functional category. As only *genes* mutated and not *mutations number* is represented per category, further information such as synonymous/non-synonymous mutation or frequency patterns could not be inserted.

Gene_ID	Gene_Function	Metabolic_pathway	COG_category	Nb_mutation
PFL01_RS26000	peptidase S9	#N/A	Amino acid transport and metabolism	1
PFL01_RS03290	potassium ABC transporter ATPase	#N/A	Amino acid transport and metabolism	1
PFL01_RS21560	hypothetical protein	#N/A	Cell wall/membrane/envelope biogenesis	1
PFL01_RS02420	polymerase	#N/A	Cell wall/membrane/envelope biogenesis	10
PFL01_RS17455	recombinase	#N/A	Cell wall/membrane/envelope biogenesis	1
PFL01_RS13015	GNAT family acetyltransferase	#N/A	Defense mechanisms	1
PFL01_RS24965	peptidase S66	#N/A	Energy production and conversion	1
PFL01_RS03075	hypothetical protein	#N/A	Function unknown	1
PFL01_RS27640	adenylate cyclase	#N/A	Function unknown	1
PFL01_RS04075	glycine/betaine ABC transporter permease	#N/A	Intracellular trafficking, secretion, and vesicular transport	1
PFL01_RS26305	N-methylproline demethylase	Purine metabolism	Nucleotide transport and metabolism	1
PFL01_RS28615	glycosyl transferase	#N/A	Replication, recombination and repair	1

Table 3.2: **Description of stop-codon mutations found on the genome of *P. fluorescens* Pf0-1 after Glucose*in vitro* evolutionary experiment**

solutions can also be in place.
The gene affected is a Polymerase (O-antigen polymerase) and belongs to an operon also supporting acyltransferase-3. Polymerase are enzymes that synthesize DNA or RNA strands. There are 40 genes in the genome of *P. fluorescens* Pf0-1 encoding for DNA or RNA polymerase, and only 3 encoding for O-antigen polymerase (the two others, associated to glycosyl transferase operon, do not carry mutations). This particular genes do not carry any other mutations than the stop-codon ones.

A closer look in the apparition and fluctuation dynamics of these polymerase stop-condon mutations (Figure 3.15) may give enlightenment on the phenomenon at stake. Indeed in the first three temporal samples, the gene concerned does not carry any of the ten mutations. Mutations start to be apparent only in the fourth sample, at diverse frequency levels (Figure 3.15). The strong increase in frequency of some of these stop-codons (e.g. green, red, blue & cyan lines) mutations from the fourth sample are concomitant with the decrease in frequency of other stop-codon mutations (orange, purple, grey & yellow lines). This suggests that (sets of) different polymerase stop-codons are carried by different subset of the population subjected to clonal interference. Indeed, even if all the stop-codon mutations are beneficial (frequency increase), some might be more beneficial (or associated to other more beneficial mutations on the genome), and thus outcompete and replace the ones already in place.

This pattern is noteworthy as several strong mutations arise in parallel, and at differentiated times, on the same gene.

As seen on Table 3.2, only one mutation is on a gene that could be associated to a metabolic pathway, the other genes are not considered directly as metabolic. Thus, except for the gene encoding N-methylproline demethylase implied in

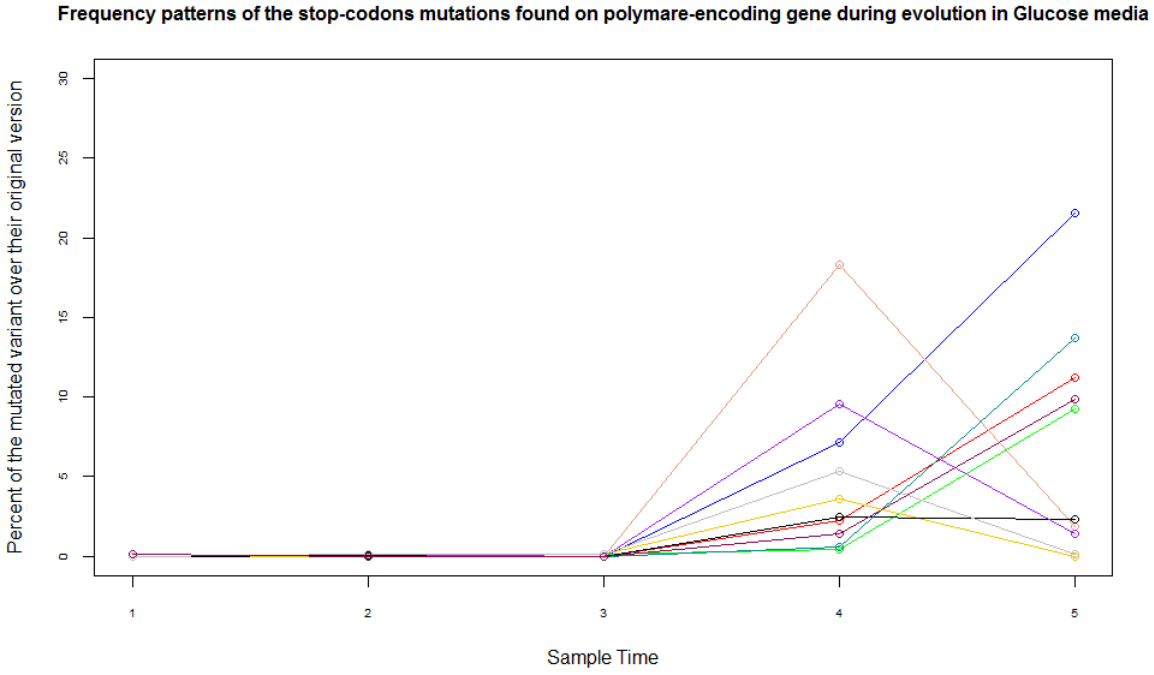


Figure 3.15: **Frequency patterns of top-Codon mutations carried by the same Polymerase-encoding gene.**

Purine metabolism, which was attributed the 'alternative status' via FBA, it is not possible to link previous modeling work to stop-codons retrieved.

The limit of the metabolic modeling approaches used is easily perceptible, as less than 20 % of the genes of *P. fluorescens* Pf0-1 are actually categorized as metabolic. Thus most of the genes stay out of our perception for prediction.

Genes accumulating mutations A total of 2007 genes (and 171 intergenic regions) were found to carry at least one mutation. The graphic 3.16 represents the number of genes carrying between 1 and 18 mutations. Obviously, fewer genes support several mutations, nevertheless, still a substantial proportion of genes support more than one mutation.

If we consider that two mutations associated to a single gene-ID are actually carried by a single gene entity, as well as if we consider that the mutations are carried by two different genes (entities), the fact that several mutations arise on a gene suggest that these genes are particularly prone to mutate.

The Chi-squared test indicates that there is a significant effect (p-value = 2.775e-06) of genes propensity to mutate according to their classification into the metabolic category or not. Metabolic genes are more prone to mutate than their counterparts (Figure 3.17). This observation is susbstantial: the fact that more metabolic genes mutate than other genes during the evolution of specialization

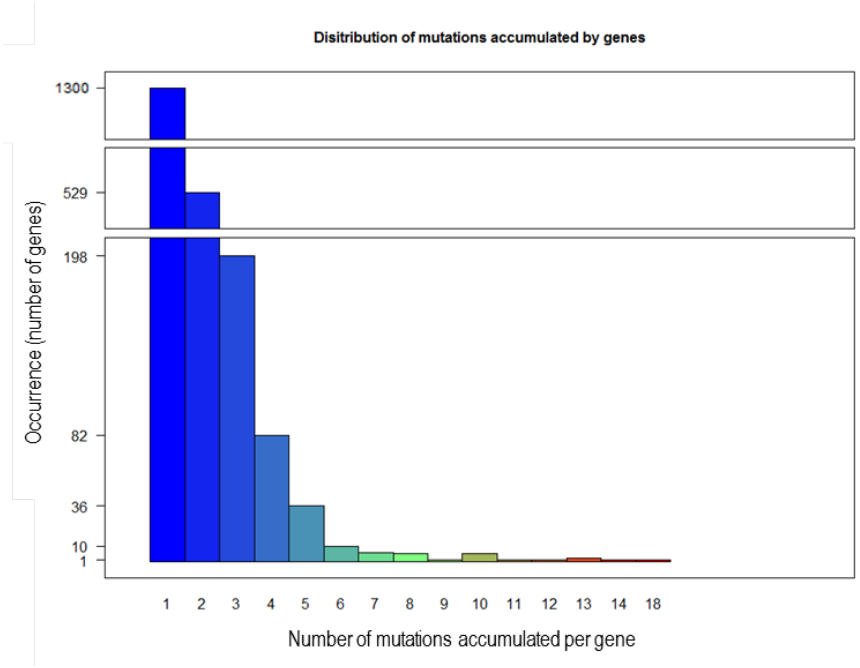


Figure 3.16: **Occurrence of genes accumulating one to eighteen mutations.** Few genes support a high number of mutations (right of the graphic), while the vast majority of genes altered support one to two mutations (left of the graphic).

strongly supports our idea that focusing on metabolism to make evolutionary prediction is essential.

Mutations Analysis for Glycerol

Mutations count and location

From the sequence analyses, a total of 4091 mutations were conserved by the filters imposed. Four-hundred and four (404) SNPs are located on 143 intergenic regions, and 3687 SNPs are positioned on 2008 coding regions.

Mutations distribution

As for Glucose, the Circos Figure 3.18 synthesize the information for Glycerol mutations observed. Here also, the whole genome carry mutations in a mostly homogeneous way. There are still 1934 portions (> to 3 bp) of the genome which do not carry mutation, and some of these regions are as big as 19367 bp.

[Legend Circos Glycerol] Synthetic description of mutations acquired by the genome of *P. fluorescens* Pf0-1 during *in vitro* evolutionary experiments in Glycerol media. The outermost circle represent the statuses of genes according to FVA modelling [Black=Blocked; Red=excluded; Green=Essential, yellow=Alternative]; The histogram represent the number of mutations accumulated by population's genes, and their function is writen in blue. The green

Gene_ID	Fq_increase	Function	FVA status	Metabolic pathway	COG_category
PFL01_RS02420	21,53	polymerase			Cell wall/membrane/envelope biogenesis
PFL01_RS02420	11,22	polymerase			Cell wall/membrane/envelope biogenesis
PFL01_RS02420	9,28	polymerase			Cell wall/membrane/envelope biogenesis
intergenic	6,06	Intergenic			
intergenic	4,19	Intergenic			
PFL01_RS27245	3,4	alkanesulfonate monooxygenase	excluded		Energy production and conversion
PFL01_RS06765	3,38	urea carboxylase		Glycine, serine and threonine metabolism	Function unknown
PFL01_RS05530	3,18	sodium:proton antiporter	alternative		Inorganic ion transport and metabolism
PFL01_RS14045	2,82	hypothetical protein			Function unknown
PFL01_RS10990	2,68	ABC transporter permease			Inorganic ion transport and metabolism
PFL01_RS15935	2,53	hemagglutinin			Intracellular trafficking, secretion, and vesicular transport
intergenic	2,28	Intergenic			
PFL01_RS15955	2,28	type II secretion system protein GspD			Intracellular trafficking, secretion, and vesicular transport
PFL01_RS21990	2,23	two-component system sensor histidine kinase			Signal transduction mechanisms
PFL01_RS24830	2,05	ATP-dependent helicase			Replication, recombination and repair
PFL01_RS09715	1,77	sugar ABC transporter ATP-binding protein	alternative		Carbohydrate transport and metabolism
PFL01_RS18155	1,63	NADH:ubiquinone oxidoreductase subunit N	alternative	Oxidative phosphorylation	Energy production and conversion
PFL01_RS09975	1,59	type II secretion system protein			Intracellular trafficking, secretion, and vesicular transport
PFL01_RS19670	1,45	alkanesulfonate monooxygenase	excluded		Energy production and conversion
PFL01_RS02770	1,33	nicotinate phosphoribosyltransferase	essential		Coenzyme transport and metabolism
PFL01_RS18760	1,24	glutathione S-transferase			Post-translational modification, protein turnover, and chaperones
PFL01_RS08815	1,22	hypothetical protein			Function unknown
PFL01_RS16350	1,11	amidase			Secondary metabolites biosynthesis, transport, and catabolism
PFL01_RS18250	1,08	MFS transporter			Carbohydrate transport and metabolism
PFL01_RS06085	0,92	DEAD/DEAH box helicase			Function unknown
PFL01_RS24210	0,86	glucose-6-phosphate isomerase	alternative	Glycolysis / Gluconeogenesis	Carbohydrate transport and metabolism
PFL01_RS09650	0,85	hypothetical protein			Post-translational modification, protein turnover, and chaperones
PFL01_RS03660	0,83	D-aminoacylase	blocked		Secondary metabolites biosynthesis, transport, and catabolism
PFL01_RS28625	0,82	FAD-binding oxidoreductase			Energy production and conversion
PFL01_RS06995	0,81	LysR family transcriptional regulator			Transcription
PFL01_RS04975	0,8	valine--tRNA ligase	essential		Translation, ribosomal structure and biogenesis
PFL01_RS23835	0,73	glutamate racemase	essential	D-Glutamine and D-glutamate metabolism	Cell wall/membrane/envelope biogenesis
PFL01_RS00055	0,69	glycine--tRNA ligase subunit beta	essential	Aminoacyl-tRNA biosynthesis	Translation, ribosomal structure and biogenesis
PFL01_RS23850	0,68	peptide chain release factor 1			Translation, ribosomal structure and biogenesis
PFL01_RS23120	0,65	histidine--tRNA ligase	essential	Aminoacyl-tRNA biosynthesis	Translation, ribosomal structure and biogenesis
PFL01_RS01105	0,63	TonB-dependent receptor			Inorganic ion transport and metabolism
PFL01_RS18915	0,61	sodium:proton antiporter			Function unknown
PFL01_RS01175	0,59	ACR family transporter			Inorganic ion transport and metabolism
PFL01_RS07660	0,43	spore coat protein			Cell wall/membrane/envelope biogenesis
PFL01_RS00660	0,42	peptide transporter			Secondary metabolites biosynthesis, transport, and catabolism

Table 3.3: **Description of genes carrying frequency increasing mutations (Glucose).**

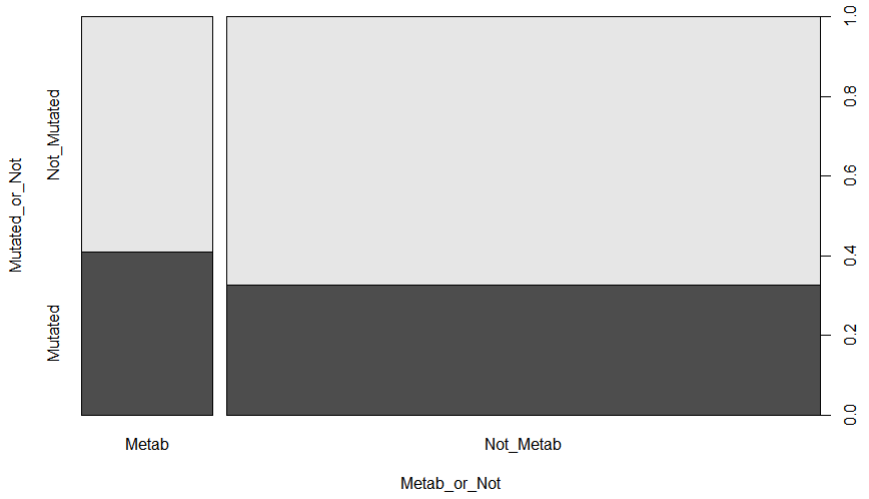


Figure 3.17: **Distribution of mutations according to the affiliation of genes to Metabolic pathways.** Metabolic genes (Metab) are more prone to mutate than non-metabolic genes (Not-Metab) even though the proportion of non-metabolic genes is higher than the proportion of metabolic genes (the width of the graph indicates the number of genes involved).

inner circle and text represent respectively mutations increasing in frequency over experimental time, and their function. When the mutations are synonymous they are represented by a sot only. Finally the inner red circle represent the stop-codon mutations, and their function. As 4091 single nucleotide modifications were detected, if mutations were totally random they should be found roughly every 1573 bases. Once again we can see that it is not the case and that multiple mutations can accumulate on genes of small sizes, while some relatively big portions of the genome (up to 19367 bp) do not carry any mutation.

Similarly as for Glucose the size of genes significantly affects the number of mutations accumulated on genes (p-value$<$2e-16, F-value =421.27), explaining 15.5% of the variance (Adjusted R-squared: 0.1549).

Mutation Types

Amongst the mutations detected after evolution in glycerol, 1151 were found to be silent (28.1 %), and 2940 are non silent (71.9 %). A total of 25 mutations were found to continuously increase in frequency in the population over time, potentially traducing of their beneficial effect on individuals. Seventeen (17) stop-codon mutations were found, hindering the associated protein production. Metabolism wise, 433 genes carrying mutations are directly associated to metabolic function. They harbor 785 mutations and are implied in 62 unique

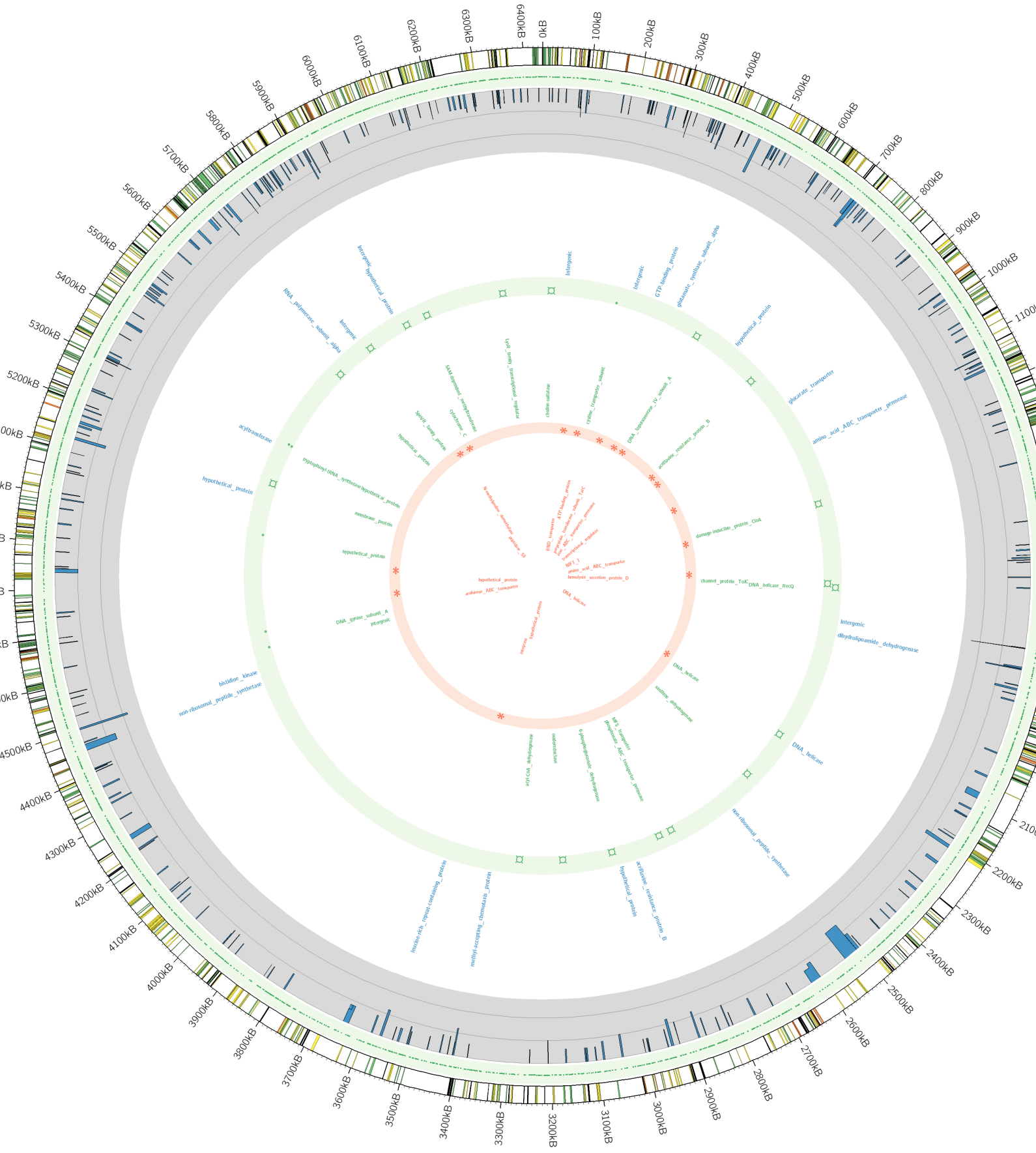


Figure 3.18: **Synthetic description of mutations acquired by the genome of *P. fluorescens* Pf0-1 during *in vitro* evolutionary experiments in Glycerol media.**

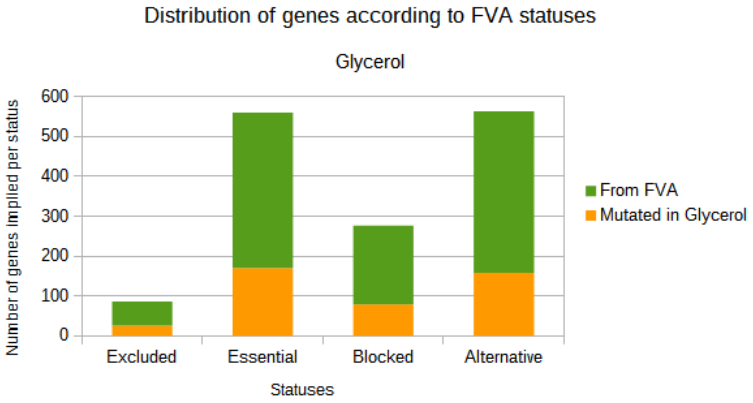


Figure 3.19: **Distribution of mutations according to FVA categories.**

metabolic pathways (without counting genes implied in multiple pathways). The FVA modelling based on Glycerol model had attributed 26 of these mutated genes the excluded status, 170 were essentials, 79 blocked and 433 alternative (Figure 3.19). The annotation of 288 genes is described as 'hypothetical protein' giving few information on the functions implied.

