Mathematical foundations of antibody affinity maturation

Irène Balelli

UNIVERSITÉ PARIS 13 | USPC Université Sorbonne Paris Cité

Université Paris 13 - Villetaneuse

Institut Galilée

Laboratoire d'Analyse, Géométrie et Applications

THÈSE

Présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PARIS 13

Spécialité: Mathématiques Appliquées

par

Irene Balelli

Fondements mathématiques de la maturation d'affinité des anticorps

Mathematical foundations of antibody affinity maturation

Soutenue le 30 Novembre 2016 devant la Commission d'examen:

M.	Julien	Berestycki	Rapporteur
M.	Jean-François	Delmas	Examinateur
M.	Vuk	Milišić	Co-directeur
M.	Thierry	Mora	Rapporteur
M.	Khashayar	Pakdaman	Examinateur
Mme.	Nadine	Varin-Blank	Examinateur
M.	Gilles	Wainrib	Co-directeur
M.	Hatem	Zaag	Directeur

UNIVERSITÉ PARIS 13 | USPC Université Sorbonne Paris Cité

Université Paris 13 - Villetaneuse

Institut Galilée

Laboratoire d'Analyse, Géométrie et Applications

THÈSE

Présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PARIS 13

Spécialité: Mathématiques Appliquées

par

Irene Balelli

Fondements mathématiques de la maturation d'affinité des anticorps

Mathematical foundations of antibody affinity maturation

Soutenue le 30 Novembre 2016 devant la Commission d'examen:

M.	Julien	Berestycki	Rapporteur
M.	Jean-François	Delmas	Examinateur
M.	Vuk	Milišić	Co-directeur
M.	Thierry	Mora	Rapporteur
M.	Khashayar	Pakdaman	Examinateur
Mme.	Nadine	Varin-Blank	Examinateur
M.	Gilles	Wainrib	Co-directeur
M.	Hatem	Zaag	Directeur

Abstract

The adaptive immune system is able to produce a specific response against almost any pathogen that could penetrate our organism and inflict diseases. This task is assured by the production of antigen-specific antibodies secreted by B-cells. The agents which causes this reaction are called antigens: during an immune response B-cells are submitted to a learning process in order to improve their ability to recognize the immunizing antigen. This process is called antibody affinity maturation.

We set a highly flexible mathematical environment in which we define and study simplified mathematical evolutionary models inspired by antibody affinity maturation. We identify the fundamental building blocks of this extremely efficient and rapid evolutionary mechanism: mutation, division and selection. Starting by a rigorous analysis of the mutational mechanism in Chapter 2, we proceed by successively enriching the model by adding and analyzing the division process in Chapter 3 and affinity-dependent selection pressures in Chapter 4.

Our aim is not to build a very detailed and comprehensive mathematical model of antibody affinity maturation, but rather to investigate interactions between mutation, division and selection in a simplified theoretical context. We want to understand how the different biological parameters affect the system's functionality, as well as estimate the typical time-scales of the exploration of the state-space of B-cell traits.

Beyond the biological motivations of antibody affinity maturation modeling, the analysis of this learning process leads us to build a mathematical model which could be relevant to model other evolutionary systems, but also gossip or virus propagation. Our method is based on the complementarity between probabilistic tools and numerical simulations.

Résumé

Le système immunitaire adaptatif est capable de produire une réponse spécifique contre presque tous le pathogènes qui agressent notre organisme. Ceci est du aux anticorps qui sont des protéines sécrétées par les cellules B. Les molécules qui provoquent cette réaction sont appelées antigènes : pendant une réponse immunitaire, les cellules B sont soumises à un processus d'apprentissage afin d'améliorer leur capacité à reconnaître un antigène donné. Ce processus est appelé maturation d'affinité des anticorps.

Nous établissons un cadre mathématique très flexible dans lequel nous définissons et étudions des modèles évolutionnaires simplifiés inspirés par la maturation d'affinité des anticorps. Nous identifions les éléments constitutifs fondamentaux de ce mécanisme d'évolution extrêmement rapide et efficace : mutation, division et sélection. En commençant par une analyse rigoureuse du mécanisme de mutation dans le Chapitre 2, nous procédons à l'enrichissement progressif du modèle en ajoutant et analysant le processus de division dans le Chapitre 3, puis des pressions sélectives dépendantes de l'affinité dans le Chapitre 4.