Gene_ID	Gene function	Metabolic pathway	Cog categories
PFL01_RS02385	carbamoyltransferase	#N/A	Post-translational modification, protein turnover, and chaperones
PFL01_RS03805	transcriptional regulator	#N/A	Transcription
PFL01_RS01160	ATP-binding protein	#N/A	Function unknown
PFL01_RS20785	arabinose ABC transporter substrate-binding protein	#N/A	Carbohydrate transport and metabolism
PFL01_RS09815	DNA helicase	#N/A	#N/A
PFL01_RS21560	hypothetical protein	#N/A	#N/A
PFL01_RS15320	hypothetical protein	#N/A	Function unknown
PFL01_RS00750	RND transporter	#N/A	Cell wall/membrane/envelope biogenesis
PFL01_RS26305	N-methylproline demethylase	#N/A	Energy production and conversion
PFL01_RS01920	preprotein translocase subunit TatC	#N/A	Intracellular trafficking, secretion, and vesicular transport
PFL01_RS07370	hemolysin secretion protein D	#N/A	Intracellular trafficking, secretion, and vesicular transport
PFL01_RS02670	iron ABC transporter permease	#N/A	Inorganic ion transport and metabolism
PFL01_RS05045	NIF3 1	#N/A	Function unknown
PFL01_RS26000	peptidase S9	#N/A	Amino acid transport and metabolism
PFL01_RS06305	amino acid ABC transporter substrate-binding protein	#N/A	Amino acid transport and metabolism
PFL01_RS04075	glycine/betaine ABC transporter permease	#N/A	Amino acid transport and metabolism
PFL01_RS15390	integrase	#N/A	Replication, recombination and repair

Table 3.4: **Description of stop-codon mutations found after Glycerol-*in vitro* evolutionary experiment.**

The COG categorization of altered genes is sustained in the Krona Figure 3.20.

Genes accumulating mutations

If most of the genes mutated (1314) carry only one mutation, a total of 693 genes (and 77 intergenic regions) were found to carry more than one mutation. The count of mutation per modified gene is described in the Figure 3.21. For glycerol, the maximum number of mutation carried by a (population's) gene is 13.

There is a significant effect (p-value = 4.449e-07) of the affiliation of genes to metabolic category or not on the propensity to mutate, and here also metabolic genes are more prone to mutate than non-metabolic genes (Figure 3.22).

Stop-Codons A total of 17 stop-codon mutations were retrieved. They are all on genes, but none are metabolic genes. Table 3.4 characterizes these stop-codon mutations.

Frequency wise, most of the mutations stay stable and below 5%, but a few of them present distinct patterns (Figure 3.23).

For example the mutation occurring on the gene involved in carbamoyltransferase (red line) production starts at really low frequency, increases abruptly before returning to being almost nonexistent. The salmon-colored curve (Figure 3.23) represent a stop-codon mutation (ATP-binding protein) of which the frequency experience an important drop on the fourth sampling time. While on the opposite the blue line (a transcriptional regulator) present a sudden increase on the fourth sampling time.

Unfortunately as no gene is associated to a metabolic pathway, it is not possible to link previous modeling work to stop-codons retrieved after Glycerol

Figure 3.20: **Distribution of mutations according to COG categories.** The figure represents the COG functional categories. The size of the category correspond to the number of genes included in this category and the percentage is the portion of COG it represents. The mutated portions represent the numbers of genes that were found to carry one or more mutation after the experiment -and that could be matched to a COG functional category. As only *genes* mutated and not *mutations number* is represented per category, further information such as synonymous/non-synonymous mutation or frequency patterns could not be inserted

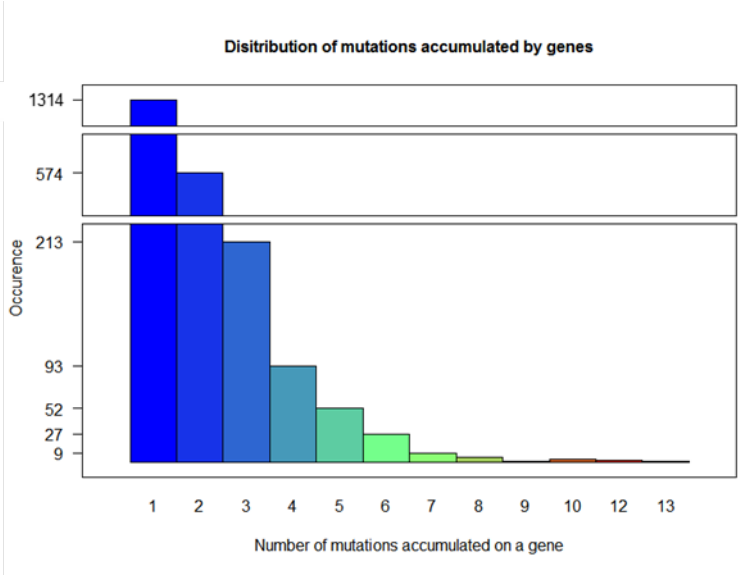


Figure 3.21: **Occurrence of genes accumulating several mutations.** Few genes support a high number of mutations (right of the graphic), while the vast majority of genes altered support one to two mutations (left of the graphic.)

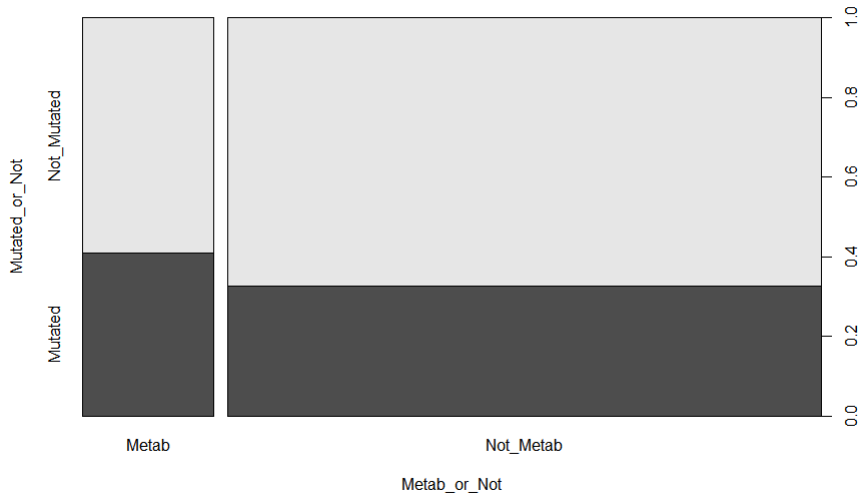


Figure 3.22: **Distribution of mutations according to the affiliation of genes to Metabolic pathways.** Metabolic genes (Metab) appear more prone to mutate than non-metabolic genes (Not-Metab) even though the proportion of non-metabolic genes is way higher than than the proportion of metabolic genes.

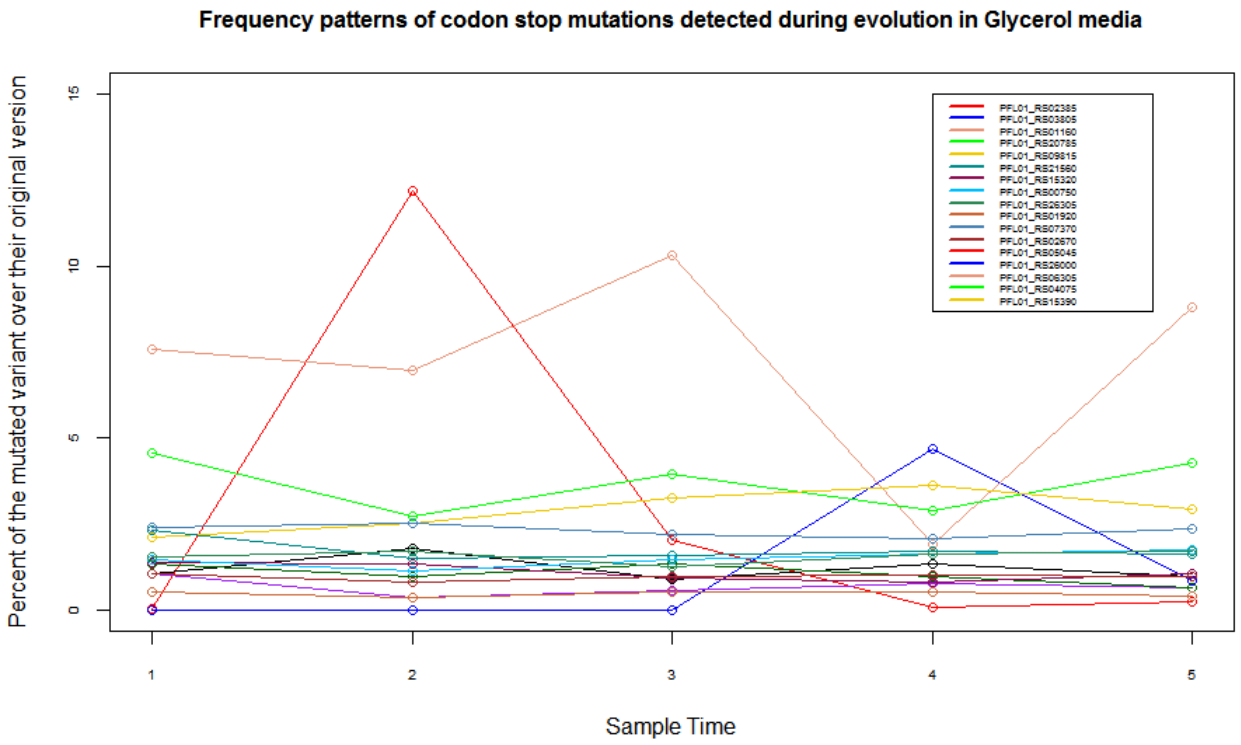


Figure 3.23: **Frequency dynamic of stop-codons retrieved in Glycerol analysis.**

experiment.

Genes with frequency increasing mutations

Of interest are also mutations which increase in frequency, as they can be considered as mutation being selected through fitness increase and traduce specialization.

From the Glycerol experiment, a total of 1856 mutations (45%) were described as being (even slightly) at a higher frequency in the last sample than in the first, but with punctual decrease at in-between sample times. The ratio of 'spreading' mutations is thus similar in Glucose and Glycerol (Figure 3.24).

None of these mutations correspond to stop-codons. Seven (7) of these genes supporting frequency increasing mutations are defined as metabolic genes and, through modelling, 3 were classified as Essential, 2 as Alternative and 2 as blocked. The rest of the genes was not considered as metabolic and are categorized in various COG categories as described in Table 3.5.

Alternatively we were expecting a rise of selection of genes implied in lipase production, and 3 genes directly implied in lipase production exclusively were detected. The genes affected carry between one and three mutations. Additionnaly, 15 more genes also implied in lipase metabolism (e.g. membrane, trans-

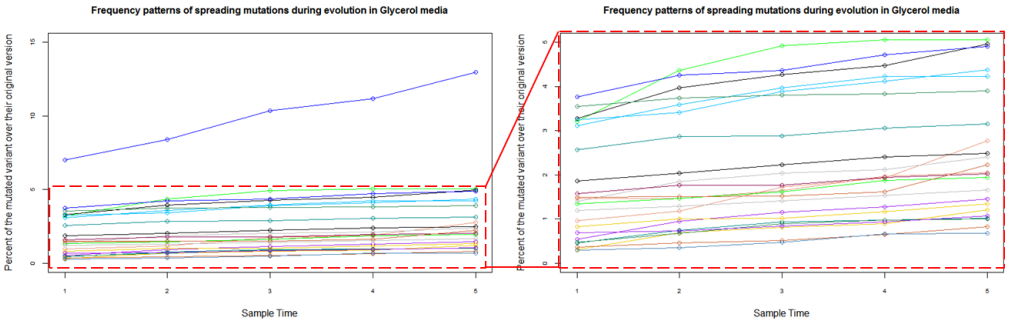


Figure 3.24: **Frequency patterns of the 25 spreading mutations (Glycerol).** The frequency presented is the occurrence of the mutated variant over the initial variant.

Gene_ID	Fq_increase	Function	FVA status	Metabolic Pathway	COG_categories
PFL01_RS12275	5,94	phosphonate ABC transporter permease			Inorganic ion transport and metabolism
PFL01_RS07380	4,85	channel protein TolC			Cell wall/membrane/envelope biogenesis
PFL01_RS13040	1,84	6-phosphogluconate dehydrogenase	alternative	Multiple	Carbohydrate transport and metabolism
PFL01_RS02535	1,81	DNA topoisomerase IV subunit A	essential		Replication, recombination and repair
PFL01_RS10630	1,68	xanthine dehydrogenase			Nucleotide transport and metabolism
PFL01_RS07405	1,14	ATP-dependent DNA helicase RecQ			Replication, recombination and repair
PFL01_RS09815	1,12	DNA helicase	alternative		
PFL01_RS03580	1,12	acriflavine resistance protein B			Defense mechanisms
PFL01_RS28295	1	LysR family transcriptional regulator			Transcription
PFL01_RS23600	0,91	hypothetical protein			Function unknown
PFL01_RS14450	0,9	acyl-CoA dehydrogenase		Multiple	Lipid transport and metabolism
PFL01_RS05930	0,74	damage-inducible protein CinA	blocked	Nicotinate and nicotinamide metabolism	Coenzyme transport and metabolism
PFL01_RS22135	0,63	hypothetical protein			Function unknown
PFL01_RS13750	0,59	oxidoreductase			Function unknown
PFL01_RS26955	0,59	SAM-dependent methyltransferase			Function unknown
PFL01_RS12025	0,58	MFS transporter			Carbohydrate transport and metabolism
PFL01_RS25155	0,56	hypothetical protein			Function unknown
PFL01_RS26540	0,52	cytochrome C			Energy production and conversion
PFL01_RS00140	0,52	choline-sulfatase	blocked		Inorganic ion transport and metabolism
PFL01_RS22990	0,47	membrane protein			Function unknown
PFL01_RS20470	0,46	DNA gyrase subunit A	essential		Replication, recombination and repair
PFL01_RS01235	0,44	cystine transporter subunit			Amino acid transport and metabolism
PFL01_RS25895	0,38	SpoVR family protein			Function unknown
intergenic	0,38	Intergenic			
PFL01_RS23605	0,35	tryptophanyl-tRNA synthetase	essential	Aminoacyl-tRNA biosynthesis	Translation, ribosomal structure and biogenesis

Table 3.5: **Description of genes carrying frequency increasing mutations (Glycerol).**

porters, lysophospholypase), but not uniquely, were also found. As lipases are excreted in the environment, this information is interesting: additionally to direct 'unused' genes modification, their diffusion system may be affected too.

Comparison between Glucose and Glycerol

As for the metabolic modelling, the results observe for the analysis of mutations in Glucose and in Glycerol follow the same trends: There is a high number of mutations (3767 for Glucose and 4091 for Glycerol) statistically not evenly distributed all over the population's genome in both experiments. In both cases the same number of genes are affected by mutations (2007 and 2008 respectively for Glucose and Glycerol) and 70% of the genes carrying mutations are common to both experiments. The patterns where some genes accumulate many mutations, while other genomic regions are conserved is also common to both populations. The ratio of synonymous mutations over non synonymous mutation is close for both analysis, and the characterization of mutations (apparition of stop-codons and of mutations that increase in frequency) present similarities.

Differences lay in the identity of genes being affected by either stop-codons or mutation that increase in frequency, without necessarily the capacity to always relate these differences to the actual evolutionary constraints. The 30% of the genes that are different between the two experiments are in both cases distributed in all the COG categories, with no particular difference noticeable. Interestingly, fewer stop-codons and frequency increasing mutations are characterized for the glycerol, despite the higher number of mutations detected.

3.4.3 Comparison with metabolic prediction

Comparison based on statuses

Classically FVA are used to test for effects of mutations on fitness variation (e.g: Papp *et.al.*, 2011). Here we explored the potential of FVA to predict the effects of environmental variation on genes sustainability.

Based on our rationale, we expected genes encoding for enzymes implied in reactions whose status changed from the null model to the constrained model to be the one the most changed. Especially for status that changed to Excluded, and *vice-versa*.

In the Glucose analysis, 80 genes of the 180 targeted genes were actually carrying mutation. For the Glycerol analysis 78 of the 180 targeted genes were carried mutation

These ratios correspond to 44% of the genes predicted. Thus targeted genes appear more prone to mutate than others genes (of which only 37% are mutated). In parallel, it was mentioned earlier (3.2.2, genes accumulating mutations) that metabolic genes were more prone to mutate than not metabolic genes. While the second statement assesses the importance of a metabolic approach to study the evolution of specialization, the first statement confirms that genes predicted to be modified (through metabolic modeling) have a higher propensity to carry mutations than others. These results suggest that the metabolic approach is a key element to forecast more accurately the emergence of mutations in the context of adaptative evolution.

Comparison based on CV2s

Glucose Altogether, 439 out of the 2533 mutated genes are directly considered as metabolic genes (approx. 17 %). As a recall the SEED metabolic model of *P. fluorescens* Pf0-1 presented 1062 metabolic genes. Thus 41.3 % of the metabolic genes are carrying at least one mutation.

Within the *P. fluorescens* population, the ratio of mutated genes within the Essential-genes, Alternatives-genes and Excluded-genes categories were of similar magnitude (i.e. 42.9%, 39.5% and 46.6% respectively). According to our hypotheses, we expected more mutations in the excluded-genes category. The results present such a tendency but the differences are thin. However we also observe a higher rate of mutation for the essential category. This evoke the possibility that acquired mutations in these genes enhance their activity. However, when focusing on the *reactions* entailed by the mutated genes retrieved, the pattern is different, and the ratio of reactions entailed by mutated genes is lower in the excluded category than in the alternative category. Their distribution according to CV2 is represented in Figure 3.25. This means that our predictions may hold to a certain degree at the genetic level, but not necessarily at the reactions level. The difference can potentially be explained by the attribution rules of genes to reactions (Material & Method) and the issue of genes being implied in several reactions.

We would have expected that more mutated reaction would accumulate in the excluded category, but slightly more mutated reactions are observed in the 'alternative' gene category.

Also when focusing on the category of genes for which the status changed between the unconstrained model and the constrained model, 42% of the genes are mutated.

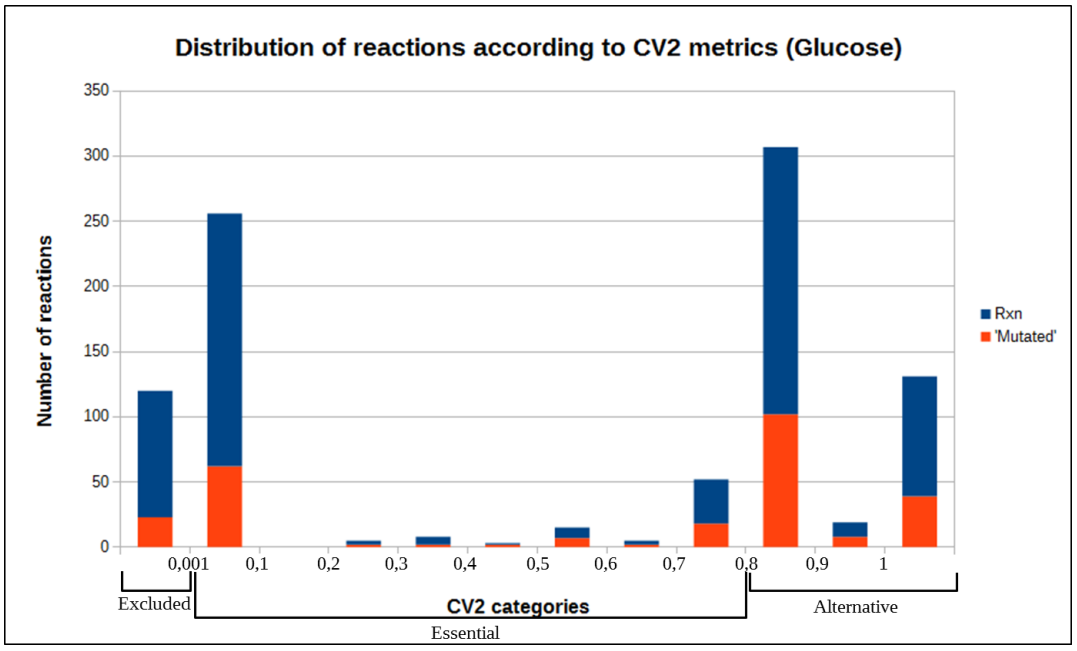


Figure 3.25: **Ratio of the number of reactions encoded by genes that carried mutations (orange) over the overall number of reactions (green)**. The distribution along the X axis is according to the CV2 metrics of the reactions. The distribution of mutated genes amongst total genes was associated to the reactions determined by the genes

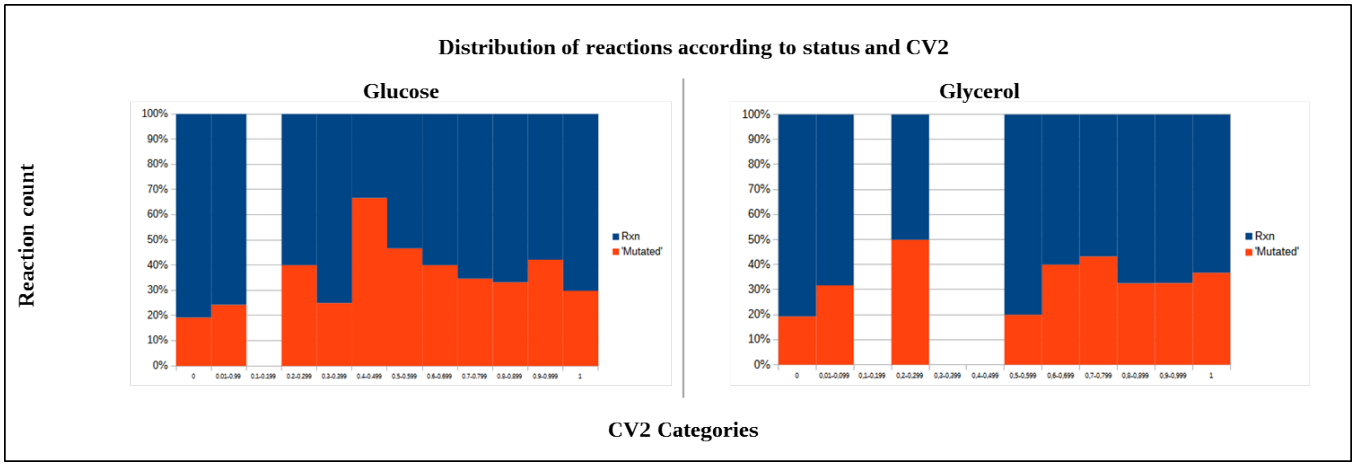


Figure 3.26: **Comparison of the distribution of reactions (mutated over all) according to CV2 metrics between glucose and glycerol FBA models.**

3.5 Conclusion

The population's genome of *P. fluorescens* was found to carry a relatively high number of mutations, and 1/3 of the genes (all genomes confounded) were carrying at least one mutation.