Notre objectif n'est pas de construire un modèle mathématique très détaillé et exhaustif de la maturation d'affinité des anticorps, mais plutôt d'enquêter sur les interactions entre mutation, division et sélection dans un contexte théorique simplifié. On cherche à comprendre comment les différents paramètres biologiques influencent la fonctionnalité du système, ainsi qu'à estimer les temps caractéristiques de l'exploration de l'espace d'états des traits des cellules B.

Au-delà des motivations biologiques de la modélisation de la maturation d'affinité des anticorps, l'analyse de ce processus d'apprentissage nous a amenée à concevoir un modèle mathématique qui peut également s'appliquer à d'autres systèmes d'évolution, mais aussi à l'étude de la propagation de rumeurs ou de virus. Notre travail théorique s'accompagne de nombreuses simulations numériques qui viennent soit l'illustrer soit montrer que certains résultats demeurent extensibles à des situations plus compliquées.

Remerciements

Je tiens tout d'abord à remercier les rapporteurs de ma thèse et tous les autres membres du jury de vous être intéressés à mon travail de recherche : j'en suis profondément honorée. Je suis certaine que vos remarques vont donner un autre regard sur cette thèse et seront pour moi de grande inspiration.

Je dois ensuite un remerciement spécial à mes directeurs. Tout d'abord merci à tous les trois, Vuk, Gilles et Hatem, pour avoir cru en ce projet et en mes capacités pour le mener à terme. Vuk et Gilles, depuis notre première collaboration lors de mon stage de fin d'études, l'idée de travailler ensemble à une modélisation de la maturation d 'affinité des anticorps m'a tout de suite passionné pour différentes raisons. Ses enjeux, la richesse de la problématique et les questions mathématiques auxquelles cela allait sûrement nous mener en sont les principales. J'ai beaucoup apprécié la façon dont on s'est approché de ce problème très complexe et dont j'ignorais encore plein de facteurs : nous avons d'abord examiné le problème biologique et comment il avait déjà été modélisé. Nous avons choisi ensuite notre point de vue personnel, qui impliquait certainement une forte simplification du problème de départ, mais qui à notre avis avait aussi de grandes potentialités. J'espère avoir pu le montrer au cours de ma thèse. Les discussions qu'on a eues au cours de ces années ont toujours été très enrichissantes pour moi, avec beaucoup d'échanges et de nouvelles idées à chaque fois, nous suggérant d'autres chemins à explorer. Vous avez toujours su respecter mes points de vue et valoriser mes idées, en me donnant des conseils précieux pour m'améliorer et résoudre les problèmes qui se présentaient au fur et à mesure. Merci aussi pour m'avoir toujours poussé à voir un petit peu plus loin. Tout cela m'a donné confiance et envie de poursuivre dans ce passionnant monde de la recherche, et je vous en remercie. Hatem, merci pour ton soutien et ta grande gentillesse et disponibilité : tu m'as donné beaucoup humainement et j'ai une forte et sincère estime pour toi. Je te suis très reconnaissante d'avoir accepté d'être mon directeur de thèse et de m'avoir soutenu au cours de ces années, en me posant toujours de bonnes questions pour réfléchir à mon futur.

Je vous remercie aussi tous les trois pour m'avoir introduite dans le labex Inflamex. J'ai adoré travailler dans ce contexte interdisciplinaire à l'interface entre Mathématiques et Biologie et avoir la possibilité d'apprendre d'un domaine pour moi inconnu et des gens, qui y travaillent et qui ont toujours montré une très grande ouverture d'esprit vis-à-vis de nous. Merci donc aux partenaires du labex Inflamex et en particulier à Nadine. J'ai trouvé très intéressant de pouvoir discuter avec des biologistes, qui, comme toi spécialement, ont eu la patience de répondre à nos questions et l'envie d'écouter nos idées, et d'en partager. Je vais sûrement continuer dans cette direction : j'estime avoir pris conscience pendant ces années de ce que j'aime faire et c'est grâce à vous.