From over 3000 mutations detected, only about 40 mutations present a pattern (increase of frequency of the mutant over the original variant over evolutionary time) that could directly traduce an increase in fitness. Replacing these raw results in known evolutionary theories, we could attribute the conservation and frequency increase pattern of the 40 mutations to natural selection: because they confer fitness benefits, these mutations will be selected for.

On the other hand, the vast majority of the mutations ($> 99\%$) have no particular patterns, and could thus be related to Kimura's Neutral Theory (see section 1.1.4, of Part I), where most of the mutations have no particular effects on fitness and are thus conserved randomly. Therefore, a clear adaptative effect can be attribute to less than 1% of the mutations detected.

In our predictive approach, the modeled metabolisms were realized under the assumption of a maximal growth function, which is a proxy of the fitness. Therefore predictions were made for adaptative genes modifications, which are obscured by outnumbering circum-neutral mutations.

Additionally as mentioned above, the metabolic genes represent a relatively small portion of the genomes (less than 20% for *P. fluorescens*) thus more than 80% of the genes belong to various other functional categories. As specialization can also occur through more processes than metabolism (modifications of transporters for nutrients uptake, modifications of life history traits such as the size or the divison rate, of mobility, of behavior, etc) it appears obvious that focusing only on metabolic genes will give limited elements of prediction. Every other function that can be altered by the change of environment but which is not described as purely metabolic (i.e. which is not implied in the production of an enzyme) will not be predicted. This is the case for functions such as motility or chemotaxis, for which the genes appeared to be frequently modified after the specialization experiments in Glucose.

Considering these results, it appears that a more advanced analysis of the mutations is necessary to improve the connection between the modelling predictions and the evolutionary outcomes. Indeed, we started by focusing on mu-

tations location and accumulation on genes, to verify if genes 'targeted' via the metabolic modelling were more prone to carry mutations than others. Now we can also take into consideration (i) the mutation effects (i.e. synonymous, non synonymous) and (ii) the location of the mutations onto promoting regions of genes or regulatory sequences of operons, as this could highly influence the genes expression. Finally, to explore further these predictive aspects the modelling could be re-examined under different assumptions, such as with conditions where nutrients are limited, or with other constraints than the ones relying on carbon sources, such as the presence of a usually excreted metabolism (e.g. siderophores, which was envisaged but extremely pricey).

3.6 Supplementary material

[The four following pages are extracted from W. Delage masters' thesis]

II.3 Methods

Each script has been developed in Python 2.7 and used Biopython Package (http://www.biopython.org). The wokflow is composed of two successive parts : one for DNA analysis and one for RNA analysis (Fig 3).

II.3.1 Automation of DNA Analysis

II.3.1.1 Database creation

The first step of the workflow for DNA analysis is to parse and format the gbff file which contains genome information in a fasta format. Unique identifier used for each genes contains: “locus_tag” (new id of the gene), ”old_locus_tag” (old id of the gene, before update), ”db_xref”(GI number assigned to protein), ”protein_id”(protein id used by gene and protein database) and “product”(function of protein). It is possible that a SNP is located in an intergenic region. A script was used to find intergenic region by identifying regions that do not correspond to genes from gbff file. Then genes and intergenic region are concatenated in one files (Genome.fasta) and this file is used to make a database with Blast++.

II.3.1.2 Blast

The second step is to blast on the database with os command (automated by import os module on Python). The output is an xml file. Then this xml file is parsed with a Python script to extract the Query_id (DiscoSNP id), the Query_sequence (DiscoSNP sequence), the Subject_id (Gene id from Blast++ database) and the Subject_sequence (Gene sequence from Blast++ database) in an output file named files.txt. Subject_id and Subject_sequence are genes id and genes sequences that matched with our query. Because Query_sequences can be located in repeated region or duplicated genes, files.txt is filtered in order to remove each query which has more than one hit matching with the database. Indeed, with these Query we can't determine where the mutation have appeared or if the SNP discovered is a false positive (same repeat region with SNP not induced by the experiment).

II.3.1.3 Detection of mutations

The third step is to recreate the mutant gene (Reconstruction_gene.py, fig3). We use Subject_sequence and Subject_id to retrieve full gene sequence in our Genome.fasta and replace

our Subject_sequence in our full gene sequence by the Query_sequence. The output contains DiscoSNP id, gene id and the mutant sequence.

The fourth step is to identify each mutation in mutant sequence by comparing mutant genes to wild-type genes (Gene_to_prot.py, fig 3). Because we are working with bacteria, we have to take into account the specific translate table for bacteria (named translate table 11, for more information see http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11). Indeed, bacteria have not one but up to seven start codon.

For each substitution we identified impact of the mutation, using several characteristics (Table 1).

Characteristics	**Example**
Codon (mutated or and his counterpart non mutated)	GTC (non mutated codon) GGC (mutated codon)
Amino acid that encodes for the codon extracted	Valine (non mutated amino acid) Glycine (mutated amino acid)
Amino acid location of the codon extracted	421/514
Physicochemical properties of the amino acid (mutated and non-mutated)	Hydrophobic / Aliphatic Hydrophobic / Aliphatic
Impact of mutation (silence or not silent)	Not silent

Table 1. Characteristics used to identify impact of mutations.

Codon and amino acids allow to the user to see if stop codon appears. Physicochemical properties inform on the plausible structural changes of the protein induced by the mutation. Mutation location help to know if mutation can have an important impact on protein functionality.

For indel (insertion or deletion) mutation we identified the location of mutation, the impact of the first codon mutated and if a stop codon prematurely appeared, the size of wild type amino acid sequence, the size of mutant amino acid and the size of amino acid mutant if a stop codon appeared prematurely. For mutation in intergenic region, only location and nucleotide of wild type and mutant were identified.

The fifth step is to eliminate mutation that have not been induced by the evolutionary experiments (filter2.py, fig 3). These mutations were detected because the strains of pf0-1 used already differs from the reference genome available. We compared every mutation that was detected for a DiscoSNP couple to the reference genome, and kept only the mutation that appeared once in the couple. If there was more than one mutation kept for one DiscoSNP id, we did not keep the mutant in our output file named protein.txt.

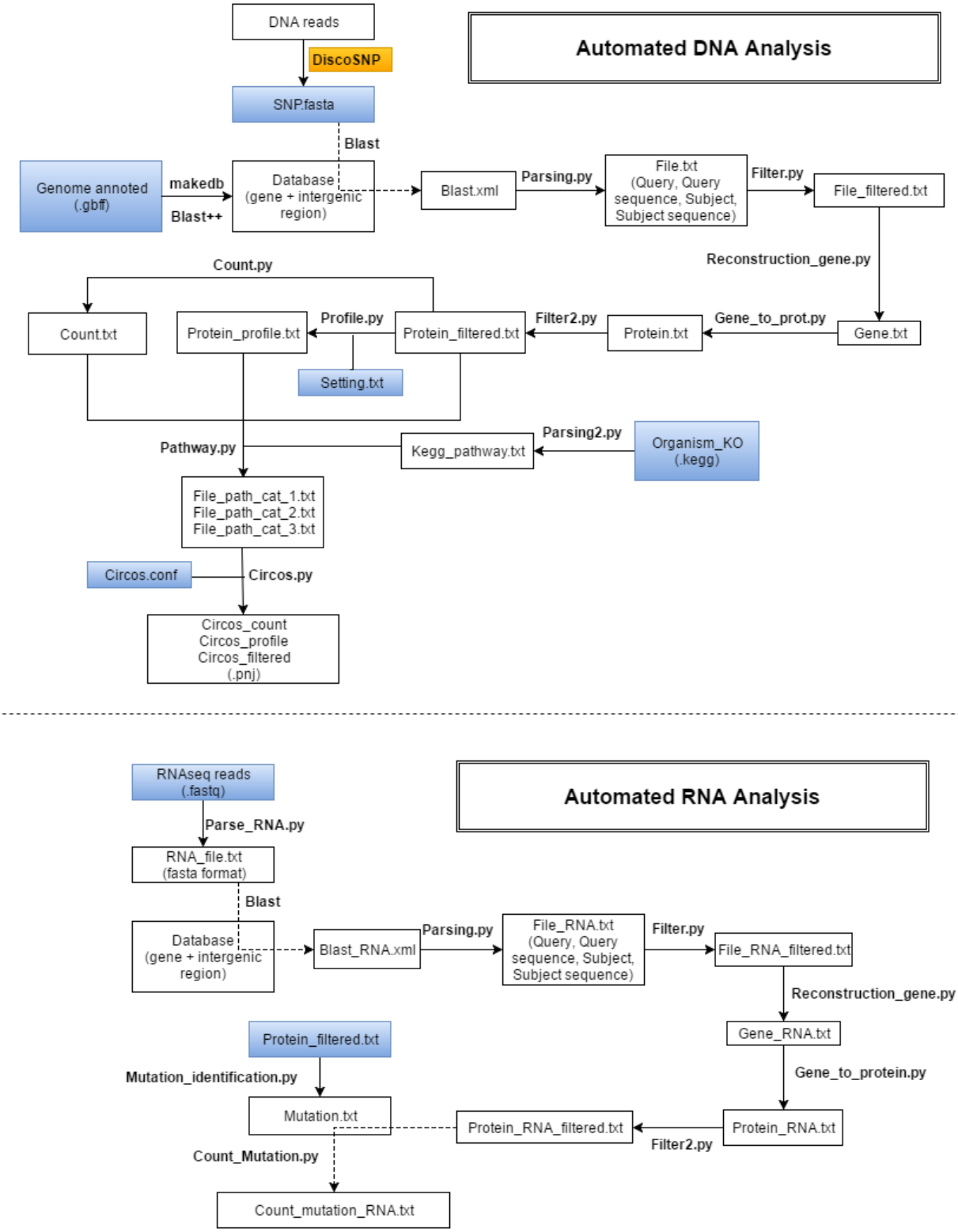


Figure 3 : RNA and DNA analysis workflow. In blue the input files required in order to run the workflow. In Yellow, DiscoSNP that the workflow doesn't run.

The sixth step is to count how many mutations appear in genes which allow to see if some genes accumulate more mutation than others genes (Count.py, fig 3).

II.3.1.4 Filtering by profile

The seventh step allows to select specific mutation by their coverage profile (Profile.py, fig 3). One can select mutations that increase or decrease in frequency between each bacteria sample, or compare only two times points (e.g: coverage C1 of sample 1 vs coverage C3 of sample 3) or even apply a threshold on coverage to eliminate undesirable data.

II.3.1.5 Pathway identification and graphic representation

The eighth step is to format KEGG files in order to get pathway for each genes (file pathway.txt, script Parsing.py, fig 3). Then we identify each pathway for each mutated genes from protein_filtered.txt, protein.txt or Count.txt by connecting these outputs to pathway.txt (Pathway.py, fig 3). Three outputs are transformed to fit the Circos input format.

The first is an output which contains a general category name about the organism pathway (File_path_cat_1.txt, fig 3). We parsed the file pathway.txt and counted which pathway contained more information and genes. The second is an output made on our data about the ratio of the first category (File_path_cat_2.txt, fig 3). The third is about the sub-category and the fourth is about the sub-sub-category (File_path_cat_3.txt, fig 3). Then an os command is used to automate graphic representation with Circos using Circos.conf as setting for the graphic representation.

In order to validate the work flow developed during the internship, we compared results obtained manually with results obtained via the work flow.

II.3.2 Automated RNA Analysis

RNA was extracted from the same samples as for DNA. In order to identify and quantify the impact of mutations on gene expression from RNA reads we proceeded as follow (fig 3):

We extracted unique id and sequences (120pb) for each read and blasted them against our database used for the DNA analysis part (see above). We filtered and proceeded to the same steps 2 and 3 as for the DNA analysis. Step 4 was slightly modified: each mutation in the mutant sequence was identified by comparing mutant genes and wild-type genes. We identified the non-mutated codon, the amino acid, the amino acid location, the physicochemical properties of non

Chapter 4

Phenotype expression via Microfluidic Experiments

Combination of genetic studies and functional assays to characterize the evolution of *Pseudomonas fluorescens'* motile behavior in a context of specialization.

[1]Alix Mas, [1]Philippe Vandenkoornhuyse, [1]Yvan Lagadeuc & [2]Pietro de Anna

1 Université de Rennes 1, ECOBIO UMR 6553.

2 Université de Lausanne, ISTE.

This work was supported by the Université de Bretagne Loire (UBL) through an outgoing mobility grant awarded in 2016.

4.1 Introduction

A defining characteristic of bacteria is their ability to be motile. Motility is the capacity for self-propelled motion in a liquid medium (motion on a (quasi)solid medium is termed gliding and requires other mechanisms). The motility behavior in bacteria was discovered 300 years ago (Mitchell and Kogure, 2006), yet its evolutionary emergence is still not well understood (Faguy and Jarrell, 1999). Motility is common to many bacteria and mostly Gram-positive bacteria. The majority of motile bacteria move by the use of flagella, which are tail-like structures protruding from the cell and functioning as rotary propellers (Jarrell and McBride, 2008). Depending on species, bacteria can have different types of flagella: a single flagellum at one pole or at each poles of the cell, or numerous flagella, either organized as a tuft at one end of the cell or all over the cell surface.

The motile behavior is mostly used for nutrients foraging (Mitchell and Kogure, 2006), through the usage of chemotaxis. Chemotaxis is the ability of organisms to sense gradients (of nutrients, chemicals, or any compounds) in their environment and adjust their motile behavior accordingly.

Motility is known as being energetically costly (Wei and Bauer, 1998) and a common analogy is that swimming in water for bacteria would be the equivalent of swimming in honey for humans (Mann & Lazier, 1991). Under the assumption that nutrient foraging is the main reason why bacteria will express motility behavior, we can hypothesize that motility functions will be reduced if bacteria are subjected to a very homogeneous environment where nutriments are continuously accessible everywhere. In the frame of evolution, if such an environment is stable long enough for the population to specialize to it, we can expect that this need for reduction in motility (energy savings) would be integrated to the genome through the alteration of genes, impairing definitively the motility behavior.

4.1.1 Preamble on bacterial motility

As described in Jarrell and McBride (2008), the displacement of bacteria is not linear : The direction of rotation of the flagella determines the movement of the cell and periodically the direction of rotation is briefly reversed, causing what is known as a “tumble”, which results in the reorientation of the cell. When anticlockwise rotation is resumed, the cell moves off in a new direction. This allows bacteria to change direction often. The current range of speeds for

swimming in fluid is 1–1000 µms/s (Marwan et al., 1991; Fenchel and Thar, 2004).

Bacteria can sense nutrients and toxic molecules and move respectively towards or away from them – this process is known as Chemotaxis. This behavior is achieved by changes in the frequency of tumbles. For example, a decreased frequency of tumbling in response to an attractant gradient resulted in migration up this gradient (Mitchell and Kogure, 2006). By consequence, when moving towards a favorable stimulus the frequency of tumbles is lower than when swimming towards an unfavorable stimulus (and reversely), allowing the cell to reorient itself and move to more suitable environments. It is noteworthy to mention that at the genetic level the expression of genes encoding flagellar proteins is highly regulated (Jarrell and McBride, 2008).

4.1.2 System studied and objectives

In order to verify that the motility behavior is modified when bacteria specialize to a stable and homogeneous medium with nutrients being constantly available, tests of functional traits related to motility were performed on a generalist population of bacteria and on a *specialized* population of the same bacteria.

To obtain two related population (a generalist and a specialized one), an *in vitro* evolutionary experiment was performed, allowing for the specialization of the initial population of *Pseudomonas fluorescens* Pf0-1, to Glucose (being the unique source of carbon in the environment) for 140 generations [1]. Samples of the initial and of the evolved population were archived through cryogenization, and could be revived at will.

Subsequent genetic analyses of the initial and of the evolved population were realized, and several mutations were detected to have emerge on genes implied in the motility and chemotaxis of bacteria.

The motility behavior compared between the two populations are characterized by the 1) proportion of individuals actively moving in the population, 2) the velocity of moving individuals, and 3) the chemotaxis capacities of the populations prior and post specialization..

Microfluidic devices and time-lapse video-microscopy allowed to analyze and compare the 2 populations as it enables the detection of the position and displacement of individual bacteria.

Relying on the hypothesis that in a stable and homogeneous environment,

[1] The *in vitro* evolutionary experiment is presented in details in the 3^d chapter of this section.

evolutionary-wise if a function is costly and that its cost is not offset by advantageous effects, the function should be lost (Lahti et al., 2009) (or unexpressed), one of our primary hypothesis was that, as the experimental environment was very homogeneous and stable nutrient wise, the energy-demanding motility behavior should be reduced as the need for nutrient foraging is offset by the ubiquity of the glucose in the environment.

4.2 Material and Methods

4.2.1 Measures of pH and viscosity of carbon solutions

The motility of bacteria can be influenced by physico-chemical parameters such as 1) the pH of the solution they are in, and 2) the viscosity of these solutions.

In many flagellar system the rotation of the flagellum is driven by gradients of protons ions across the membrane (Atsumi et al, 1992; Manson et al, 1977) and thus will influence in which way the flagella is activated (Jarrell and McBride, 2008). The pH (concentration of H+) can modulate the motile capacity of organisms (Minamino et al., 2003). The viscosity of the solution can also modify the swimming behavior, the more viscous a solution is, the harder it is to move in it. In order to ensure that observations of motility behavior of the samples in different carbon sources was not influenced by either pH or viscosity, these parameters were measured for some of the solutions.

pH: The pH of the Glucose and Glycerol solutions was measured with a classic pH probe, 2 times for each solution.
Viscosity: The viscosity of Glucose, Glycerol and Acetate solutions were measured with a rheometer. As the detection threshold for viscosity of the rheometer was too high to measure the viscosity of the solutions at the concentration of carbon sources used for our experiments, the solutions were concentrated 10 folds for the viscosity measures.

The table for viscosity (Table 4.1) and pH below shows that no noticeable difference exist in viscosity between the solutions and that these parameters can be set aside in the interpretation of the observations. The slight variations (2.37mPa/S to 2.56mPa/S) observed are negligible, especially when we recall that the solutions were ten times concentrated for the viscosity-tests than for the actual experiments. At the bacterial scale, such sensible variation should not affect the velocity.

Carbon Source	pH	Viscosity (10e-3 Pa/S)
Glucose	6,85	2,37
Glycerol	6,91	2,38
Fructose		#
Lactose		#
Galactose		#
Acetate		2,56
Fumarate		#
Succinate		2,51
Xylose		#

Table 4.1: **pH and viscosity for tested solutions**

pH for Glucose (6.85) and Glycerol (6.91) are very similar, and a difference in behavior between these two solution would not be explained by a difference of pH of the media. The pH of the other solution was not assessed and could vary slightly as for example compounds such as acetate are known to acidify the medium.

4.2.2 Microfluidic devices

Until a few decades ago various strategies existed to appraise bacterial motility (capilarity in tubes, colonization on agar plates, counting bacteria in initially sterile medium with chemo-attractant (Adler, 1966; Meyer et al, 2002; Kearns et al, 2001). Since less than a decade (Binz et al., 2010), microfluidics has appeared as a convenient tool to quantify and characterize motility of bacteria. Even though it is not yet the most spread one (the equipment is still being very costly), microfluidic instrumentation to study motility and chemotaxis is one of the most accurate one as it measures motility *in vivo*, in direct, at the individual scale and in a limited volume of liquid (Ahmed et al., 2010; Rusconi et al., 2014).

Device used in Motility Assays

Polydimethylsiloxane (PDMS, Sylgard 184 Dow Corning, MI, USA) microfluidics were designed using an in-house software generating 12 parallel straight channels of width w = 1 mm and length L = 40 mm. The generated geometry is printed in chrome onto transparent glass with high resolution (JD-photodata). Micro-channels were fabricated by prototyping against a silicon master with positive relief features using standard soft lithography techniques. The PDMS layer was patterned with one channel that is 0.1 mm deep. PDMS channel was ob-

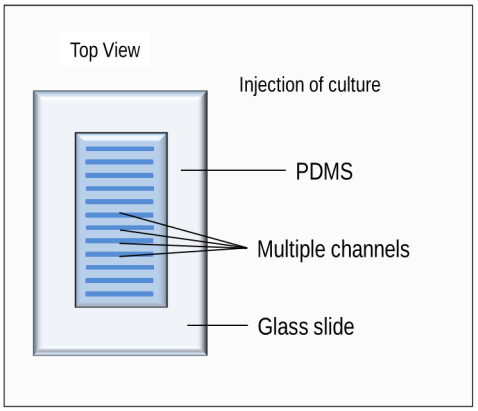


Figure 4.1: **Microfluidic devices used to test the motility behavior of bacteria**

tained by molding against the silicon master, baking at 65 °C for 6 h, peeling off the hardened PDMS, cutting it to size, and punching inlets and outlets apertures and sticking it to a soda-lime glass slide (size of 75 mm x 25 mm x 1 mm) (Figure 4.1).

Having several parallel channels permitted to couple data acquisition of each sample in the various carbon sources to maximize the homogeneity of conditions and enable inter-sources comparisons.

Device used in Chemotaxis Assays

The gradient generator used is made of 3 layers (Figure 4.4): a polydimethylsiloxane (PDMS, Sylgard 184 Dow Corning, MI, USA) layer on top (impermeable to solutes), an agarose-gel layer in the middle (permeable to the dissolved carbon sources), and a glass slide at the bottom (for support). PDMS cuts were created through the same protocole as above but with only two irrigation channels carved this time. The channel hosting the microbial populations is carved in the agarose layer. The position of the PDMS layer is adjusted to ensure that the channel carved in agarose are located in the middle (at equal distance) of the two PDMS channels. The whole system is put under adjusted press in order to ensure a perfect adhesion of each layer, without squishing the agarose membrane.

4.2.3 Culture preparation for microfluidic experiments

The samples mentioned are issued from the *in vitro* evolutionary experiments presented in Chapter 3. Each experiment described below has been repeated in 5 independent replicates.

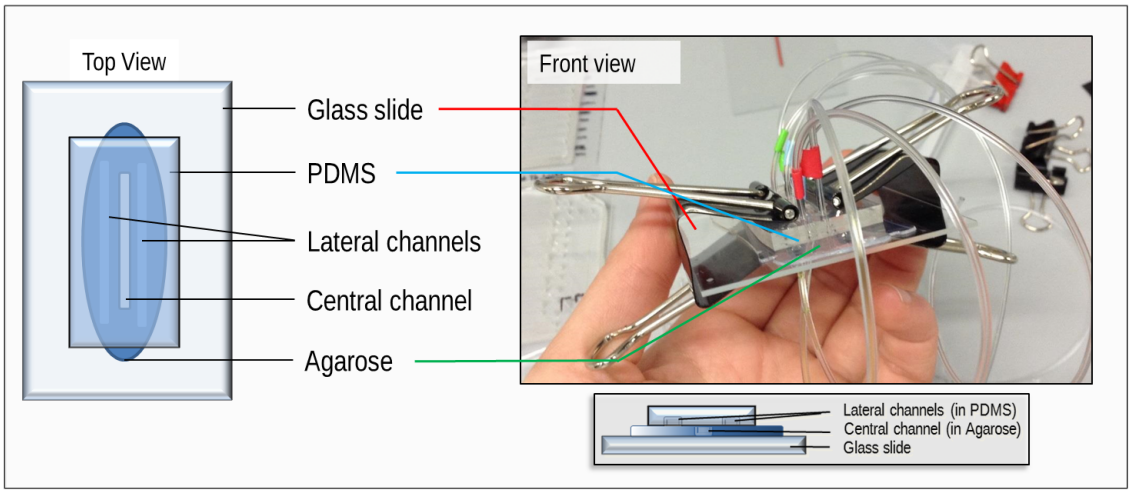


Figure 4.2: **Microfluidic device allowing for the creation of a nutrient gradient to test for chemotaxis**

Culture preparation for Motility Assay

Samples from a glycerol stocks of the initial population and of the population evolved in Glucose were revived and incubated overnight in LB at 29°C, 60 rpm, for 21 hours. Nine carbon sources were tested: Glucose, Glycerol, Acetate, Succinate, Lactose, Galactose, Fructose, Fumarate and Xylose. Before performing the microfluidic experiments, for each carbon source tested (concentration of 1,2 g of Carbon-equivalent per Liter) 1 mL of the overnight cultures was re-suspended in 7 ml of LB solutions supplemented with one of each of the 9 carbon sources. The fresh mixes were put back in the incubator for 10-15 min to prevent modification due to thermal conditions. Then, the 9 suspensions were successively injected in the channels through micro-pipettes, and the apertures were taped to avoid evaporation during data acquisition.

Culture preparation for Chemotaxis Assays

Samples from glycerol stocks of the initial population and of the population evolved in Glucose were revived and incubated overnight in LB at 29°C, 60 rpm, for 24 hours. We tested the chemotactic response of the 2 microbial populations (initial and mutated population) to a steady, spatially homogeneous, gradient of a solution of Glucose PMM (Pseudomonas Minimal Medium).

Before performing the microfluidic experiments, the optical density of the population samples was measured through a spectrophotometer at wavelength 600 nm to estimate the population size. Once the DO measured, 100 microL of the overnight cultures were re-suspended in 900 µL of pseudomonas mini-

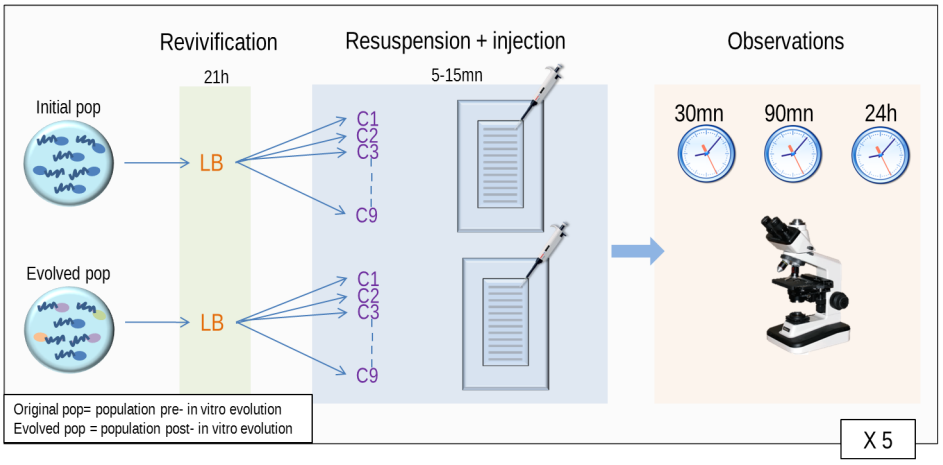


Figure 4.3: **Protocol to perform the motility assays.** Cryogenized samples of the initial and evolved population are revived overnight in LB (Lyseogeni Broth) medium. Samples are then re-suspended in different carbon-source solutions before being injected in the microfluidic devices where short set of pictures are realized for each samples. Each experiment is repeated independently 5 times.

mal media without carbon source (called neutral-PMM afterwards). The fresh mixes were put back in the incubator for 1-2 hours to acclimate to conditions of the experimental runs (Figure 4.4). Meanwhile, the fluxes of neutral PMM and Glucose-PMM were started and active for at least 30 minutes to soak the agarose membrane; the flows of neutral and carbon solution stay active all along the experimental run. Then for each experimental run, a bacterial sample is pumped within the mid-channel carved in agarose, and the flow of culture is stopped.

4.2.4 Microscopy, image acquisition and processing

Workflow for Motility Assays

All samples were observed immediately after injection of the cultures in the micro-channels, and also at 30 minutes and 90 minutes. The observations were made at mid-depth of the channels for motility assays and at the bottom of the channel for biofilm assays. To track the position of the bacteria between constant time intervals, we set the microscope (a fully automated Nikon Ti-E) and the camera (NIKON QI-2) to capture 30 images every second for 10 second at the selected location, saving gray scale images of 8-bit pixel depth. We perform the experiments with 40X magnification that allow for detection of suspended microbes position. Gray scale pictures are taken in phase contrast using a phase ring much larger (Ph3) than the one needed for the optics used (Ph1), mimicking a dark field imaging, so that the image background is dark and bacteria are

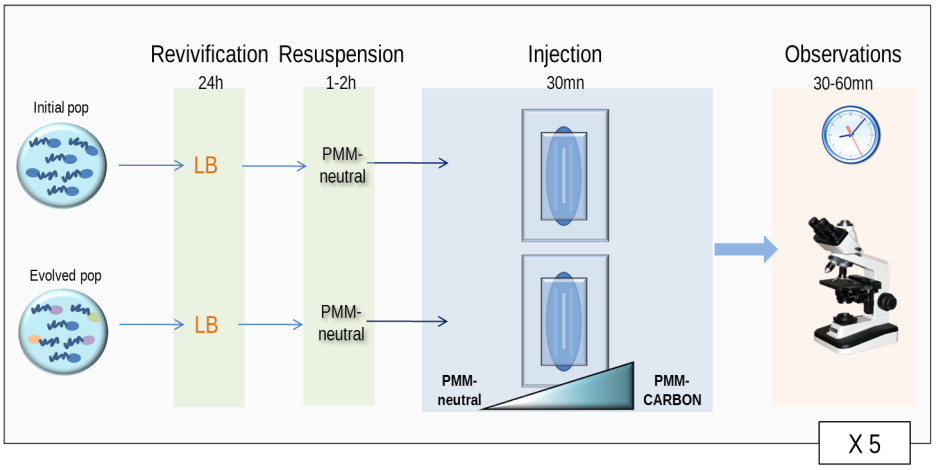


Figure 4.4: **Protocol to perform the chemotaxis assays.** Cryogenized samples of the initial and evolved population are revived overnight in LB (Lysogeny Bprth) medium. Samples are then re-suspended in neutral medium (no carbon source added) to initiate starvation and accustom bacteria to the experimental medium before being injected in the microfluidic devices. A set of pictures spanning from the benginning up to 60 minutes are realized for each samples. Each experiment is repeated independently 5 times.

detected as bright objects.

Then, collected images are processed on MATLAB. The trajectory of each microbe is considered like a time series of positions (from an image to the next). The position of each bacteria is detected at each picture and from the image analysis (supplementary) we obtain the trajectories of bacteria detected in motion (Figure 4.5). The number of bacteria mentioned is not the total number of bacteria visible along the time series of an experiment, but the number of detected, moving, bacteria. From these acquired trajectories it is possible to determine the velocity and tumble times associated to each moving bacteria.

For a complete and detailed presentation of image processing, please refer to the supplementary material.

Wicrofluidic workflow for Chemotaxis Assays

The observations were made in the channel carved in the agarose membrane which contains bacterial population within a gradient of carbon media.

The carbon solution (Glucose-PMM) in one of the two irrigation channels and the buffer solution (neutral PMM) in the other channel were continuously flown: relying on the molecular diffusion of the carbon solution through agarose, this device produces a steady, linear carbon concentration profile within the underlying agarose membrane and into the microbial suspension channel. After

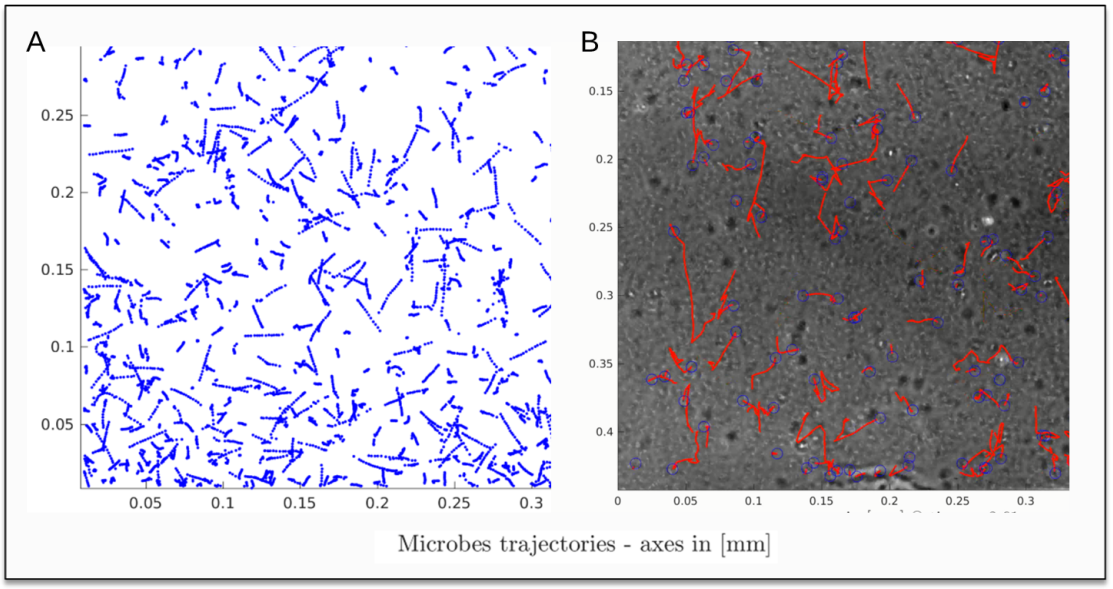


Figure 4.5: **Steps in the determination of bacteria trajectories**
A Each single blue point correspond to the localization of a bacterium on one image; here the figure resumes the successive localization taken by bacteria, as extracted from several successive images. **B** The blue circles localize the bacteria, the red lines correspond to the trajectories of the bacteria based on the successive blue points-trajectories determined in A.

a time sufficient to ensure the establishment of the steady, homogeneous gradient in the culture (at least 30 minutes), pictures were taken during 30 to 60 minutes without delay between images. The magnification was 15X, enabling the picture of the entire channel's width.

The characterization of chemotaxis is classically done by measuring the spatial accumulation of bacteria along the gradient of the chemo-attractant crossing the channel in its width, as represented in Figure 4.6. Once again collected images are processed on MATLAB. A set of the 100th last images before 30 minutes are selected and concatenated, and the mean spatial distribution of bacteria in the channel is characterized.

The microbes spatial distribution is expected to equal the stationary solution of the advection-diffusion equation where the advection is not due to fluid motion, but to chemotactic migration. The characteristic exponent is given by the chemotactic velocity divided by the cells diffusivity D which was measured in the tracking experiment.

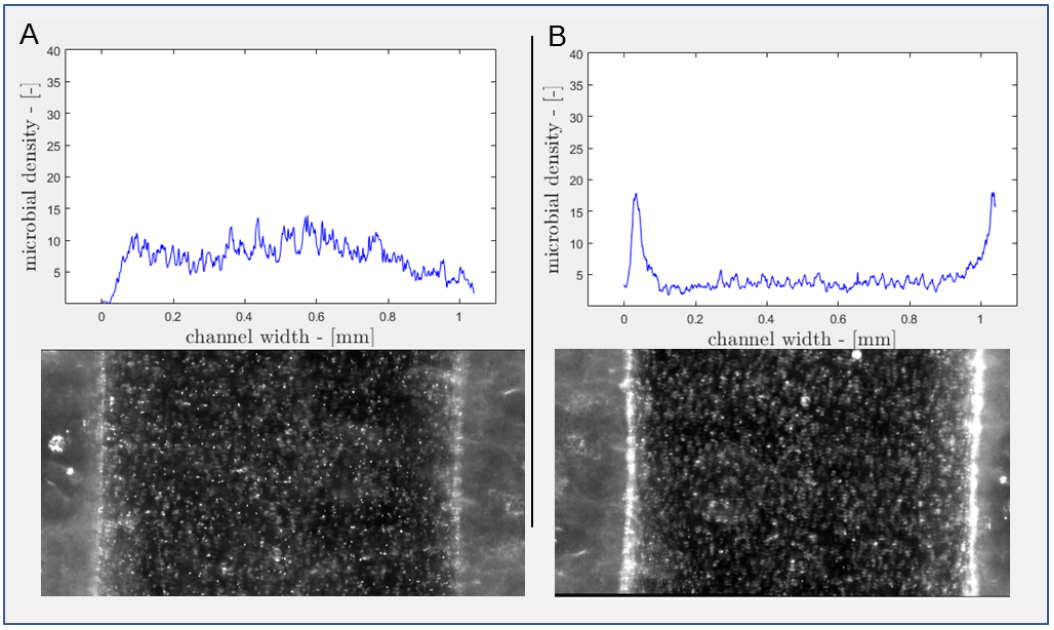


Figure 4.6: **Calcul of the concentration of bacteria across channel width** Top of the figure: Microbial density calculated over channel width; the limits of the X axis correspond to the limits of the channel of the picture on the bottom of the figure; **A** measures at injection time and **B** after 30 minutes in the channel, the bright white dots are bacteria, we can see how their accumulation is represented by the peaks on the graph above.

4.2.5 Statistical Analysis for motility assays

For a finer investigation, complementary statistics analysis were ran for both ratio of swimming bacteria and velocity of swimming bacteria, in order to test whether the evolved population presented a significantly different behavior than the initial population under 9 different Carbon Sources.

Ratio of swimming bacteria over non-swimming bacteria: An ANOVA analysis was performed using a linear mixed model procedure in R 3.1.358 with packages 'nlme'59 and 'car'60 in order to test the effect of Population (2 modalities: initial-evolved) and Carbon Source (9 modalities) on the variance of swimming bacteria ratio observed between experimental runs. The treatments (Populations and Carbon Source modalities) were considered as fixed factors while the replicates (5) were considered as a random factor.

Mean velocity of swimming bacteria: The velocity of each particle tracked per experimental run was retrieved and an ANOVA analysis was performed using the procedure and packages described above. When a significant effect of treatment was detected by the ANOVA, post-hoc contrast tests were performed using the 'Tukey's test' to test for significant differences between modalities.

4.3 Results and Discussion

4.3.1 Mutations Characterized

From the evolutionary experiments and genetic analysis detailed in Chapter 3, 45 genes implied in flagellar synthesize and chemotaxis were found to carry mutations after *in vitro* evolution in Glucose-medium. It represents 44% of the total genes implied in motility/chemotaxis in the genome. The 45 genes were affected by a total of 79 mutations (Table 4.2, Supplementary Material). Additionally, twelve genes implied in pilus formation also carry 23 mutations. Some pilus are known to be active in twitching motility (Sampedro et al., 2014; Burrows, 2012; Piepenbrink and Sundberg, 2016). Therefore a total of 102 mutations implied in the motility functions emerged during the evolutionary experiment. Non of these mutation lead to the apparition of a stop-codon (which would block the genes activity), yet 85% of the mutations are non-synonymous, suggesting that they have potential effects on protein functions. The frequency patterns of these mutations are mostly stable, i.e. they don't vary much along experimental time. Only one mutation (flagellar biosynthesis protein FlhB) increases substantially in frequency over time (from 0 to 12%) (Table 4.2).

Knowing that from our experiment, overall 34% of the genes carried at least one mutation (even if synonymous), and that here 44% of the genes implied in motility/chemotaxis carry mutations, this later category seems more prone to mutate than at random. This result comforts our hypothesis that the motility behavior will be modified when organism specialize to a given environment.

4.3.2 Results on Motility tests

Ratio of swimming bacteria

The ANOVA statistic test, run on the mixed linear model to assess the effect of population (initial population or evolved population) and Carbon Source on the ratio of swimming bacteria per population, permits to assess that there is a significant (p-value=0.0004, F-value =14.210) effect of the population type on the ratio of bacteria swimming, but no significant effect of the Carbon Source tested (p-value=0.352, F-value =1.141). The evolved population has a higher ratio of swimming bacteria, all carbon sources confounded. The evolved population can be assumed to have more active bacteria than the initial population.

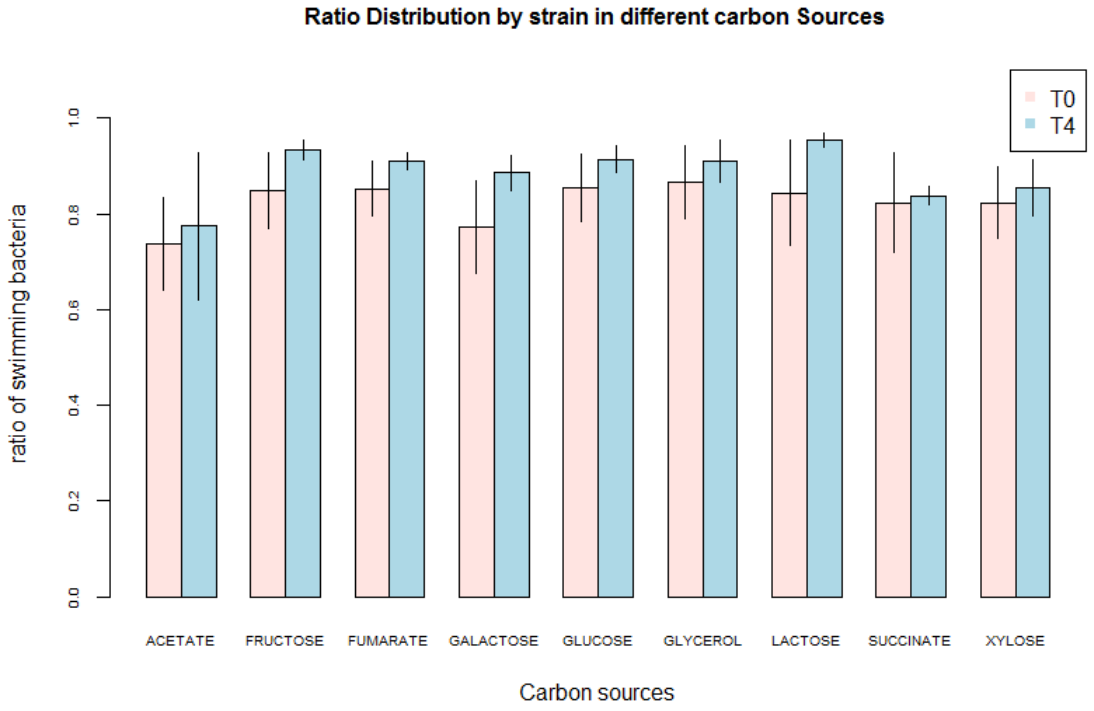


Figure 4.7: **Ratios of swimming bacteria according to their population of origin and in various carbon sources**. Pink is the initial population (T0 in the legend), blue is the evolved population (T4 in the legend). The bar error represents standard deviation. For all carbon sources confounded, the evolved population has a higher ratio of moving bacteria than the initial population and this is explained by the populations' type (p-value=0.0004).

Velocity of swimming bacteria

The ANOVA statistic test, run on the mixed linear model to assess the effect of population (initial population or evolved population) and carbon sources on the velocity of swimming bacteria, permits to assess that the population type (F-value =401.4124), the carbon source (F-value =26.2410) and the interaction of both variable (F-value =6.1181) have significant effects (for all modalities: p-value=<.0001) on the velocity of swimming bacteria.

For each carbon source tested, the velocity of bacteria is significantly different between the evolved population and the initial population (for all combination : Tukey's test p-value=<.0001).

The same difference pattern between initial and evolved pattern is observable in each source, the velocity increase expressed by the evolved population is of the order of 1 µm/s.

The differences in motility observed between carbon sources clearly indicate that the dispersion behavior of *P. fluorescens* Pf0-1 is related to the evolution of the population, and to the available C-source.

Carbon sources added in the neutral medium might impact the physico-chemical characteristics (such as pH, viscosity) but did not correlate with a modification of viscosity (see M & M section).

Alternatively, these differences observed could be explained by different metabolic capacities of the cells (uptake and use) when confronted to different carbohydrates molecules. This hypothesis has to be tested further through complementary experiments.

Mitchell and Kogure (2006) mentioned that swimming speed would decrease with an increasing nutrient concentration. We can interpret this by taking the problem from a different angle: if we consider the evolved population as more efficient in uptaking the glucose present, then it may deplete the environment from its glucose faster, leading it to increase its motility.

Interestingly, the differences observed are quasi-common to all carbon source, and not uniquely to Glucose as would have been expected in the case of a punctual acclimation to the media. The modification expressed by the evolved population is thus mostly intrinsic to the population itself, and not dependable on the medium it is subjected to only. This comforts our hypothesis of specialization to the experimental evolutionary medium.