Je veux ensuite remercier l'équipe de Modélisation et Calcul Scientifique qui m'a accueillie, et aussi l'équipe de Probabilités et Statistiques. Je remercie en particulier Clément, Isabelle et Yueyun, les responsables des principaux modules pour lesquels j'ai donné des TDs au cours de ces trois années. J'ai beaucoup apprécié de travailler avec vous et je considère l'expérience de l'enseignement très formative pour moi et complémentaire à la recherche, même si on se sent parfois découragé vis-à-vis de certains étudiants. Un remerciement particulier à Clément, avec qui j'ai beaucoup discuté des questions plus ou moins professionnelles : on va finalement se boire un coup !

Il y a beaucoup d'autres personnes que je veux remercier pour leur présence dans ma vie pendant ces années de thèse, quelques-uns même depuis très longtemps et d'autres que depuis quelques mois. J'espère n'oublier personne, mais ça va être compliqué !

Je veux remercier premièrement mes amis historiques, qui même avec les kilomètres qui nous séparent savent être toujours présent, et vers qui vont tout de suite mes pensées quand quelque chose arrive dans ma vie et qui m'écoutent toujours quand j'ai quelques chose à raconter ! À chaque fois qu'on se voit je ressens pour vous la même affection et en même temps j'aimerais tellement vous voir plus souvent ! Du coup, un gros merci à Silvia Al, à Silvia Amò, à Margo et à Marirosa. Vous me manquez !!

Un merci aussi aux amis niçois, Djé et Mathieu, et à Nina bien sûr (dans la section des niçois, même si hollandaise, tu peux me pardonner !). Merci pour les bons moments passés ensemble pendant ces années parisiennes, pour m'avoir aidé quand j'en ai eu besoin, et pour être devenus très naturellement des tontons quand on est devenu parents ! Et un merci de plus à Mathieu pour

avoir plusieurs fois pris du temps et de la patience pour m'aider à comprendre certains articles "un peu trop bio" : tu ferais un bon prof, je pense !

J'ai toujours trouvé une très bonne ambiance au LAGA, ce qui a beaucoup aidé à faire de ces trois années de thèse une expérience vraiment positive. Je vais d'abord remercier mes collègues du bureau A307 de ma première année : Giovanni cuore pesante, Thomas, Cuong et Roland, qui m'ont accueillie en premier au LAGA et que je revoie toujours avec grand plaisir ! Merci aussi à tous les autres que j'ai rencontré aux cours des années : Guillaume, Peppe, Didier, Daniel, Carlos, Mattia, Tom, Delphin, Marion, Annalaura (le chef des doctorants, mais aussi une amie, et une personne que j'estime beaucoup et que j'apprécie pour son humour et son intelligence, et avec qui on peut rigoler et parler des choses les plus sérieuses, très naturellement), Bruno, Pierre, Eva, Liza, Julien, Nicolas. J'ai été ravie de vous connaître et j'ai bien apprécié les moments qu'on a partagés, les repas et les cafés, les discussions sérieuses et celles qui virent rapidement vers l'absurde, les conseils, les séminaires et aussi les petits gâteaux et les apéros ! Merci aussi à Ian pour nos conversations et ses conseils, pour son optimisme et pour sa disponibilité. Un grand merci à Alberto et à ses mises en perspectives des mes problèmes. Un remerciement spécial pour Michele (avec une seule "l", pour cette fois), qui a été mon point de repère à Paris, même s'il se perd dans 40 m^2 ! Tu es vraiment une très belle personne, toujours disponible et pleine d'attentions pour tous ! Je te souhaite tout le mieux dans le futur et j'espère continuer à en faire partie !

Merci aussi aux autres copines et copains avec qui j'ai partagé des buts de ma vie au cours de ces années, en particulier à Ele Leo, qui a rapidement pris une place dans ma petite famille avec son énergie et franchise ! Et merci aussi à Filomena, toujours vivante et positive, qui va bien nous manquer l'année prochaine !

Merci à ma belle famille, bruyante et pleine d'amour : j'ai de la chance d'avoir une place parmi vous ! Un merci spécial à Hervé et Béa, pour leur soutien et encouragement, et leur joie de partager. Et une pensée à Pascale, qui nous a quitté trop tôt, mais qui a bien laissé ses marques.

Je veux ensuite remercier ma famille (belle aussi !), et en particulier mes parents, qui me soutiennent toujours et m'encouragent, et qui sont simplement toujours là. J'aime mon indépendance mais j'aime aussi mes racines. Vous m'avez donné une base solide qui m'a permis de choisir et de construire, et aussi de partir, mais sans quitter.

J'ai laissé en dernier les remerciements aux deux personnes qui signifient le plus pour moi : mon copain et mon fils. Vous êtes ma première et dernière pensée, ma joie à la fin d'une très mauvaise journée, et aussi d'une très belle ! Vous me donnez l'envie de faire des projets, et la détermination nécessaire pour les poursuivre. Vous me donnez les plaisirs de profiter de la vie et la force de faire face aux moments difficiles. Et je vous aime.

Contents

Chapter 1

Introduction

This thesis is devoted to the construction and the study of a simplified mathematical evolutionary model of antibody affinity maturation. Our strategy consists in analyzing and successively coupling fundamental building blocks of this learning process: mutation, division and selection, that we study through a rigorous mathematical analysis.

Antibody affinity maturation is a key process in adaptive immunity, leading to the production of high-affinity antibodies upon immunization. This task is assured by B-cells, special lymphocytes which are activated by the encounter with an antigen and then directed through the peripheral lymphoid follicles. There they give rise to germinal centers, transient high specialized micro environments in which they undergo multiple rounds of mutation, division and selection. Once B-cells have improved their affinity with respect to the presented antigen, they successfully complete the germinal center reaction and differentiate into memory or plasma B-cells.

B-cell antigen-dependent affinity maturation is a key mechanism of adaptive immunity. Perturbations or malfunctions in this mechanism lead to various pathologies. One of them is the Chronic Lymphocytic Leukemia (CLL), the starting point of our project. CLL is a disease derived from antigen-experienced B-cells that differ in the level of mutations in their receptors [31]. It is the commonest form of leukemia in the Western world, with an incidence of $4.2 : 100000$/year, increasing up to more than $30 : 100000$/year among people older than 80 years [44]. In CLL, leukemia B-cells can mature partially but not completely, and survive longer than normal cells, crowding out healthy B-cells. Even if major progresses have been made in the identification of molecular and cellular markers predicting the expansion of this disease in patients, the pathol-

ogy remains incurable [40, 44]. Understanding how the immune system works in a healthy individual would certainly provide suggestions about the causes that lead to CLL, and motivation for further research on possible treatments.

Beside this initial motivation, improving our knowledge of the functioning of immune system is one of the fundamental research axes both in Biology and in Medicine, equally from a physiopathological (*e.g.* autoimmune diseases) and therapeutical (*e.g.* vaccination, immunotherapy) points of view. In the last few decades immunotherapy has become an important part of treating some types of diseases such as cancers. The development of these treatments has been possible thanks to the spectacular advances in our understanding of adaptive immunity over the past 30 years. Immunotherapy consists in the treatment of diseases either by stimulating the patient's immune system to work harder or smarter, or by giving to the immune system extra components, such as artificially synthesized proteins. There already exists a variety of strategies in this direction, new immune treatments are now under investigation and may impact cancer treatment in the future. One can think for instance to immune checkpoint therapies [84], or to adoptive cell therapies [113]. Their development is extending and saving lives of thousands of patients suffering from cancer. Moreover, since they are highly personalized therapies, they offer the promise of high specificity and safety [118], having significantly fewer side effects than existing drugs. Immunotherapies have been shown to be really promising also for the treatment of other diseases, such as autoimmune diseases or allergic asthma, the commonest form of asthma, which still causes significant morbidity (and sometimes mortality), particularly in the pediatric population [87].