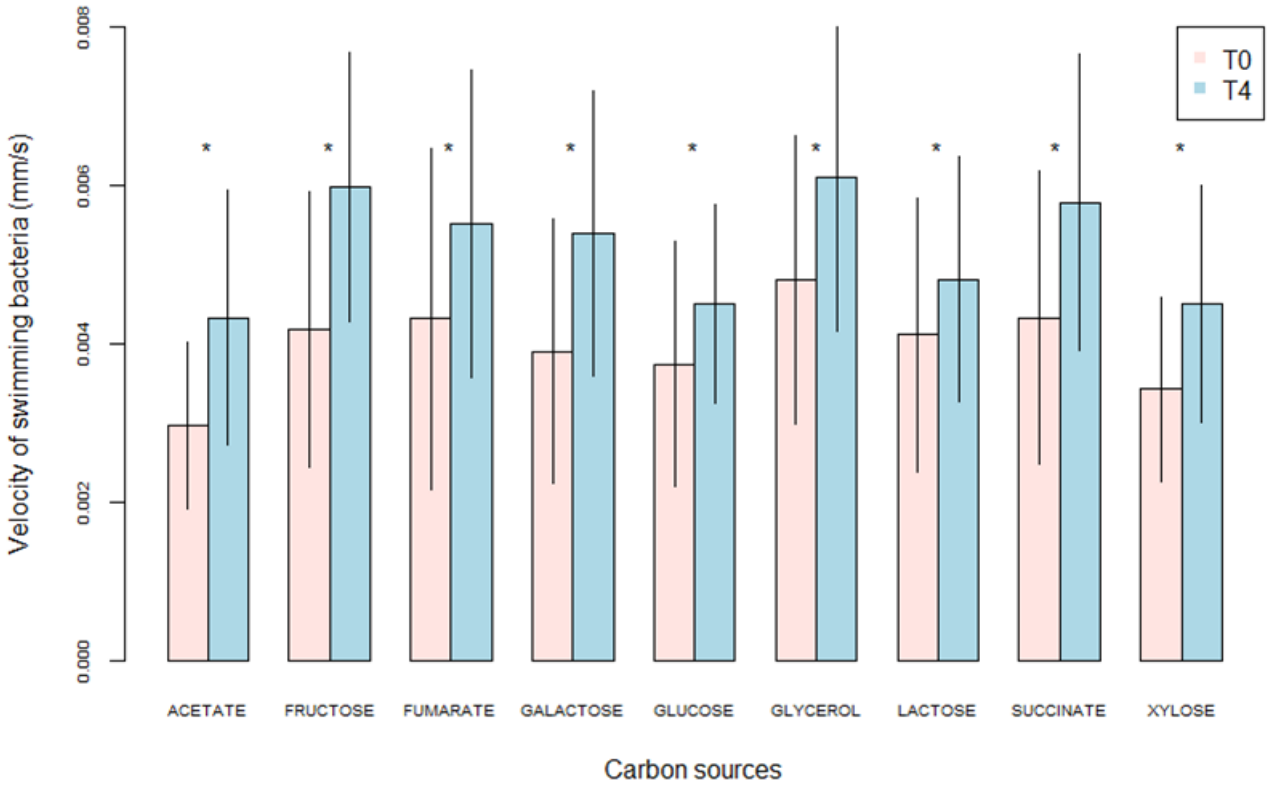


Figure 4.8: **Velocity of swimming bacteria according to their population of origin.** Pink is the initial population, blue is the evolved population. The bar error represents standard deviation. The stars means significant difference between the population tested. For each carbon source tested, we can see that the bacteria in the evolved population (T4 in the legend) swim significantly faster than the initial population (T0 in the legend). This difference between the two population is significantly (p-value=<.0001) explained by the populations' origin.

4.3.3 Results on Chemotaxis tests

By studying chemotaxis of the initial and evolved population in a Glucose media, we want to determine if the evolved individuals developed an enhanced capacity to detect and/or find its way towards Glucose. The rationale behind this hypothesis is that if the evolved population actually specialized to the Glucose present in its growth media, it should be more efficient at exploiting it.

To do so, we observed the chemotactic response of the to population by subjecting them to a gradient of Glucose.

As presented in Figure 4.9, which represents before/after chemotaxis of the same population, we can already notice that individuals tend to accumulate on both surfaces of the channel borders. This accumulation on the sides can be due to both physical and chemical reasons.

It is most probably due to the fact that fresh media with trace elements (and potentially more oxygen) is provided on each side. This makes the detection of the accumulation of bacteria due to the nutrient gradients harder to define, as there is always an accumulation of bacteria on the brim of the channels at some point in every experimental run (Figure 4.6).

Preliminary results displayed on Figure 4.10 show the distribution of bacteria on the width of the channel after 30 minutes in the gradient for the five replicates of each population (initial and evolved). There is a high variance between the 5 replicates in each population, and it is even more compelling for the initial population where there are both very flat profiles (bacteria are distributed all over the channel) and highly skewed profile (accumulation of bacteria on the left side of the channel).

It is also apparent that there is a difference between the initial population and the evolved population. The initial population when it perform chemotaxis, seems to mostly accumulate towards the left edge of the channels (glucose), with a small but still consistent accumulation on the right side too (no glucose). On the other hand, the profile of the evolved population are distinctly skewed toward the right side (no glucose). The slope is much less abrupt, meaning that the distribution of the bacteria in the channel is more dispersed. The difference in distribution observed indicates a difference in the behavior of the the populations.

The chemotaxis profile are determined after 30 minutes in the microfluidic. The choice of this timing was based on previous chemotaxis assays (e.g. Sampedro et al. (2014); Rico-Jim?nez et al. (2016)).

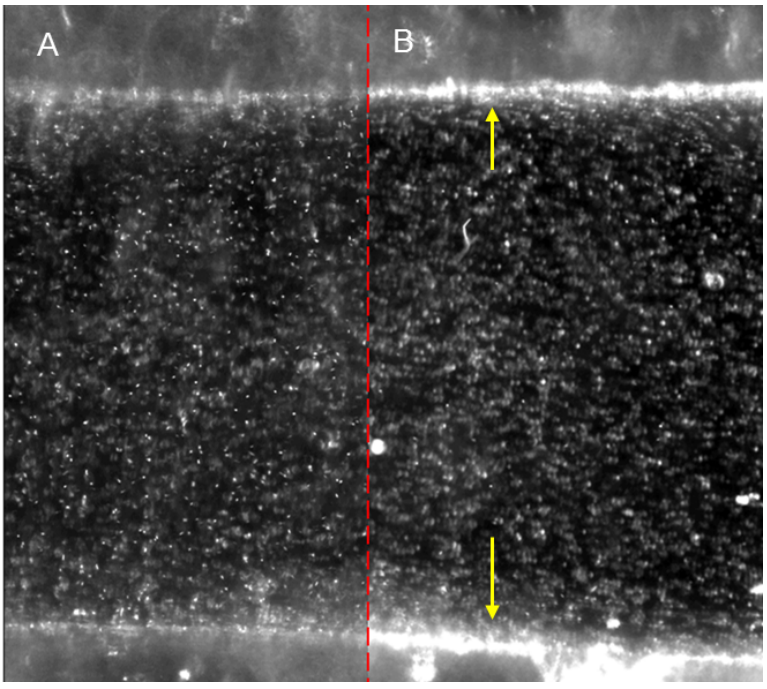


Figure 4.9: **Picture of the channel containing bacteria, before and after chemotaxis**. The top of the channel is where the concentration in glucose is higher, the bottom of the channel is where the concentration is low. On the left side of the picture (A), the bacteria were just injected in the channel, and their distribution is still rather homogeneous. On the right side of the picture (B) which is more than 30 minutes after injection, it is noticeable that the borders of the channel are much brighter than for A, because the bacteria have accumulated there.

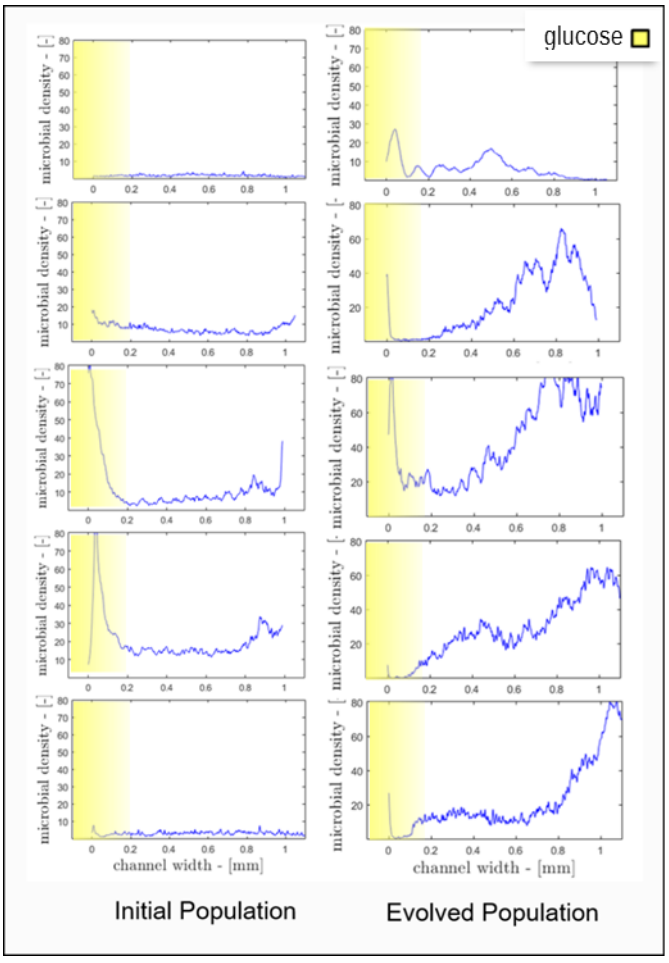


Figure 4.10: **Distribution profiles of bacteria resulting from chemotaxis behavior**. On the left of the figure are the 5 replicates of bacteria distribution for the inital population over the channel width. On the right side are the five replicates for the Evolved population. Relative to the X axis of each image, zero is were the glucose-PMM filters by, and one is were the neutral PMM filters by. Peaks of bacterial density traduces an active accumulation of bacteria through chemotaxis.

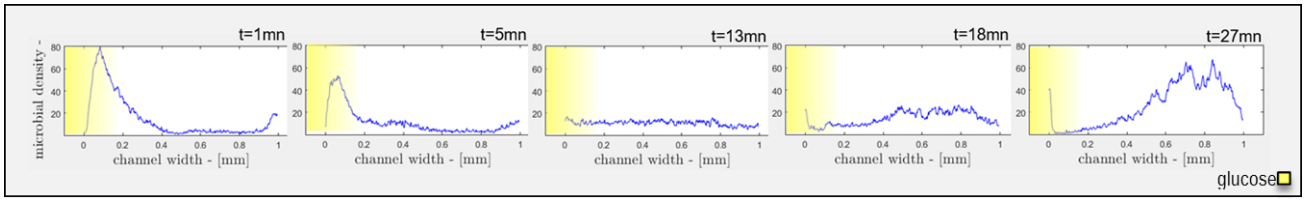


Figure 4.11: **Chemotaxis dynamics of the evolved population of *P. fluorescens* duringthe first 30min of the experiment**. It is noticeable that during the first ten minutes there is a high concentration of bacteria toward the side liberating Glucose, and this concentration slowly diminishes when the bacteria start moving to the other side of the channel (where the neutral medium filters by).

Nevertheless, and most interestingly, when observing the distribution patterns of the evolved population of bacteria within the first 30 minutes of the experiments, there is always, during the first 5 to 8 minutes a very strong chemotaxis toward glucose, before the population start slowly switching side (as can be seen in the example sequence Figure 4.11).

A possible explanation to this dynamic is that, first bacteria very rapidly detect and move toward the side where glucose is, and then start changing side within the first 10 minutes.

Because a high density of bacteria may have covered the surface where Glucose arrives from, the access to the nutrient for other bacteria is limited. It can thus be hypothesized that the bacteria need to forage elsewhere in the channel.

An alternative explanation might be that the bacteria stocked up on glucose, and roam around to balance the stoechiometry of other needed nutrients in order be able to metabolize the stored glucose. Indeed bacteria, amongst which *Pseudomonad*, have the capacity to store carbon, and more specifically glucose under the form of PHB (Polyhydroxybutyrate) (Wilkinson, 1963; Yu, 2001). The carbon is stored under this form when other nutrients such as oxygen or azote, necessary for the metabolism, are limiting (Yu, 2001).

Finally, another (complemetary) explanation could be that the glucose utilization during the first 5 minutes may have led to the production of metabolites in the environment which could act as chemo-repellent to present bacteria.

Thus when studying chemotaxis with microfluidics, it is also interesting to study the dynamic of the chemotactic response, instead of the results of the chemotactic response at a given time-point, as it may help characterize better the behavior of bacteria toward nutrients (or else) gradient.

4.4 Conclusion & perspectives

The results presented herein clearly highlight the differences in the motile behavior between the population that evolved in glucose in comparison to the initial unevolved population. The bacteria from the evolved population have a more developed motility than their initial counterparts, as more bacteria are active, and the active bacteria also swim faster. These patterns are not only true in glucose medium, which is the medium in which bacteria evolved, but they are also observable for all the carbohydrates tested (sugars or else). Thus we can infer that the behavior modification observed are not only punctual acclimation through gene regulation, but more likely characteristics acquired by the evolved population.

These modifications can be put in parallel with the numerous mutations (104) found on 44% of the genes implied in flagella formation and chemotaxis.

Not mentioned here but presented in chapter 3 on the genetic analysis of the two population, high level of genes implied in membrane transport were also affected by mutations. These membrane transport changes, coupled with the chemotaxis and flagellar modification could also explain the difference in chemotaxis observed.

Indeed, the results of the chemotaxis analysis presented herein suggest that the evolved population has a stronger chemotaxis (and more efficient use) toward glucose.

Further analyses enabling a stronger characterization of the differences between the two population will be performed. The chemotaxis velocities of *P. fluorescens* will be defined, after having determined their diffusivity in the chemotaxis environment. The temporal dynamic of the chemotaxis behavior will also be analyzed in more details to specify the duration of the chemotaxis and its intensity.

This work stirred up many question, and opened a door on several ideas of experiments and analyses. For example, we observed that the the evolved population had a very different type of biofilm formation when observing the surface of the channels 24 hours later (Figure 4.12), which we had not necessarily predicted.

Actually recent research on Pseudomonas found that mutations affecting flagellar regulation also influenced the formation of biofilms (Mastropaolo et al., 2012). It was determined that single genes are implied in both motility and biofilm formation (Mastropaolo et al., 2012), or both in adhesion and flagellar

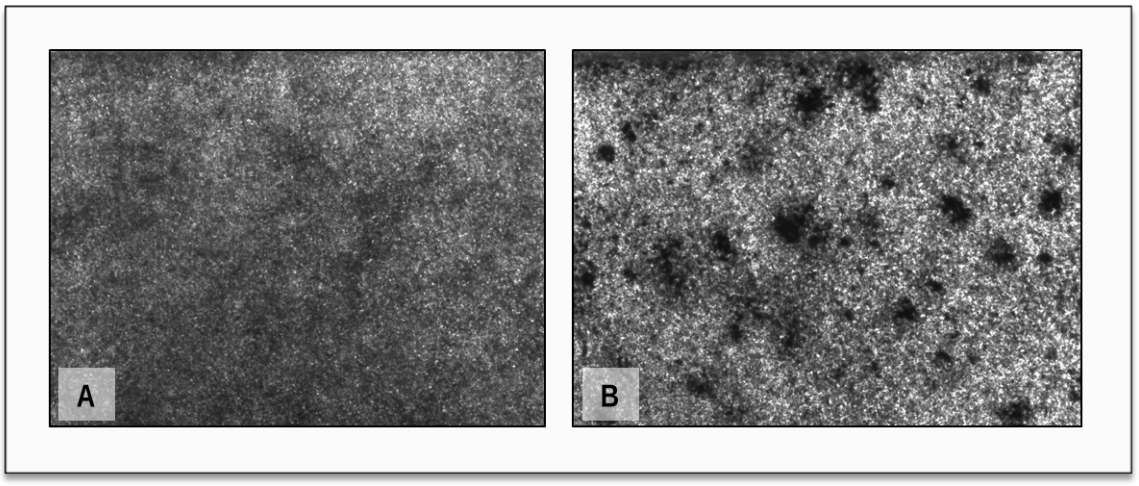


Figure 4.12: **Picture of the biofilm formed by the intial population (A) and evolved population (B)** The biofilm of the evolved population (B) appears more thick and more patchy than the biofilm of the initial population (A).

production (Casaz et al., 2001).

Spatial statistics to define biofilm formation in our experiment could comfort and illustrate these findings at the functional level.

It was also casually observed that the motility of bacteria was varying largely according to the replication phase they were going through. While this can influence the observation of motility of our two population if the growth rate was modified for the evolve population, it also give space for a detailed characterization of the evolution of motility along the growth of a bacterial population.

Finally, as motility is a costly behavior, and because the evolutionary experiment was realized in an homogeneous and stable environment, we expected a reduction in the motility of evolved bacteria. Yet the motile behavior was not impaired as we expected it to be after the evolutionary experiment even if clear modifications emerged from the specialization to the experimental environment.

The specialization has resulted in a more efficient motility, uptake of glucose, and different biofilm formation, which may be explained by an improvement of the trans-membrane import of the glucose and its utilization by the bacteria.

Additionally, the use of motility is not exclusively linked to chemotaxis but also to swim away from disadvantageous environments. In an environment where bacteria are thriving, the accumulation of individuals can lead to a higher rate of competition for the nutrients available. Therefor, an alternative hypothesis is that the motility of individuals in a constant environment is conserved or developped to escape competition and enhance nutrient uptake to optimize individual fitness.

4.5 Supplementary Information

4.5.1 Genes of the evolved population, implied in motility or chemotaxis and which carry at least one mutation

4.5.2 Notes on the statistical analysis of microbes trajectories

A-Trajectory statistics

The trajectory of each microbe ($'b'$ in the following equations) is a time series of positions: a position is a vector, an object composed by two elements describing the position with respect to the horizontal (x axis) and vertical (y axis) directions. The position of each microbe is detected at each step (picture) i and it is represented here by $(x_b(i), y_b(i))$. From the image analysis (described below) we obtain N_b trajectories of N_b microbes detected in motion. As explained below the number N_b is not the total number of bacteria visible along the time series of an experiment, but the number of detected, moving, bacteria. Every quantity defined based on trajectories will be called Lagrangian, in opposition to quantities defined on a reference grid which are said Eulerian.

The Lagrangian displacement and its curvilinear coordinate

A microbe trajectory can be thought as a curve in space (two dimensional space), as it should appear clear from the observation the movies we realized. We define the displacement $\vec{d}$ traveled by a microbe between two time steps as a vector of two components: the position variation in x, dx and the one in y, dy

$$\begin{aligned} dx_b(i) &= x_b(i) - x_b(i-1) \\ dy_b(i) &= y_b(i) - y_b(i-1) \end{aligned} \tag{4.1}$$

The distance s traveled between the two steps is thus the Euclidean distance traveled:

$$s_b(i) = \sqrt{dx_b(i)^2 + dy_b(i)^2} \tag{4.2}$$

Along each trajectory it is now defined the curvilinear coordinate s as the distance from the original point: s can only increase as time goes on, even if the microbe goes back and forth passing over the same location periodically.

Genes Mutated	Function	Number of mutation	Frequency Change
PFL01_RS26770	twitching motility protein PilT	1	0,19
PFL01_RS23395	methyl-accepting chemotaxis protein	3	-2,56
PFL01_RS18910	methyl-accepting chemotaxis protein	2	-1,96
PFL01_RS14950	methyl-accepting chemotaxis protein	1	1,64
PFL01_RS18920	methyl-accepting chemotaxis protein	1	0,48
PFL01_RS23685	methyl-accepting chemotaxis protein	1	1,08
PFL01_RS23905	methyl-accepting chemotaxis protein	2	-1,19
PFL01_RS03685	methyl-accepting chemotaxis protein	2	-0,99
PFL01_RS01910	methyl-accepting chemotaxis protein	2	-0,38
PFL01_RS16695	methyl-accepting chemotaxis protein	2	-0,46
PFL01_RS03140	methyl-accepting chemotaxis protein	1	0,74
PFL01_RS03995	methyl-accepting chemotaxis protein	1	-0,42
PFL01_RS21715	methyl-accepting chemotaxis protein	1	-0,44
PFL01_RS08215	methyl-accepting chemotaxis protein	1	-0,1
PFL01_RS07590	flagellar rod assembly protein FlgJ	1	-0,07
PFL01_RS07795	flagellar motor switch protein FliM	1	0,35
PFL01_RS07750	flagellar motor switch protein FliG	1	-0,23
PFL01_RS07905	flagellar motor protein MotD	1	1,05
PFL01_RS02585	flagellar motor protein MotB	1	-0,65
PFL01_RS07940	flagellar hook-length control protein FliK	3	0,65
PFL01_RS07785	flagellar hook-length control protein	1	1,01
PFL01_RS07865	flagellar biosynthesis regulator FlhF	1	3,48
PFL01_RS07825	flagellar biosynthesis protein FlhB	2	-0,26
PFL01_RS07570	flagellar basal body rod protein FlgF	2	0,1
PFL01_RS21355	flagellar basal body rod modification protein FlgD	1	-0,23
PFL01_RS27395	flagellar basal body protein FliL	1	-0,75
PFL01_RS00240	chemotaxis protein CheY	4	0,47
PFL01_RS21220	chemotaxis protein CheY	4	12,72
PFL01_RS13335	chemotaxis protein CheY	3	-0,75
PFL01_RS07775	chemotaxis protein CheY	2	-0,11
PFL01_RS24365	chemotaxis protein CheY	2	-0,32
PFL01_RS04145	chemotaxis protein CheY	2	1,7
PFL01_RS15420	chemotaxis protein CheY	1	-3,16
PFL01_RS21415	chemotaxis protein CheY	1	0,64
PFL01_RS05315	chemotaxis protein CheW	2	-0,94
PFL01_RS22040	chemotaxis protein	5	-1,14
PFL01_RS23950	chemotaxis protein	2	-0,38
PFL01_RS03315	chemotaxis protein	1	-0,14
PFL01_RS02270	chemotaxis protein	2	-0,6
PFL01_RS22285	chemotaxis protein	2	-0,71
PFL01_RS27750	chemotaxis protein	2	0,47
PFL01_RS02665	chemotaxis protein	2	0,66
PFL01_RS26695	chemotaxis protein	1	-0,22
PFL01_RS21070	chemotaxis protein	1	0,39
PFL01_RS09235	chemotaxis protein	2	-0,84

Table 4.2: **Genes of the evolved population implied in motility or chemotaxis and carrying mutations**

(Lagrangian) Average position

Since the fluid surrounding the microbes is at rest (there is no flow during the experiments), the average positions of the microbes is supposed to be constant due to their random motion (as much they move on a direction, as much they move on the opposite one). the average position at step i can be computed as

$$\begin{aligned} \mu_x(i) &= \frac{1}{N_b}\sum_{b=1}^{N_b} x_b(i) \\ \mu_y(i) &= \frac{1}{N_b}\sum_{b=1}^{N_b} y_b(i) \end{aligned} \tag{4.3}$$

Since the average position in this configuration does not have a particular physical meaning, it makes sense to simplify our analysis imposing the average position at $(0,0)$, by removing from each trajectory the initial position:

$$\begin{aligned} \mu_x(i) &= \frac{1}{N_b}\sum_{b=1}^{N_b} \big(x_b(i) - x_b(1)\big) \\ \mu_y(i) &= \frac{1}{N_b}\sum_{b=1}^{N_b} \big(y_b(i) - y_b(1)\big) \end{aligned} \tag{4.4}$$

The so-defined average position is expected to be zero (or smaller than the detection limit which is a bit below the μm).

The (Lagrangian) velocities

We define the velocity of a microbe along its own trajectory as the displacement (difference of position) between step i and the step $i-1$, divided by the time gap between the two steps. It is a vector, an object composed by two elements, one for the direction x and one for y:

$$\begin{aligned} vx_b(i) &= \frac{x_b(i) - x_b(i-1)}{t(i) - t(i-1)} \\ vy_b(i) &= \frac{y_b(i) - y_b(i-1)}{t(i) - t(i-1)} \\ v_b(i) &= \frac{\sqrt{dx_b(i)^2 + dy_b(i)^2}}{t(i) - t(i-1)} \end{aligned} \tag{4.5}$$

The latter corresponds to the temporal variation of the displacement s_b (curvilinear coordinate), defined above.

The average velocity along a microbe trajectory is the arithmetic average over all bacteria tracked in the image i:

$$\mu_{vx} = \frac{1}{N_b}\sum_{b=1}^{N_b} vx_b(i)$$
$$\mu_{vy} = \frac{1}{N_b}\sum_{b=1}^{N_b} vy_b(i)$$
$$\mu_v = \frac{1}{N_b}\sum_{b=1}^{N_b} \sqrt{vx_b(i)^2 + vx_b(i)^2} \tag{4.6}$$

A way to measure the time that a microbe spend moving with the same velocity along this trajectory, is the computation of the auto-correlation function of the Lagrangian velocities:

$$\chi(j) = \frac{1}{N_b}\sum_{i=1}^{N_b} \frac{(v_b(i+j) - \mu_v(i+j)) \cdot (v_b(i) - \mu_v(i)))}{\sqrt{\sigma_v(i+j) \cdot \sigma_v(j)}} \tag{4.7}$$

This is a function of the time lag (of time) between two position along a trajectory: as long as the velocity along the trajectory is the same (so it is correlated), χ stays close to 1. As time passes and the microbe along its trajectory change velocity and *looses memory of its past velocities*, $\chi(t)$ slows down to zero. Typically this slowing down is exponential in time $\chi(t) \sim e^{-t/\tau}$. the characteristic time for the exponential decay is τ, that can be quantified by fitting and exponential decay on the measured $\chi(t)$.

The swimming direction To measure the direction of the swimming displacement of a microbe b, we use the velocity vector (vx_b, vy_b): the absolute orientation is defined as the arctangent of the y to x displacement ratio

$$\theta = \text{atan}\left(\frac{vy_b}{vx_b}\right) = \text{atan}\left(\frac{sy_b}{sx_b}\right) \tag{4.8}$$

B-Tracking the microbes

The main file

The code for tracking the microbes is Swimming_microbes.m: a matlab script that it is composed by 5 parts (as also described in the comments within the

code).

1. Definition of the physical parameters like the pixel size etc...
2. Particle Tracking parameters
3. Picture files parameters (like folder, name etc...)
4. A *for* loop over all the images (with a given name within the selected folder) where each image is read, treated, analyzed and the microbes positions detected and saved.
5. The particle tracking code that read the microbes positions pictures by pictures and track them between consecutive images.

In each image recorded during the experiments the background will be bright while the microbes in focus will be darks spots. In order to track the microbe's positions we must:

In the image processing, each picture is represented by a matrix M of double floating values between 0 and 212-1. Each image is analyzed using an in-house code written in Matlab. The collected images are systematically read and converted into a matrix M of double floating values. We define a background matrix as the arithmetic average over all the matrices associated to the images. This background is, then, a matrix B of double floating numbers whose value is everywhere smaller than the signal left by a moving fluorescent particle. By removing from each image this background (I = M-B) and multiplying the result by the mask (I = I x mask), we obtain an image I which is cleaned from experimental noise.

- Clean each image from impurities (like shadow of dust on the microfluidic surface etc...).
- Determine what pixels are covered by a microbe.
- For each microbe detected in a single picture, determine its coordinates (x, y); thus for each picture we have a umber m (different for each picture) of coordinates couples.
- For each couple of consecutive pictures i and $i+1$ track the microbes: determine which couple (x, y) in picture i moved into the couple (x_2, y_2) in picture $i+1$.

The last step is the more complicated. We use the code developed by John C. Crocker, Eric Dufresne and Daniel Blair from 1993 until today.

In practice we read each image k (k varies from 1 to the last image) using the matlab function imread:

```
picture = imread([directory,suffix,'.tif'],k);
```

where *directory* is the folder where the images are saved and *suffix* is the name of the file (in our case the name of the sugar dissolved).

Every image is saved at n-bit (in our case 16), this means that in each image the black color is represented by a 0, the white by $2^n - 1$ and all the gray scale with values in between. We normalize the image to unity by dividing each pixel of the image by $level = 2^{16} - 1$. To do this we must also convert the image into a matrix of real numbers (called *double* for matlab, which stands for double floating):

```
picture = double(picture) ./ level;
```

In order to invert the image having the microbes bright and the background dark (as required by the particle tracking code), we take the opposite of the current image by

```
picture = 1 - picture;
```

we now remove from the image the background image (*back*) that has been previously computed as the arithmetic average or all the pictures:

```
imageBW = picture - back;
```

and we binarize this image by setting every pixel that do not overcome a given threshold to zero and to 1 all the others:

```
imageBW(imageBW < tr) = 0;
```

We now perform 4 steps qhich are required by the particle tracking code to feed

it with *optimized images*. First we smooth a bit the binary image by applying a bandpass filter of size 10:

```
image = bpass(imageBW,1,10);
```

Thus *image* is no longer a binary image. Then we normalize each picture by its maximum value in order to avoid differences between pictures due to some fluctuation in the illumination / detection / camera conversion:

```
image = image ./ max(max(b));
```

The image we look for the position of the peaks (the microbes) in the image, imposing that the peak cannot be larger than 11 pixels

```
pk = pkfnd(image,tr,11);
```

And we finish by finding the coordinates of the *centroid* that fit best each peak (each microbe) at location pk, imposing that the centroid cannot be larger than 15 pixels

```
cnt = cntrd(image,pk,15);
```

The data analysis

Once the Swimming_microbes.m code has run for each data set, we have saved the results of the particle tracking. Now we can run the code Microbes_Tracking_Analysis.m that read the saved data and compute the statistics described above. The analysis is done by reading the particle tracking results as:

```
tracking_result = load([ directory,'tracking_result_',suffix,'.dat'],'tracking_result','-
ASCII');
N_max = max(tracking_result(:,4));
T_max = max(tracking_result(:,3));
```

where N_max and T_max are the maximum number of microbes that have been tracked simultaneously and the maximum number of time steps for which a single microbe has been tracked.

Then the script microbes_trajectories.m extract the tracked microbes coordinates $(x_b(i), y_b(i))$ for each image at step i and the curvilinear coordinate s_b (in the code it is dist).
The script microbes_velocities.m compute the Lagrangian velocities as defined above, based on the coordinates $(x_b(i), y_b(i))$ and the curvilinear coordinate s_b for each microbe b.
Finally, the script Lagrangian_stats.m compute the statistics described above.

Chapter 5

The 2StEP approach

The third section is a dense association of modeling work, experimental work and genomic analyses. All these approaches can be coupled into an integrative, multi-level set of work that I called "2StEP", for Strategy to Study Evolution and Prediction, Figure 5.1. The metabolic modeling performed (chapter 2) permits to make partially testable predictions on target genes for modifications along specialization. These predictions and additional working hypotheses were tested through *in vitro* experiments. First, marks of evolution at the genetic level were detected after the evolution of bacterial population under given constraints (chapter 3). Then these marks of evolution were compared directly to the predictions made through the metabolic flux analyses (also chapter 3). Finally, functional traits (determined via both genetic results and functional hypotheses) were tested via microfluidic experiments (chapter 4). The ambition of the 2STEP workflow is to predict, and subsequently test for these predictions at variable biological levels. As presented along this section, a first 'cycle' of the 2StEP approach permitted to explore the limits of the strategy in place, while still confirming its valuable potential to ameliorate the prediction of evolutionary trajectories.

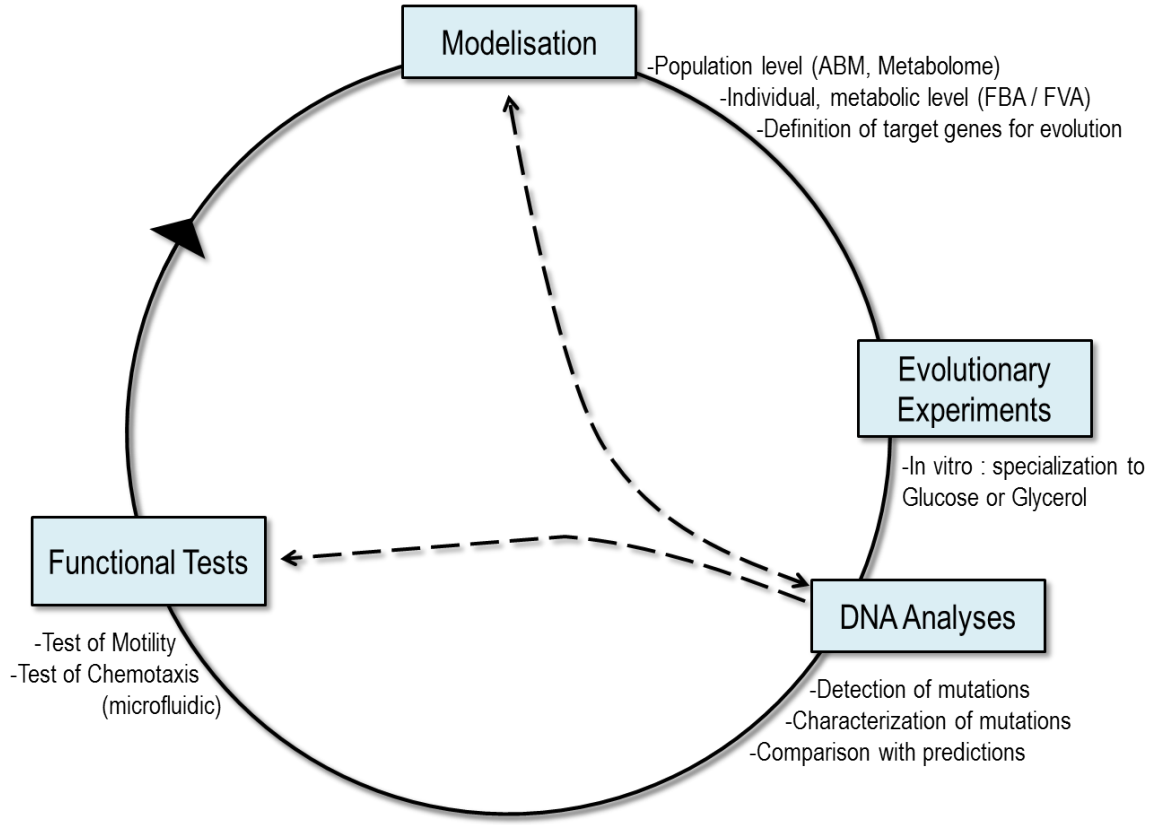


Figure 5.1: 2StEP project Cycle

Part IV

General Discussion and Perspectives

Chapter 1

General Discussion

"The theory of evolution stands up as the landmark of fundamental knowledge in life sciences." Martin-Delgado (2012).

In 2009, the bicentenary anniversary of Charles Darwin and the 150th jubilee of his work on the 'Origin of Species' have been celebrated. The modern synthesis of evolution has been introduced in 1942 by J. Huxley, and was refined within the 'Extended modern synthesis' by M. Pigliucci in 1987. All this work and the field of population genetics have produced a very important corpus of knowledge. Since the end of the 90's and the emergence of genomics, new understandings have challenged the corpus of knowledge in place, notably for population genetics. For example, the understanding of genome functioning through transcriptomic approaches have clearly highlighted that the link between a genotypic modification and the resulting expressed phenotype is not as direct as previously accepted and is in fact accurate for a limited number of features only. This calls for a 'post-modern' synthesis, as strongly argued by E. Koonin for example (Koonin, 2009).

A persisting question which also challenges the classical perception of evolution, is whether or not evolution can be predicted. Even though recent work tends toward evermore predictable genetic events (such as regions of the genome being more susceptible to mutate than others) the characteristic stand-off to prediction still holds: mutations are random, thus evolution is unpredictable. However, in a reduced environment favoring specialization through functions losses, the solutions existing to optimize the fitness of an organism are supposed to be restricted and could the open the way for finer predictions.

In the first part of the thesis, a conception focusing on the evolution of specialization in a reduced environment revisited the notion of predictability of evolution, and put forward the need to develop further the consensual evolutionary theory by integrating recently acquired knowledge. Because biological systems are constrained on several levels (genetic, metabolic, phenotypic, functional) the range of possible modifications of an individuals are (relatively) limited (see Figure 2.1, part I, chapter 2). Moreover, when focusing on a *specialization* context, where instead of evolving new features, organisms tend to lose superfluous ones to focus on essential functions, the expected modifications are all the more perceptible. In this context, we proposed a strategy based on a metabolic approach (interface between the environment and the individual) to predict the modifications occurring in metabolic activity and related genes during a specialization event.

In the second part of the thesis, we showed that focusing on metabolic insights such as the rise of dependency on a common good (Morris et al., 2012), we could also predict the dynamic of a population undergoing such evolutionary trajectory (Mas et al., 2016). This pinpoints the importance of interaction based on the sharing of metabolic function as a strong motor of coevolution. Additionally, with yet another theoretical approach, we demonstrated that through such metabolic complementarity between organisms, we could (at least partially) predict the co-occurrence of species within a community, and the opportunities for Black Queens types of dependency to emerge. These two approaches, both relying on metabolic resolution can help predict dynamics of evolution at the population and community scales.

The third part of the thesis also propose an approach to enhance the prediction of evolution by focusing on metabolism, but at the genetic level. Indeed metabolic functions are determined by genes encoding for the enzymes necessary to chemical reactions constituting the metabolic network of an organism. Therefore models predicting the activity of such metabolic functions can be used to predict the activity of the genes coding for (the enzymes implied in) these functions. The flux analyses models used present a great potential to characterize the activity of the metabolism reflecting actual environmental conditions, which enable the prediction of "specialized metabolism". Yet the complexity of the network between genes and metabolic reactions represent a limit in the strategy developed prediction for precise gene mutations.

Concretely our linear rationale needs to integrate the complexity of genome expression to predict accurately the location of mutation emergence. Yet the metabolic approach proved to be consistent with evolutionary predictions as metabolic genes were shown to be the category of genes that was the most modified during experimental evolution despite the fact that they represent less than 20% of *P. fluorescens* genes. In parallel, a complementary approach is also needed to forecast the evolution of genes that are not metabolic, as they represent most of the genome, and were also subjected to mutations (this is the case for a high number of genes involved in membrane transport of metabolites, in motility and chemotaxis, and in secondary metabolites production).

The various approaches presented above permit an integrative perception of the systems, as some component give information at the genetic and metabolic levels, while others give information at the metabolic oreven populational level. Eventually a **metabolic perceptions** of the biological systems could be considered as the cornerstone of an integrative and functional approach to enhance

prediction both at the genetic, functional, populational and community levels. Until then, each study presented here contributed by itself to increase the knowledge on observable patterns of evolution, and on evolutionary dynamics along phenomenon of specialization.

As each step of this work also sparked parallel questioning and ideas to test further our hypotheses, the next part present the main concrete perspectives envisaged to continue this exciting project on evolution and its predictability.

Chapter 2

Perspectives

2.1 Sequencing RNA to determine mutations effects

A first set of RNA sequencing was realized for the population samples of the Glucose experiments which also corresponded to the samples for which the DNA was analyzed. The results of such an analysis are important to better understand the mutations detected as it will give a quantitative information on genes expression. The relation between mutated genes and their enhanced, unchanged or reduced activity could help understand the evolution of the population, since we had sampled the evolving population along time Yet the noise present in our analysis makes it hardly usable. These transcriptomic analyses required a very high sequencing depth, because changes related to mutations are very low compared to the majority of genes expressed similarly within the population.

2.2 Complementary modelling approaches

An important step to understand better the evolutionary trajectories taken by the population of *P. fluorescens* Pf0-1 during the experimental *in vitro* evolution would be to run the flux analyses models while integrating the results of mutated genes. This way we could infer the metabolic consequences of the mutations observed. This feedback of the metabolic changes observed back into the model could also help define the models parameters better for further predictive analyses. As in the seminal paper of Bordron et al. (2016) it could also be possible to reconstitute *de novo*, from the sequenced genomes, the metabolic network of the bacteria and then compare this *de novo* pathways with the reference metabolic network. Nevertheless, a question to tackle before such modeling can be implemented is to know whether several mutations are carried on single genomes, in order to decide if the models should test every mutations separately or together. It could also be interesting to reconstitute *de novo*, from the sequenced genomes, the metabolic network of some variants, and then compare the *de novo* pathways occurring in variants of the population with the reference metabolic network. Nevertheless, a question to tackle before such modeling can be implemented is to know whether several mutations are carried on single genomes, in order to decide if the models should test every mutations separately or together.

To this purpose, the genotyping of individuals can be realized, it would allow to determine the genomic variants at the individual scale rather than in the population. The interest of such genotyping is double. It would enable to

understand the distribution of the mutations in the populations (i.e. if several mutations are on a singular genome) and it would also enable to retrieve specific mutants, and subsequently test for the specific mutations detected. For example, if a singular genome from the evolved population is known to carry a mutation implied in chemotaxis, we could run further microfluidic experiments on these particular mutants only. The observed results might give a stronger signals on the mutation effects than when testing the entire population where the mutants represent a small percentage of the genomes present.

2.3 New microlfuidic experiments

The microfluidic experiments performed during the PhD were testing for the motility and the chemotaxis. They also gave rise to a large set of question that could be experimentally tested with the same tool. First and as mentioned above, it would be very interesting to run similar experiments as performed before, but using particular (and known) mutants, in order to associate with precision the consequences of a given (set of) mutations on the motile behavior of the bacteria. Moreover, preliminary experiments showed that the motility of *P. fluorescens* Pf0-1 was variable according to the growth phase of the bacteria. A deep description of these phenomenon could be achieved through microfluidic experiments to increase the knowledge on the model-organism.

2.4 Testing the Black Queen hypothesis *in vitro*

The ultimate experiment we would like to perform is to test for the rise of a Black Queen type of dependency for micro-organisms evolving together. The question was tackled and many parameters have to be taken into account. We would need two co-existing species but which have not already co-evolved together. These two types of bacteria should produce a similar common good which is necessary to their metabolism and costly to produce (to favor its loss). But the bacteria should be distinct enough in their genome composition to be easily recognize after shotgun sequencing. Many of these parameters make the realization of this experiment uneasy. The use of genetically engineered bacteria could help palliate these current issues.

Actually genetically engineering organisms (i.e. knock out mutants) may also be of great insights regarding the effects of the mutations detected during

the evolutionary experiments: assessing the metabolic capacities, or the motility capacities of a wild-type organisms after knocking-out genes (i.e. to mimic observed stop mutation) would enable a direct information on mutations effects and potential fitness increase.

As can be seen from the results already gathered, and from the perspectives presented herein, the work realized during this PhD has a high potential for further interdisciplinary research on evolution, which could eventually lead to accurate prediction of organism evolution and better understanding of microbial systems.

On a fundamental line, this knowledge is essential to understand the evolution of interactions between organisms and of the diversity in ecosystems.

In this frame, the conceptual work presented herein on the reductive evolution and specialization opens new windows for understanding the complexity of microbiota and the observed 'self-organisation' (Momeni et al. (2013), Wider et al, 2016). An important emerging prospect is based on the hypothesis that co-occurrence of different members of the microbial community is a consequence of multiple evolution of dependencies leading to stabilize the microbiota complexity.

On a more applied (but also more distant in time) note, this predictive approach could be fundamental in the field of medical Sciences where the evolution of resistance to antibiotic in bacteria is becoming more problematic and urging every day. Imagine, a medical world where the mutations of bacteria facing antibiotics could be anticipated!

Bibliography

Bibliography of -Beyond The Black Queen Hypothesis

Cooper VS, Schneider D, Blot M, Lenski RE. (2001). Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J Bacteriol* **183:**2834–2841.

De Mazancourt C, Schwartz MW. (2010). A resource ratio theory of cooperation. *Ecol. Let.* **13:**349–359.

Driscoll WW, Pepper JW, Pierson LS, Pierson EA. (2011). Spontaneous gac mutants of Pseudomonas biological control strains: Cheaters or mutualists? *Appl. Env. Microb.* **77:**7227–7235.

D'Souza G, S Waschina, Pande S, Bohl K, Kaleta C, Kost C. (2014). Less is more: Selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution* **68:**2559–2570.

Dufresne A, Garczarek L, Partensky F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6:**R14.

Ellers J, Kiers TE, Currie CR, Mcdonald BR, Visser B. (2012). Ecological interactions drive evolutionary loss of traits. *Ecol. Let.* **15:**1071–1082.

Estes JA, Brashares JS, Power ME. (2013). Predicting and detecting reciprocity between Indirect Ecological Interactions and Evolution. *Am. Nat.* **181**:S76-S99.

Estrela S, Trisos CH, Brown SP. (2012). From Metabolism to Ecology: Cross-Feeding Interactions Shape the Balance between Polymicrobial Conflict and Mutualism. *Am. Nat.* **180:**566–576.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:**1242–1245.

Giovannoni SJ. (2012). Vitamins in the sea. *Proc. Natl. Acad. Sci.* **109:**13888–13889.

Giovannoni SJ, Thrash JC, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J.* **8:**1–13.

Hairston NG, Ellner SP, Geber MA, Yoshida T, Fox JA. (2005). Rapid evolution and the convergence of ecological and evolutionary time. *Ecol. Let.* **8:**1114–1127.