Beyond the fundamental understanding of physiological processes and their associated pathologies, the study of directed evolution mechanisms at the heart of antibody affinity maturation have been inspiring many methods for the synthetic production of specific antibodies for drugs, vaccines or cancer immunotherapy [6, 79, 122]. Indeed, this production process involves the selection of high affinity peptides and requires smart methods to generate an appropriate diversity [34]. Besides the biomedical motivations, the study of this learning process has recently given rise to a new class of bio-inspired algorithms (*e.g.* [30, 107]), mainly addressed to solve optimization and learning problems [25].

The study of the immune system, their components and mechanisms, is therefore an important subject of intense investigation, from an experimental, medical and theoretical points of view. For this reason, we believe that it is important to establish solid mathematical foundations of this extremely com-

plicated biological process: this has still not been done rigorously, to our knowledge. Moreover, this would bring us to investigate interesting mathematical problems which go beyond initial modeling purposes. For instance, our analysis suggested to model the mutation-division process of B-cells in germinal centers as branching random walks on graphs, a type of branching processes which have not been deeply investigated so far, despite the growing number of applications in biological, chemical, physical and economical systems [90, 28, 29].

Chapter 1 details the biological background and gives a panorama of the existing models of germinal center reaction and antibody affinity maturation. It provides as well an overview on the main results obtained in this thesis.

Chapter 2 focuses on pure mutational models. We set the state-space of B-cell traits and define several mutational mechanisms on it. The aim of this part is to understand how the typical time-scales of state-space exploration change depending on the choosen mutational rule. Namely, for each rule, we derive explicit formulas to evaluate the expected hitting time to reach a specific configuration. This allows to compare the impact of the rule on the efficiency of antibody affinity maturation.

In Chapter 3 we introduce a branching process over the state-space of B-cell traits, modeling the division of B-cells. We apply the theory of expander graphs to establish results about the ability of different mutational rules to make the exponentially growing population fill the state-space of all possible B-cell traits. We observe an unexpected saturation phenomenon: increasing the mutation rate above a certain threshold has only marginal effects on the speed of state-space covering.

In Chapter 4, we study more comprehensive models including mutation, division, death and affinity-dependent selection mechanisms. We formalize these models by opportunely using multi-type Galton Watson processes. Investigating how the interaction of different parameters affects the system functionality, we identify an optimal selection rate which maximizes the production of output cells.

Finally in Chapter 5 we suggest some limitations and possible extensions of our models, providing motivation for further research.

Throughout the project we pursue three fundamental objectives:

i) we reflect upon the modeling assumptions and methods,

ii) we make a rigorous mathematical analysis of the objects that we introduce, which leads to new theoretical results. Then, we provide the corresponding biological interpretation,

iii) we perform for each Chapter extensive numerical simulations: on the one hand, they validate our theoretical results and, on the other, they conjecture how these results extend to cases which we are not able to study mathematically.

Each chapter is self-contained and can be read independently from the others. Chapters 2, 3 and 4 have been collected into three papers, [10, 11, 12] respectively.

1.1 The germinal center reaction

Antibody Affinity Maturation (AAM) is defined as the increasing of the average affinity of serum specific antibodies during the course of an immune response [132]. This is achieved through an evolutionary Darwinian process of B-lymphocytes, which takes place in Germinal Centers (GCs) in secondary lymphoid follicles.

The initiation and development of the GC Reaction (GCR) is assured by a coordinated cascade involving different cell types which move dynamically within and between GCs [36]. The GCR starts with the activation of B-cells after the encounter of an antigen. This encounter takes place in the secondary lymphoid organs, which include lymph nodes, the spleen and the mucosal-associated lymphoid tissue [126]. Here the antigen arrives either via blood or lymphatic vessels or transported by conventional dendritic cells (cDCs). All secondary lymphoid organs contain lymphoid follicles, which are critical for the functioning of the adaptive immune system. In the absence of an immune response to an antigen, the follicle appears as a primary lymphoid follicle, a loose