Hanson NW, Konwar KM, Hawley AK, Altman T, Karp PD, Hallam SJ. (2014). Metabolic pathways for the whole community. *BMC Genomics* **15:** 619.

Hosoda K, Habuchi KM, Suzuki S, Miyazaki M, Takikawa G, Sakurai T, Kashiwagi A, *et al.* (2014). Adaptation of a cyanobacterium to a biochemically rich environment in experimental evolution as an initial step toward a chloroplast-like state. *PLoS ONE* **9:** e98337.

Hottes AK, Freddolino PL, Khare A, Donnell ZN1, Liu JC, Tavazoie S. (2013). Bacterial Adaptation through Loss of Function. *PLoS Genet.* **9:**e1003617.

Hussa E, Goodrich-Blair H. (2013). It takes a village: ecological and fitness impacts of multipartite mutualism. *Annu. Rev. Microb.* **67:**161–178.

Johnson MT, Stinchcombe JR. (2007). An emerging synthesis between community ecology and evolutionary biology. *Trends Ecol. Evol.* **22:**250–257.

Kreft JU, Bonhoeffer S. (2005). The evolution of groups of cooperating bacteria and the growth rate versus yield trade-off. *Microbiology* **151:**637–641.

Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, *et al.* (2009). Relaxed selection in the wild. *Trends Ecol. Evol.* **24:**487–496.

Lee MC, Marx CJ. (2012). Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* **8:**2–9.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. (2014). Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. *ISME J.* **8:**1428–1439.

McCutcheon JP, Moran NA. (2011). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microb.* **10:**13–26.

McGinty SE, Rankin DJ, Brown SP. (2011). Horizontal gene transfer and the evolution of bacterial cooperation. *Evolution* **65:**21–32.

Morris JJ, Lenski RE, Zinser ER. (2012). The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *MBio* **3:** e00036-12.

Morris JJ, Papoulis SE, Lenski RE. (2014). Coexistence of evolving bacteria stabilized by a shared black queen function: experimental evolution of a black queen community. *Evolution* **68**:2960-2971.

Nadell CD, Xavier JB, Foster KR. (2009). The sociobiology of biofilms. *FEMS Microb. Rev.* **33:**206–224.

Ochman H, Moran N. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292:**1096–1099.

Pande S, Merker H, Bohl K, Reichelt M, Schuster S, de Figueiredo L, *et al.* (2014). Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J.* **8:**953–962.

Partensky F, Hess WR, Vaulot D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microb. Mol. Biol. Rev.* **63:**106–127.

Porter ML, Crandall KA. (2003). Lost along the way: The significance of evolution in reverse. *Trends Ecol. Evol.* **18:**541–547.

Richards TA, Talbot NJ. (2013). Horizontal gene transfer in osmotrophs: playing with public goods. *Nat. Rev. Microb.* **11:**720–727.

Sachs JL, Hollowell C. (2012). The Origins of Cooperative Bacterial Communities. *MBio* **3**.

Schoener TW. (2011). The newest synthesis: understanding the interplay of evolutionary and ecological dynamics. *Science* **331:**426–429.

Smith J. (2001). The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. B* **268:**61–69.

Turcotte MM, Corrin MSC, Johnson MTJ. (2012). Adaptive Evolution in Ecological Communities. *PLoS Biol.* **10:** e1001332.

Van Valen L. (1973). A new evolutionary law. *Evol. Theory* **1,** 1–30.

Visser B, Le Lann C, den Blanken FJ, Harvey JF, van Alphen JJM, Ellers J. (2010). Loss of lipid synthesis as an evolutionary consequence of a parasitic lifestyle. *Proc. Natl. Acad. Sci. USA* **107:**8677–8682.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. (2002). Phylogenies and Community Ecology. *Annu. Rev. Ecol. Syst.* **33:**475–505.

Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci. USA* **112:**6449–6454.

Bibliography

Aguilar-Rodríguez, J., Payne, J. L., and Wagner, A. (2017). A thousand empirical adaptive landscapes and their navigability. *Nature Ecology & Evolution*, 1(2):0045.

Ahmed, T., Shimizu, T. S., and Stocker, R. (2010). Microfluidics for bacterial chemotaxis. *Integrative Biology*, 2(11-12):604.

Albesa, I., Barberis, L. I., Pajaro, M. C., and Eraso, A. J. (1985). Pyoverdine production by Pseudomonas fluorescens in synthetic media with various sources of nitrogen. *Journal of general microbiology*, 131(12):3251–3254.

Arendt, J. and Reznick, D. (2008). Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology & Evolution*, 23(1):26–32.

Bailey, S. F., Blanquart, F., Bataillon, T., and Kassen, R. (2016). What drives parallel evolution?: How population size and mutational variation contribute to repeated evolution. *BioEssays*.

Bailey, S. F., Hinz, A., and Kassen, R. (2014). Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nature Communications*, 5.

Bank, C., Matuszewski, S., Hietpas, R. T., and Jensen, J. D. (2016). On the (un-)predictability of a large intragenic fitness landscape. Technical Report biorxiv;048769v4.

Barrett, R., MacLean, R., and Bell, G. (2005). Experimental Evolution of *Pseudomonas fluorescens* in Simple and Complex Environments. *The American Naturalist*, 166(4):470–480.

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C., and Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature*, 462(7269):90–93.

Binz, M., Lee, A. P., Edwards, C., and Nicolau, D. V. (2010). Motility of bacteria in microfluidic structures. *Microelectronic Engineering*, 87(5-8):810–813.

Blank, D., Wolf, L., Ackermann, M., and Silander, O. K. (2014). The predictability of molecular evolution during functional innovation. *Proceedings of the National Academy of Sciences*, 111(8):3044–3049.

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237.

Bono, L. M., Smith, L. B., Pfennig, D. W., and Burch, C. L. (2017). The emergence of performance trade-offs during local adaptation: insights from experimental evolution. *Molecular Ecology*.

Bordron, P., Latorre, M., Cortés, M.-P., González, M., Thiele, S., Siegel, A., Maass, A., and Eveillard, D. (2016). Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *MicrobiologyOpen*, 5(1):106–117.

Boscaro, V., Felletti, M., Vannini, C., Ackerman, M. S., Chain, P. S. G., Malfatti, S., Vergez, L. M., Shin, M., Doak, T. G., Lynch, M., and Petroni, G. (2013). Polynucleobacter necessarius, a model for genome reduction in both free-living and symbiotic bacteria. *Proceedings of the National Academy of Sciences*, 110(46):18590–18595.

Bowler, P. J. (1989). *Evolution: the history of an idea*. History of Sciencie. University of California Press.

Brockhurst, M. A. and Koskella, B. (2013). Experimental coevolution of species interactions. *Trends in Ecology & Evolution*, 28(6):367–375.

Brundrett, M. C. (2002). Coevolution of roots and mycorrhizas of land plants. *New phytologist*, 154(2):275–304.

Buckling, A. and Rainey, P. B. (2002). Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1494):931–936.

Budinich, M., Bourdon, J., Larhlimi, A., and Eveillard, D. (2017). A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLOS ONE*, 12(2):e0171744.

Burrows, L. L. (2012). *Pseudomonas aeruginosa* Twitching Motility: Type IV Pili in Action. *Annual Review of Microbiology*, 66(1):493–520.

Casaz, P., Happel, A., Keithan, J., Read, D. L., Strain, S. R., and Levy, S. B. (2001). The Pseudomonas fluorescens transcription activator AdnA is required for adhesion and motility. *Microbiology*, 147(2):355–361.

Castellanos, M. C., Wilson, P., and Thomson, J. D. (2004). 'Anti-bee' and 'pro-bird' changes during the evolution of hummingbird pollination in Penstemon flowers. *Journal of Evolutionary Biology*, 17(4):876–885.

Christin, P.-A., Weinreich, D. M., and Besnard, G. (2010). Causes and evolutionary significance of genetic convergence. *Trends in Genetics*, 26(9):400–405.

Comeault, A. A., Carvalho, C. F., Dennis, S., Soria-Carrasco, V., and Nosil, P. (2016). Color phenotypes are under similar genetic control in two distantly related species of *Timema* stick insect: GENETIC ARCHITECTURE OF COLOR IN *TIMEMA*. *Evolution*, 70(6):1283–1296.

Connell, J. H. (1980). Diversity and the Coevolution of Competitors, or the Ghost of Competition Past. *Oikos*, 35(2):131.

Conte, G. L., Arnegard, M. E., Peichel, C. L., and Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):5039–5047.

Conway Morris, S. (2010). Evolution: like any other science it is predictable. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):133–145.

Cooper, V. S., Schneider, D., Blot, M., and Lenski, R. E. (2001). Mechanisms Causing Rapid and Parallel Losses of Ribose Catabolism in Evolving Populations of Escherichia coli B. *Journal of Bacteriology*, 183(9):2834–2841.

Danchin, t., Charmantier, A., Champagne, F. A., Mesoudi, A., Pujol, B., and Blanchet, S. (2011). Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nature Reviews Genetics*, 12(7):475–486.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. Murray, London. or the Preservation of Favored Races in the Struggle for Life.

Dawkins, R. (1990). *The Selfish Gene*. Oxford University Press.

Dawkins, R. (1999). *The extended phenotype : the long reach of the gene*. Oxford University Press, Oxford ;;New York, rev. ed. edition.

de Visser, J. A. G. and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490.

de Vries, H., Darbishire, A., and Farmer, J. (1901). *The mutation theory; experiments and observations on the origin of species in the vegetable kingdom*, volume 2. Chicago,Open Court Publishing Company; [etc.]. http://www.biodiversitylibrary.org/bibliography/4634.

Dobler, S., Dalla, S., Wagschal, V., and Agrawal, A. A. (2012). Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. *Proceedings of the National Academy of Sciences*, 109(32):13040–13045.

Dobzhansky, T. (1973). Nothing in Biology Makes Sense except in the Light of Evolution. *The American Biology Teacher*, 35(3):125–129.

Doebeli, M. and Ispolatov, I. (2014). Chaos and unpredictability in evolution. *Evolution*, 68(5):1365–1373.

Duarte, J., Rodrigues, C., Januário, C., Martins, N., and Sardanyés, J. (2015). How Complex, Probable, and Predictable is Genetically Driven Red Queen Chaos? *Acta Biotheoretica*, 63(4):341–361.

Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome biology*, 6(2):R14.

Dykhuizen, D. E. (1990). Experimental studies of natural selection in bacteria. *Annual Review of Ecology and Systematics*, pages 373–398.

Dykhuizen, D. E. (1993). [45] Chemostats used for studying natural selection and adaptive evolution. In *Methods in Enzymology*, volume 224, pages 613–631. Elsevier. DOI: 10.1016/0076-6879(93)24046-W.

Ehrlich, P. R. and Raven, P. H. (1964). Butterflies and plants: a study in coevolution. *Evolution*, 18(4):586–608.

Ellers, J., Toby Kiers, E., Currie, C. R., McDonald, B. R., and Visser, B. (2012). Ecological interactions drive evolutionary loss of traits. *Ecology Letters*, 15(10):1071–1082.

Faguy, D. M. and Jarrell, K. F. (1999). A twisted tale: the origin and evolution of motility and chemotaxis in prokaryotes. *Microbiology*, 145(2):279–281.

Feldman, C. R., Brodie, E. D., Brodie, E. D., and Pfrender, M. E. (2012). Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. *Proceedings of the National Academy of Sciences*, 109(12):4556–4561.

Fenchel, T. and Thar, R. (2004). "Candidatus Ovobacter propellens": a large conspicuous prokaryote with an unusual motility behaviour. *FEMS Microbiology Ecology*, 48(2):231–238.

Ferriere, R. and Legendre, S. (2012). Eco-evolutionary feedbacks, adaptive dynamics and evolutionary rescue theory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1610):20120081–20120081.

Finn, T. J., Shewaramani, S., Leahy, S. C., Janssen, P. H., and Moon, C. D. (2017). Dynamics and genetic diversification of *Escherichia coli* during experimental adaptation to an anaerobic environment. *PeerJ*, 5:e3244.

Forsberg, S. K. G., Bloom, J. S., Sadhu, M. J., Kruglyak, L., and Carlborg, r. (2017). Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature Genetics*, 49(4):497–503.

Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature*, 459(7244):193–199.

Fussmann, G. F., Loreau, M., and Abrams, P. A. (2007). Eco-evolutionary dynamics of communities and ecosystems. *Functional Ecology*, 21(3):465–477.

Gaut, B. S. (2015). Evolution Is an Experiment: Assessing Parallelism in Crop Domestication and Experimental Evolution: (Nei Lecture, SMBE 2014, Puerto Rico). *Molecular Biology and Evolution*, 32(7):1661–1671.

Giovannoni, S. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738):1242–1245.

Giovannoni, S. J., Thrash, J. C., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME journal*.

Goldman, A. D. and Landweber, L. F. (2016). What Is a Genome? *PLOS Genetics*, 12(7):e1006181.

Gompel, N. and Prud'homme, B. (2009). The causes of repeated genetic evolution. *Developmental Biology*, 332(1):36–47.

Gould, S. J. (1989). *Wonderful life: The burgess shale and the nature of history*. W. W. Norton, New York.

Goymer, P. (2007). Synonymous mutations break their silence. *Nature Reviews Genetics*, 8(2):92–92.

Gravel, D., Bell, T., Barbera, C., Bouvier, T., Pommier, T., Venail, P., and Mouquet, N. (2011). Experimental niche evolution alters the strength of the diversity–productivity relationship. *Nature*, 469(7328):89–92.

Gross, J. B., Borowsky, R., and Tabin, C. J. (2009). A Novel Role for Mc1r in the Parallel Evolution of Depigmentation in Independent Populations of the Cavefish Astyanax mexicanus. *PLoS Genetics*, 5(1):e1000326.

Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):489.

Hairston, N. G., Ellner, S. P., Geber, M. A., Yoshida, T., and Fox, J. A. (2005). Rapid evolution and the convergence of ecological and evolutionary time: Rapid evolution and the convergence of ecological and evolutionary time. *Ecology Letters*, 8(10):1114–1127.

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, 210(9):1518–1525.

Hardin, G. (1968). The tragedy of the commons. *Science*, 162:1243–47.

Harmon, L. J., Kolbe, J. J., Cheverud, J. M., and Losos, J. B. (2005). Convergence and the multidimensional niche. *Evolution*, 59(2):409–421.

Herbert, S. (1863). *The principles of biology*, volume 1. New York :D. Appleton,. http://www.biodiversitylibrary.org/bibliography/29113.

Herron, M. D. and Doebeli, M. (2013). Parallel evolutionary dynamics of adaptive diversification in Escherichia coli. *PLoS biology*, 11(2):e1001490.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, page 16048.

Huneman, P. (2012). Determinism, predictability and open-ended evolution: lessons from computational emergence. *Synthese*, 185(2):195–214.

Isea, R. (2015). The Present-Day Meaning of the Word Bioinformatics. *Global Journal of Advanced Research*, 2:70–73.

Jarrell, K. F. and McBride, M. J. (2008). The surprisingly diverse ways that prokaryotes move. *Nature Reviews Microbiology*, 6(6):466–476.

Jerison, E. R. and Desai, M. M. (2015). Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. *Current Opinion in Genetics & Development*, 35:33–39.

Johnson, G. T. and Hertig, S. (2014). A guide to the visual analysis and communication of biomolecular structural data. *Nature reviews. Molecular cell biology*, 15(10):690.

Kaplan, J. (2008). The end of the adaptive landscape metaphor? *Biology & Philosophy*, 23(5):625–638.

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., and Chisholm, S. W. (2007). Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. *PLoS Genetics*, 3(12):e231.

Kim, J., Lee, S., Shin, H., Kim, S. C., and Cho, B.-K. (2012). Elucidation of bacterial genome complexity using next-generation sequencing. *Biotechnology and Bioprocess Engineering*, 17(5):887–899.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

King, J. L. and Jukes, T. H. (1969). Non-darwinian evolution. *Science*, 164(3881):788–798.

Koonin, E. V. (2009). *Towards a postmodern synthesis of evolutionary biology*. Taylor & Francis.

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Reviews in Microbiology*, 55(1):709–742.

Kruger, N. J. and von Schaewen, A. (2003). The oxidative pentose phosphate pathway: structure and organisation. *Current Opinion in Plant Biology*, 6(3):236–246.

Kryazhimskiy, S., Rice, D. P., Jerison, E. R., and Desai, M. M. (2014). Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522.

Lahti, D. C., Johnson, N. A., Ajie, B. C., Otto, S. P., Hendry, A. P., Blumstein, D. T., Coss, R. G., Donohue, K., and Foster, S. A. (2009). Relaxed selection in the wild. *Trends in Ecology & Evolution*, 24(9):487–496.

Lai, C.-Y., Baumann, L., and Baumann, P. (1994). Amplification of trpEG: adaptation of Buchnera aphidicola to an endosymbiotic association with aphids. *Proceedings of the National Academy of Sciences*, 91(9):3819–3823.

Laland, K. N. and Sterelny, K. (2006). Perspective: seven reasons (not) to neglect niche construction. *Evolution*, 60(9):1751–1762.

Lapidot, I. and Conley, D. W. (2015). Evolution Predictability, Lamarck, Altshuller, Darwin and Chaos. *Procedia Engineering*, 131:115–122.

Laubichler, M. D. and Renn, J. (2015). Extended evolution: A conceptual framework for integrating regulatory networks and niche construction: EXTENDED EVOLUTION. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(7):565–577.

Le Gac, M., Cooper, T. F., Cruveiller, S., Médigue, C., and Schneider, D. (2013). Evolutionary history and genetic parallelism affect correlated responses to evolution. *Molecular Ecology*, 22(12):3292–3303.

Lee, M.-C. and Marx, C. J. (2012). Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. *PLoS Genetics*, 8(5):e1002651.

Leiby, N. and Marx, C. J. (2014). Metabolic Erosion Primarily Through Mutation Accumulation, and Not Tradeoffs, Drives Limited Evolution of Substrate Specificity in Escherichia coli. *PLoS Biology*, 12(2):e1001789.

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. (1991). Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2, 000 generations. *American Naturalist*, 138(6):1315–1341.

Levin, B. R., Perrot, V., and Walker, N. (2000). Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics*, 154(3):985–997.

Lobkovsky, A. E. and Koonin, E. V. (2012). Replaying the Tape of Life: Quantification of the Predictability of Evolution. *Frontiers in Genetics*, 3.

Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2011). Predictability of Evolutionary Trajectories in Fitness Landscapes. *PLoS Computational Biology*, 7(12):e1002302.

Losos, J. B. (1998). Contingency and Determinism in Replicated Adaptive Radiations of Island Lizards. *Science*, 279(5359):2115–2118.

Losos, J. B. (2011). CONVERGENCE, ADAPTATION, AND CONSTRAINT. *Evolution*, 65(7):1827–1840.

Louca, S. and Doebeli, M. (2015). Calibration and analysis of genome-based models for microbial ecology. *eLife*, 4:e08208.

Ludewig, M. and Fehlhaber, K. (2009). Investigations on the generation time of selected gramnegative bacteria species. *Archiv für lebensmittelhygiene*, 60:56–60.

Maddamsetti, R., Hatcher, P. J., Green, A. G., Williams, B. L., Marks, D. S., and Lenski, R. E. (2017). Core Genes Evolve Rapidly in the Long-term Evolution Experiment with Escherichia coli. *Genome biology and evolution*, 9(4):1072–1083.

Malthus, T. R. (1826). *An Essay on the Principle of Population*. John Murray, London, sixth edition edition.

Martin-Delgado, M. A. (2012). On Quantum Effects in a Theory of Biological Evolution. *Scientific Reports*, 2.

Marwan, W., Alam, M., and Oesterhelt, D. (1991). Rotation and switching of the flagellar motor assembly in Halobacterium halobium. *Journal of bacteriology*, 173(6):1971–1977.

Mas, A., Jamshidi, S., Lagadeuc, Y., Eveillard, D., and Vandenkoornhuyse, P. (2016). Beyond the Black Queen Hypothesis. *ISME Journal*, 10(9):2085–2091.

Mastropaolo, M. D., Silby, M. W., Nicoll, J. S., and Levy, S. B. (2012). Novel Genes Involved in Pseudomonas fluorescens Pf0-1 Motility and Biofilm Formation. *Applied and Environmental Microbiology*, 78(12):4318–4329.

Matthews, B., Narwani, A., Hausch, S., Nonaka, E., Peter, H., Yamamichi, M., Sullam, K. E., Bird, K. C., Thomas, M. K., Hanley, T. C., and Turner, C. B. (2011). Toward an integration of evolutionary biology and ecosystem science: Integration of evolutionary biology and ecosystem science. *Ecology Letters*, 14(7):690–701.

Maughan, H., Callicotte, V., Hancock, A., Birky, C. W., Nicholson, W. L., and Masel, J. (2006). The population genetics of phenotypic deterioration in experimental populations of Bacillus subtilis. *Evolution*, 60(4):686–695.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

McDaniel, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., and Paul, J. H. (2010). High Frequency of Horizontal Gene Transfer in the Oceans. *Science*, 330(6000):50–50.

Minamino, T., Imae, Y., Oosawa, F., Kobayashi, Y., and Oosawa, K. (2003). Effect of Intracellular pH on Rotational Speed of Bacterial Flagellar Motors. *Journal of Bacteriology*, 185(4):1190–1194.

Mitchell, J. G. and Kogure, K. (2006). Bacterial motility: links to the environment and a driving force for microbial physics: Bacterial motility. *FEMS Microbiology Ecology*, 55(1):3–16.

Miton, C. M. and Tokuriki, N. (2016). How mutational epistasis impairs predictability in protein evolution and design: How Epistasis Impairs Predictability in Enzyme Evolution. *Protein Science*, 25(7):1260–1272.

Mitri, S. and Foster, K. R. (2016). Pleiotropy and the low cost of individual traits promote cooperation: BRIEF COMMUNICATION. *Evolution*, 70(2):488–494.

Momeni, B., Waite, A. J., and Shou, W. (2013). Spatial self-organization favors heterotypic cooperation over cheating. *Elife*, 2:e00960.

Morozov, A. (2013). Modelling biological evolution: recent progress, current challenges and future direction. *Interface Focus*, 3(6):20130054–20130054.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012a). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio*, 3(2):e00036–12–e00036–12.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012b). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio*, 3(2):e00036–12–e00036–12.