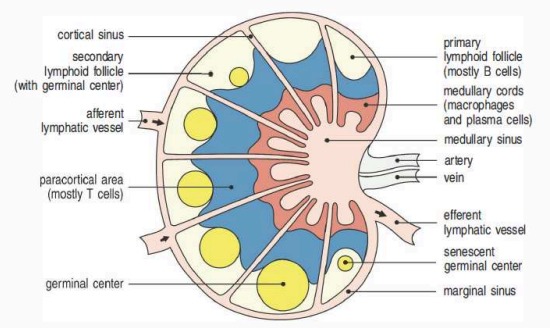
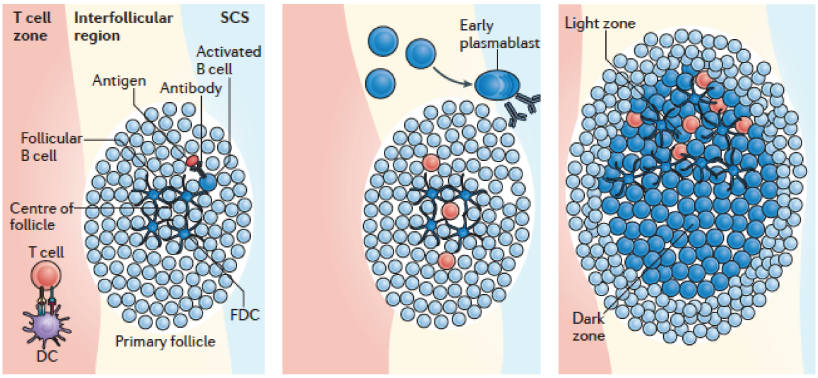


Figure 1.1: Organization of a lymph node (source [102])

of follicular dendritic cells (FDCs) embedded in regions rich in B-cells. The T-cell zone borders follicles. When the surface immunoglobulins (Igs) of a B-cell, called B-cell Receptors (BCRs), succeed in binding the antigen, the B-cell become activated. Simultaneously, also antigen-specific T-cells get activated thanks to the interactions with cDCs, which have processed the antigen and display it to T-cells as peptides in the context of major histocompatibility complex class II (pMHCII) [136]. Hence they start to proliferate. Activated B-cells upregulate chemokine receptor CCR7 and migrate toward the border of the T-cell zone. There they present antigen to T-cells via pMHCII molecules and compete for T-cells help. A successfully interaction with cognate T-cells promotes B-cell proliferation. At this stage, a subset of fully activated B-cells



(a) Day 0 **(b)** Day 3 **(c)** Day 7

Figure 1.2: Initiation of the GCR (source [36])

travels to the center of the follicle to seed the early GC, while other activated B-cells differentiate into early memory B-cells and short-lived plasma cells, and start to produce antibodies with a low affinity for the presented antigen. Recent evidence shows that ten to hundreds B-cells seed each GC [132]. Follicles with GCs are called secondary lymphoid follicles. Interactions between B and T-cells have effects over T-cells phenotype as well. Indeed, within 3 days post immunization most of activated T-cells differentiate into T follicular helper cells (Tfh cells) and migrate into the B-cell follicle. T-cells which are going to become Tfh cells start to upregulate B-cell lymphoma 6 (BCL-6), a transcriptional repressor and the master regulator of both B-cell and Tfh cell program during a GCR. BCL-6 is essential for the initiation of the GCR since it controls a transcriptional program which facilitates the migration of both B and Tfh cells to the center of the follicle. Tfh cells are characterized by high expression of CXC-chemokine receptor 5 (CXCR5) and programmed cell death protein 1 (PD1) [36, 136].

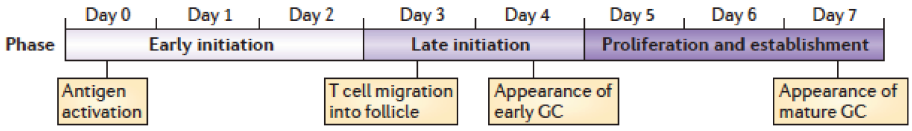


Figure 1.3: Initiation phase of the GC (source [36])

The GC size increases rapidly due to the intense proliferation of B-cells, which reaches thousands when mutations and affinity-dependent selections are turned on [148]. By day 7 after immunization, the GC is fully established and splits into two distinct microenvironments, named Dark Zone (DZ) and Light Zone (LZ). This polarization relies on gradients of the CXC-chemokine ligand 13 (CXCL13), produced in the LZ, and CXCL12, highly expressed in the DZ, whose receptors, CXCR5 and CXCR4, are upregulated in LZ and DZ B-cells respectively [137]. The DZ is characterized by densely packed proliferating B-cells within an interconnected network of reticular cells expressing CXCL12: DZ B-cells divide at a very high speed, since they can double every 6-7 hours. Moreover, DZ B-cells upregulate activation-induced cytidine deaminase (AID) and error-prone DNA polymerase η (Polη). This induces random Somatic Hypermutations (SHMs) of the binding region of BCRs, and Class Switch Recombinations (CSRs), which allows for switch between different Ig isotypes, each with a distinct effector function [133]. The LZ is more sparsely populated by B-cells and contains FDCs, Macrophages (TBM) and a high density of Tfh cells, whose role is to select B-cells based on affinity.

Each round of DZ B-cell division is predicted to produce 1 somatic mutation per 10^3 base pairs in Ig variable genes [55], which affects specific regions of Ig genes targeted in the local DNA sequence. This results in an increased differential mutability in the regions encoding the antigen binding domain, triggering the generation of mutant clones having a broad range of affinities for the immunizing antigen [74, 36]. Since SHM is mostly a random process, it can either increase affinity of the BCR for the presented antigen - a critical event in AAM - or cause loss of affinity and even lead to autoimmunity [74]. In particular, it has been estimated that about 50% of all mutations are silent, 30% lethal and lead to B-cell death and only 20% having an effect on affinity, which can be either of an improving nature, or of a worsening one [121, 148]. Therefore, stringent selection mechanisms acting on mutated clones are clearly needed.

This takes place in GC LZ, which is the site of two main B-cell developmental processes: the selection of B-cells that produce high affinity antibodies

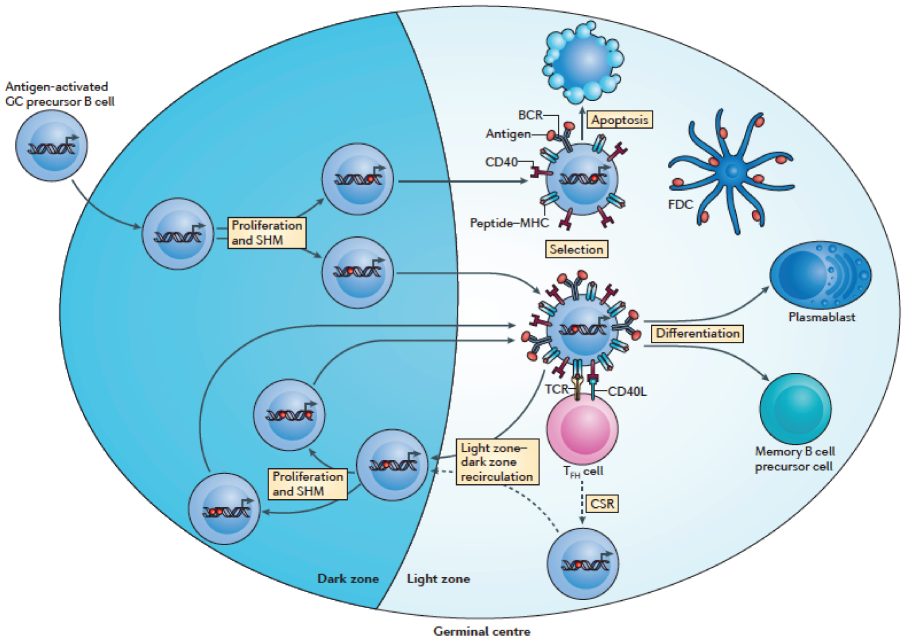


Figure 1.4: Dynamics of the GCR once mutations and affinity-dependent selections are turned on (source [36])