Morris, J. J., Papoulis, S. E., and Lenski, R. E. (2014). COEXISTENCE OF EVOLVING BACTERIA STABILIZED BY A SHARED BLACK QUEEN FUNCTION: EXPERIMENTAL EVOLUTION OF A BLACK QUEEN COMMUNITY. *Evolution*, 68(10):2960–2971.

Muschick, M., Indermaur, A., and Salzburger, W. (2012). Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. *Current Biology*, 22(24):2362–2368.

Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. D., and Andersson, D. I. (2005). Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(34):12112–12116.

Ochman, H. (2001). Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. *Science*, 292(5519):1096–1099.

Orth, J. D., Thiele, I., and Palsson, B. . (2010). What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248.

Overbeek, R. (2005). The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research*, 33(17):5691–5702.

Pande, S., Merker, H., Bohl, K., Reichelt, M., Schuster, S., de Figueiredo, L. F., Kaleta, C., and Kost, C. (2013). Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *The ISME journal*.

Piepenbrink, K. H. and Sundberg, E. J. (2016). Motility and adhesion through type IV pili in Gram-positive bacteria. *Biochemical Society Transactions*, 44(6):1659–1666.

Pigliucci, M. (2006). Phenotypic plasticity and evolution by genetic assimilation. *Journal of Experimental Biology*, 209(12):2362–2367.

Pigliucci, M. and Finkelman, L. (2014). The Extended (Evolutionary) Synthesis Debate: Where Science Meets Philosophy. *BioScience*, 64(6):511–516.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., and Schwartz, J.-M. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC systems biology*, 4(1):114.

Rainey, P. B. and Travisano, M. (1998). Adaptive radiation in a heterogeneous environment. *Nature*, 394(6688):69.

Ribeck, N. and Lenski, R. E. (2015). Modeling and quantifying frequency-dependent fitness in microbial populations with cross-feeding interactions: BRIEF COMMUNICATION. *Evolution*, 69(5):1313–1320.

Rico-Jim?nez, M., Reyes-Darias, J. A., Ortega, l., D?ez Pe?a, A. I., Morel, B., and Krell, T. (2016). Two different mechanisms mediate chemotaxis to inorganic phosphate in Pseudomonas aeruginosa. *Scientific Reports*, 6(1).

Riley, M. S., Cooper, V. S., Lenski, R. E., Forney, L. J., and Marsh, T. L. (2001). Rapid phenotypic change and diversification of a soil bacterium during 1000 generations of experimental evolution. *Microbiology*, 147(4):995–1006.

Rosenberg, E. and Zilber-Rosenberg, I. (2016). Microbes Drive Evolution of Animals and Plants: the Hologenome Concept. *mBio*, 7(2):e01395–15.

Rosenblum, E. B., Parent, C. E., and Brandt, E. E. (2014). The Molecular Basis of Phenotypic Convergence. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):203–226.

Ross-Gillespie, A., Dumas, Z., and Kümmerli, R. (2015). Evolutionary dynamics of interlinked public goods traits: an experimental study of siderophore production in *Pseudomonas aeruginosa*. *Journal of Evolutionary Biology*, 28(1):29–39.

Rundle, H. D., Nagel, L., Boughman, J. W., and Schluter, D. (2000). Natural selection and parallel speciation in sympatric sticklebacks. *Science*, 287(5451):306–308.

Rusconi, R., Garren, M., and Stocker, R. (2014). Microfluidics Expanding the Frontiers of Microbial Ecology. *Annual Review of Biophysics*, 43(1):65–91.

Sampedro, I., Parales, R. E., Krell, T., and Hill, J. E. (2014). *Pseudomonas* chemotaxis. *FEMS Microbiology Reviews*, pages n/a–n/a.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., and Palsson, B. . (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, 6(9):1290–1307.

Schimel, J., Balser, T. C., and Wallenstein, M. (2007). Microbial stress-response physiology and its implications for ecosystem function. *Ecology*, 88(6):1386–1394.

Schomburg, I., Chang, A., Placzek, S., Sohngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M., and Schomburg, D. (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research*, 41(D1):D764–D772.

Schulte, R. D., Makus, C., Hasert, B., Michiels, N. K., and Schulenburg, H. (2010). Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. *Proceedings of the National Academy of Sciences*, 107(16):7359–7364.

Signor, S., Liu, Y., Rebeiz, M., and Kopp, A. (2016). Genetic Convergence in the Evolution of Male-Specific Color Patterns in Drosophila. *Current Biology*, 26(18):2423–2433.

Silby, M. W., Cerdeño-Tárraga, A. M., Vernikos, G. S., Giddens, S. R., Jackson, R. W., Preston, G. M., Zhang, X.-X., Moon, C. D., Gehrig, S. M., Godfrey, S. A., and others (2009). Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. *Genome biology*, 10(5):R51.

Stearns, S. C. (1989). Trade-Offs in Life-History Evolution. *Functional Ecology*, 3(3):259.

Steinberg, B. and Ostermeier, M. (2016). Environmental changes bridge evolutionary valleys. *Science Advances*, 2(1):e1500921–e1500921.

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics*, 14(11):751–764.

Stern, D. L. and Orgogozo, V. (2008). THE LOCI OF EVOLUTION: HOW PREDICTABLE IS GENETIC EVOLUTION? *Evolution*, 62(9):2155–2177.

Sturmbauer, C., Salzburger, W., Duftner, N., Schelly, R., and Koblmüller, S. (2010). Evolutionary history of the Lake Tanganyika cichlid tribe Lamprologini (Teleostei: Perciformes) derived from mitochondrial and nuclear DNA data. *Molecular Phylogenetics and Evolution*, 57(1):266–284.

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., Luo, H., Wright, J. J., Landry, Z. C., and Hanson, N. W. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences*, 110(28):11463–11468.

Szendro, I. G., Franke, J., de Visser, J. A. G. M., and Krug, J. (2013). Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences*, 110(2):571–576.

Taylor, M. B. and Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, 31(1):34–40.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D., and Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615):165–170.

Theis, K. R., Dheilly, N. M., Klassen, J. L., Brucker, R. M., Baines, J. F., Bosch, T. C. G., Cryan, J. F., Gilbert, S. F., Goodnight, C. J., Lloyd, E. A., Sapp, J., Vandenkoornhuyse, P., Zilber-Rosenberg, I., Rosenberg, E., and Bordenstein, S. R. (2016). Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes. *mSystems*, 1(2):e00028–16.

Thiele, I. and Palsson, B. . (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121.

Toll-Riera, M., San Millan, A., Wagner, A., and MacLean, R. C. (2016). The genomic basis of evolutionary innovation in Pseudomonas aeruginosa. *PLoS Genet*, 12(5):e1006005.

Travisano, M. and Lenski, R. E. (1996). Long-term experimental evolution in Escherichia coli. IV. Targets of selection and the specificity of adaptation. *Genetics*, 143(1):15–26.

Tripp, H. J., Bench, S. R., Turk, K. A., Foster, R. A., Desany, B. A., Niazi, F., Affourtit, J. P., and Zehr, J. P. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, 464(7285):90–94.

Tripp, H. J., Kitner, J. B., Schwalbach, M. S., Dacey, J. W. H., Wilhelm, L. J., and Giovannoni, S. J. (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature*, 452(7188):741–744.

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo, P. (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, 43(2):e11–e11.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary theory*, 1:1–30.

Vandenkoornhuyse, P., Dufresne, A., Quaiser, A., Gouesbet, G., Binet, F., Francez, A.-J., Mahé, S., Bormans, M., Lagadeuc, Y., and Couée, I. (2010). Integration of molecular functions at the ecosystemic level: breakthroughs and future goals of environmental genomics and post-genomics: Environmental genomics. *Ecology Letters*, 13(6):776–791.

Varma, A. and Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and environmental microbiology*, 60(10):3724–3731.

Venail, P. A. and Vives, M. J. (2013). Positive Effects of Bacterial Diversity on Ecosystem Functioning Driven by Complementarity Effects in a Bioremediation Context. *PLoS ONE*, 8(9):e72561.

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., and Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, 116(5):882–892.

Visser, B., Le Lann, C., den Blanken, F. J., Harvey, J. A., van Alphen, J. J. M., and Ellers, J. (2010). Loss of lipid synthesis as an evolutionary consequence of a parasitic lifestyle. *Proceedings of the National Academy of Sciences*, 107(19):8677–8682.

Wade, M. J. (2011). The neo-modern synthesis: The confluence of new data and explanatory concepts. *BioScience*, 61(5):407–408.

Wei, X. and Bauer, W. D. (1998). Starvation-induced changes in motility, chemotaxis, and flagellation of Rhizobium meliloti. *Applied and environmental microbiology*, 64(5):1708–1714.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583.

Wilkinson, J. F. (1963). Carbon and Energy Storage in Bacteria. *Microbiology*, 32(2):171–176.

Wiper-Bergeron, N. and Skerjanc, I. S. (2009). Transcription and the Control of Gene Expression. In Krawetz, S., editor, *Bioinformatics for Systems Biology*, pages 33–49. Humana Press, Totowa, NJ. DOI: 10.1007/978-1-59745-440-7_2.

Wood, T., Burke, J., and Rieseberg, L. (2005). Parallel genotypic adaptation: when evolution repeats itself. *Genetics of Adaptation*, pages 157–170.

Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A., and Lenski, R. E. (2006). Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. *Proceedings of the National Academy of Sciences*, 103(24):9107–9112.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the XI International Congress of Genetics*, 8:209–222.

Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F., and Hairston Jr, N. G. (2003). Rapid evolution drives ecological dynamics in a predator-prey system. *Nature*, 424(6946):303.

Yu, J. (2001). Production of PHA from starchy wastewater via organic acids. *Journal of Biotechnology*, 86(2):105–112.

ABSTRACT

Is evolution predictable?
While the usual response is an almost unanimous No, a growing corpus of knowledge suggests it is time to seriously revisit this answer. Even though mutations are still assumed to be random, the detection of genetic patterns underlying evolutionary events opens the door on potential strategies to forecast the evolutionary trajectories followed by organisms when they adapt to changing constraints. When organisms undergo functional specialization (through genes and function loss) to adapt to given environmental cues, the possible evolutionary paths they can take are restrained and should thus be more easily predictable. In this context of reductive evolution through specialization, the objectives of this thesis are to understand better the interaction between environmental constraints, metabolism, genetic evolution and functional adaptations, and in a second time to predict, for given environmental constraints, the evolutionary trajectories which will be followed by organisms to adapt to these constraints. A first approach focuses on the importance of biotic interactions as being determinants of evolutionary trajectories, and how by modelling a beneficial rise of dependency on a common good, we could predict the dynamic of a population undergoing such evolutionary events . A second approach investigates how changes entailed in metabolisms and functions by a change in environmental constraints could be forecast and tested. Based on a metabolic-centered view, we combined modelling and experimental work to encompass the evolution of specialization at the genetic, metabolic and functional levels. We show that evolution trajectories can partially be predicted according to specific environmental conditions, but that these predictions are limited due to the intricacy of the genetic expression network. This exploratory and interdisciplinary work increases the knowledge on evolutionary determinants and trajectories followed by organism during specialization. It also demonstrates a great potential for predictions, notably through a metabolic perception of the systems.

L'évolution est-elle prédictible?

Alors que la réponse habituelle est un non presque unanime, un corpus croissant de connaissances suggère qu'il est temps de revoir cette réponse. Même si les mutations sont toujours considérées comme aléatoires, la détection de patterns génétiques sous-jacents aux événements évolutifs ouvre la porte sur des stratégies potentielles permettant de prévoir les trajectoires évolutives suivies par les organismes lorsqu'ils s'adaptent à des contraintes changeantes. Quand les organismes subissent une spécialisation fonctionnelle (à travers la perte de gènes et de fonctions) pour s'adapter à des signaux environnementaux donnés, les trajectoires évolutives possibles qu'ils peuvent emprunter sont restreintes et devraient donc être plus facilement prévisibles. Dans ce contexte d'évolution réductive par spécialisation, les objectifs de cette thèse sont de mieux comprendre l'interaction entre contraintes environnementales, métabolisme, évolution génétique et adaptations fonctionnelles, et dans un deuxième temps de prédire, pour des contraintes données, les trajectoires évolutives qui seront suivies par les organismes pour s'adapter à ces contraintes. Une première approche met l'accent sur l'importance des interactions biotiques en tant que déterminants des trajectoires évolutives, et comment, en modélisant une hausse bénéfique de dépendance envers un bien commun, il est possible de prédire la dynamique d'une population subissant de tels évènements évolutifs. Une deuxième approche étudie comment les changements observés au niveau métabolique et fonctionnel et engendrés par des modifications de contraintes environnementales pourraient être prévus et testés. Sur la base d'une vision centrée sur le métabolisme, des travaux de modélisation et des travaux d’expérimentions ont été combinés pour étudier l'évolution de la spécialisation aux niveaux génétiques, métaboliques et fonctionnels. Nous montrons que les trajectoires évolutives suivies peuvent-être partiellement prédites en fonction de conditions environnementales spécifiques, mais que ces prédictions sont limitées en raison de la complexité du réseau de l'expression génétique.

– – –

L’évolution en biologie est déterminée par deux processus clefs: l’émergence de variation via des phénomènes stochastiques, et la sélection naturelle, où la fitness d’un organisme dans un environnement donné va déterminer sa probabilité de survivre et donc la probabilité de transmission de son patrimoine génétique aux générations futures. La combinaison de ces mutations stochastiques et de la sélection font, qu’a priori, l’évolution est un processus qui semble imprédictible. Néanmoins, et de façon frappante, un certain nombre d'exemples d'évolution convergente (traits morphologique et fonctionnels convergents, mais aussi processus génétiques sous-jacents parallèles) ont été observés et mieux déterminés, ce qui remet d’actualité la question de la prédictibilité de l’évolution.

Ici, cette question de la prédictibilité est explorée par un travail aussi intégratif que possible, qui engage une approche interdisciplinaire combinant des modèles

computationnel à des expériences d'évolution in vitro et à des analyses à l'échelle du génome, mais aussi du comportement des populations de bactéries. Dans le contexte de la spécialisation, l'objectif principal de ce travail est de déterminer dans quelle mesure les trajectoires évolutives des bactéries, tant au niveau des interactions génétiques, fonctionnelles, qu'écologiques, pourraient être prévues.

Cette thèse, qui compte plus de 200 pages, bibliographie exclue, comprend 4 sections.
La première section fournit, dans un premier chapitre, un historique de la recherche en évolution et des théories aujourd'hui débattues, avec un focus sur l'évolution dans le monde bactérien. Cette partie permet une introduction des points d'évolution qui seront traités dans la suite du manuscrit.
Le second chapitre de cette section explique le raisonnement général de cette thèse et présente les éléments de conceptualisation nécessaires aux hypothèses posées sur la prédictibilité de l'évolution : à savoir, comment la compréhension des phénomènes de convergence et de parallélisme évolutifs mènent à penser qu'ils existent des contraintes et mécanismes sous-jacents à l'évolution. La compréhension de ces mécanismes pourra mener à une prédiction plus fine de l'évolution. Notamment, l'approche des systèmes biologique par leur **composante métabolique** est ici considérée comme une pierre angulaire du raisonnement suivi, permettant d'élucider certaines contraintes qui façonnent l'évolution (aux niveaux fonctionnel et génomique) ainsi que la structure des communautés.
Puis, le troisième chapitre de cette section précise le contexte dans lequel ce projet de recherche sur la prédictibilité de l'évolution se développe : celui d'une évolution de la **spécialisation** d'organismes initialement généralistes (ici les bactéries) à des contraintes environnementales particulières.

La seconde section s'appesantit sur l'importance des **interactions** entre organismes comme contrainte évolutive. Toujours en s'appuyant sur une perception métabolique des systèmes, les questions suivantes sont abordées : 1) comment ces interactions 'métaboliques' entre organismes (à l'échelle de la population ou de la communauté) peuvent influencer les patterns d'évolution ; et 2) la possibilité de prédire l'émergence de ces patterns d'évolution en se basant sur les interactions potentiellement existantes au sein d'une communauté.
A cet effet, le premier chapitre de cette seconde section présente une extension intéressante de l'Hypothèse de la Reine Noire (The Black Queen Hypothesis, Morris et al 2012) qui à été publiée en tant que mini review dans le journal ISME en 2016 (Mas et al, 2016). Cet article présente une approche de modélisation basée agent qui considère une amélioration de fitness acquise par un organisme lorsque son métabolisme est modifié (mutants ayant perdu certaines fonctions) et qu'ils développent une dépendance à d'autres organismes de la population pour compenser ces fonctions particulières perdues. En analysant *in silico* l'augmentation d'une dépendance métabolique, nous avons cherché à explorer les états stables atteints par les populations mutantes et non mutantes sous

différentes contraintes. Ce travail permet une nouvelle compréhension de l'Hypothèse de la Reine Noire.
Le second chapitre de cette section présente les résultats préliminaires d'une étude considérant le métabolisme des organismes comme l'interface des interactions entre ces organismes, de la même espèce ou d'espèces différentes. En reconstruisant le métabolisme d'une communauté, les fonctions métaboliques réalisées à l'échelle de la communauté microbienne ont étés comparées avec l'ensemble des fonctions métaboliques qui pourraient être réellement réalisées par les individus séparément. Il est apparu que certaines fonctions ne peuvent être réalisées par des génomes uniques, c'est-à-dire par des organismes isolés. De telles fonctions, qui ne sont soutenues dans leur intégralité par aucun organisme, suggèrent fortement que leur réalisation dépend de l'interaction et de l'association du métabolisme de plusieurs organismes au sein de la communauté. Qui plus est, ces analyses préliminaires de données soutiennent que ces fonctions accomplies au niveau de la communauté sont associées à de fortes cooccurrences de micro-organismes particuliers.

La troisième section du manuscrit présente l'essence expérimentale de la thèse, sous forme de trois chapitres complémentaires. Ces trois chapitres correspondent aux étapes d'un workflow appelé 2STEP présenté plus en détail ci-dessous.
Le premier chapitre présente de façon approfondie le raisonnement mis en place pour l'utilisation de l'analyse de flux métabolique comme outil de prédiction de l'évolution au niveau génétique. En effet, l'une des stratégies développées pour étudier l'évolution de la spécialisation invoque la vision métabolique du système présentée ci-dessus. Le métabolisme est à l'interface des conditions environnementales d'un organisme et du fonctionnement de cet organisme. C'est l'un des premiers éléments impactés par un changement d'environnement, et donc considéré comme un élément clé de l'adaptation et de l'évolution. De plus, grâce aux descriptions récentes des processus génétiques, il est possible de lier étroitement les gènes d'un organisme à son fonctionnement métabolique. L'interaction entre ces deux composantes est à l'origine de prédictions potentielles: l'environnement affecte le métabolisme qui peut être relié au fonctionnement des gènes.
Est ensuite expliqué comment, à partir du modèle métabolique de la bactérie *Pseudomonas fluorescens* Pf0-1, des analyses de flux métaboliques (flux balance and flux variability analyses) sont réalisée pour déterminer 1) les conditions *in silico* et *in vitro* du milieu, et 2) déterminer l'activité des différentes réactions métaboliques selon leurs statuts (avec une expression des réactions plus ou moins obligatoire, alternative ou exclue) et la variabilité de leurs flux (flux span et CV2). A l'issu de cette modélisation, l'identité des gènes impliqués dans ces réactions est retrouvée, et des hypothèses évolutives sont faites sur l'avenir évolutif de ces gènes en fonction de l'activité prédite des réactions métaboliques qu'ils déterminent (expression des gènes conservée, réduite ou augmentée).
Le second chapitre de cette section présente les deux expériences d'évolution *in vitro* réalisées en continue, sur plus de 150 générations. Une première expérience ou une population de bactéries *P. fluorescens* Pf0-1 se spécialise à un milieu où la source de

carbone unique est du glucose, et la seconde expérience où la même population initiale se spécialise à un environnement où la source de carbone unique est le glycérol. Le raisonnement et la mise en place de chaque expérience est détaillée.
Suite à ces expériences, des analyses génomiques de l'échantillonnage issues d'un séquençage de l'ADN de la population de *Pseudomonas fluorescens* (suivi de l'évolution sur 5 pas de temps) permettent de détecter l'émergence de mutations ponctuelles (single nucleotide polymorphisme) *de novo* et de suivre la dynamique de ces mutations dans la population au cours du temps évolutif. Les mutations détectées et leur fitness sont interprétées à la lumière du fonctionnement du métabolisme. Une partie de ces mutations sont analysées en détail pour en déterminer l'effet potentiel sur le fonctionnement des individus qui les portent (présentation de l'outil informatique GetoPe développé a cette fin). Finalement ce chapitre présente aussi la comparaison entre les mutations observées et les prédictions faites par modelisation métabolique au préalable. Cette comparaison est néanmoins limitée de par la complexité du réseau d'expression génétique et de la faible proportion de gène métabolique sur lesquelles la prédiction ont pues être faites.
Le troisième chapitre de la section permet la caractérisation phénotypique des populations évoluées et non évoluée, afin de déterminer et quantifier certaines modifications évolutives. Les traits phénotypiques testés étaient ceux de la mobilité (vélocité, difusibilité, trajectoires, tumble time) et de la chémotaxie face à différentes sources de carbone. Les résultats obtenus montrent de claires différences entre la population initiale (non évoluée) et la population évoluée, pour chaque type de test. Néanmoins les tendances des résultats observés sont parfois contre-intuitives et ouvrent sur de nouvelles questions de recherche.
La section se termine par une représentation graphique synthétique du workflow 2STEP, qui intègre l'ensemble de la réflexion et des expériences/analyses menées à bien.

Finalement la quatrième section de ce manuscrit reprend brièvement dans une discussion générale le raisonnement et les grandes lignes de résultats trouvés dans chacune des parties avant de s'étendre sur les perspectives conceptuelles et expérimentales qui pourraient permettre de développer de nouvelles lignes de recherche basées sur ce travail, avec notamment l'étude de l'expression génétique (ARN, épigénétique) au cours de l'évolution et la possibilité de modifier génétiquement, selon les gènes ciblés par la modelisation, les organismes avant expérimentation. Cette partie traite également des ponts construits entre la génétique des populations, le fonctionnement génomique et le fonctionnement écologique des systèmes vivants en combinant les approches *in silico* et *in vitro*. Il souligne finalement les portes ouvertes par cette approche multidisciplinaire pour répondre à de nouvelles questions sur l'évolution.

Pour conclure, ce travail exploratoire et interdisciplinaire augmente les connaissances sur les déterminants évolutifs et les trajectoires suivies par l'organisme au cours d'un phénomène de spécialisation. Il démontre également un grand potentiel de prédiction, notamment grâce à une perception métabolique des systèmes.