and the initiation of B-cells differentiation into plasma cells (*i.e.* antibody-secreting cells) and memory B-cells. Memory B-cells are antigen-experienced B-cells which assure a faster and more efficient response in case of a new encounter with the same antigen: they express high-affinity antibodies and can quickly differentiate into plasma cells in antigen-recall response. After dividing and mutating in the DZ, B-cells travel to the LZ: the transition from DZ to LZ is triggered in most or all B-cells, based on a timed, cell-intrinsic program. Within a period of 4 to 6 hours, about 50% of DZ B-cells switch their phenotype, presenting slightly higher levels of CXCR5, and lower levels of CXCR4 [55]: this allows them to escape the pull of chemokine CXCL12, hence pass to the LZ. Here the selection of B-cells for improved antigen binding depends on the ability of B-cells to first capture antigen held on FDCs through their surface Igs and then compete to present the processed antigen as pMHCII to cognate Tfh cells. B-cells with higher affinity succeed in capturing a greater amount of antigen from FDCs and consequently present higher densities of pMHCII. This allows them to successfully outcompete lower affinity B-cells in receiving signals from Tfh cells, such as CD40L binding to CD40 proteins expressed on the surface of B-cells. Indeed, Tfh cells form the largest and longest contact with those B-cells which are able to process the higher quantity of antigen to present as pMHCII. Moreover, since these contacts are transient, Tfh cells can

test many different LZ B-cells and progressively increase the strength of the selection pressure over GC B-cells [36]. Therefore Tfh cells have a crucial role in the selection of high-affinity antibodies. In addition, the positive selection of B-cells in GCs is fine tuned by antigen masking on FDCs via antibodies secreted by B-cells which have already differentiate into plasma cells. Since antibodies can infiltrate in neighboring GCs and Tfh cells can freely move between GCs, a coordination between several GCs can be achieved and contributes in improving AAM [36].

Lower-affinity B-cells that fail to receive proper selection signals from Tfh cells die by apoptosis and are rapidly cleaned by TBM: this mechanism eliminates not only B-cells which have lost antigen binding, but also those that have acquired autoreactive specificities [74]. Positive selected B-cells can either exit the GC differentiating into later plasma cells and memory B-cells, or re-enter the DZ upregulating CXCR4. In this case, they undergo further rounds of division and mutation. Apparently the differentiation of a GC B-cell into a plasma cell is driven by the acquisition of a high-affinity BCR and can be triggered by signals from Tfh cells. On the contrary, the differentiation process into memory B-cells seems to be stochastic, as throughout GCR, GC B-cells are constantly selected to enter the memory pool [102, 126]. LZ to DZ transition after positive selection signals is triggered in about 10 to 30% of high-affinity B-cells, and the magnitude of Tfh cells help provided in the LZ determines the behavior of the LZ B-cells when they reenter the DZ. Indeed, recent evidence [55] suggests that the number of B-cell divisions per DZ cycle is variable (from 1 to 6), and proportional to the strength of B-Tfh-cell interaction in the LZ. Therefore, higher-affinity B-cells gain a proliferative advantage leading them to dominate the GC B-cell population [74]. Moreover, since each cell division is associated with mutations of the Ig genes, the finding that Tfh cells regulate the number of division cycles in the DZ suggests that they also regulate SHM [55].

AAM is therefore achieved by multiple rounds of division and random SHM in the DZ followed by a Darwinian competition for Tfh cells help in the LZ, which selects B-cells with increasing affinity for the presented antigen. Recirculation between the two zones, in which B-cells alternate distinct genetic programs, facilitate the production of high-specialized antibodies, essential for the effectiveness of the immune response [136, 36, 132, 55]. The GCR reaches its peak within approximately 2 weeks [144] then after about 3 weeks the GC begins to dissipate and disappears in a time which can vary greatly, passing from a few days to several weeks.

Although substantial progress has been made in adaptive immunity and the key dynamics of GCR are now well characterized [89, 36, 55, 132], there are still facts that remain unclear, owing the absence of experimental data. For instance the precise mechanisms which govern the selection of LZ B-cells, and whether cell division and mutation are regulated are still unknown [14]. Moreover very little is known about which factors determine GC termination [144], and many hypotheses are currently debated through various mathematical models. GC termination can be due to progressive decrease of available antigen through antigen consumption or masking on FDCs [77, 130], or caused by an increasing differentiation of GC B-cells into output cells as a consequence of an increasing signaling by FDCs and Tfh cells [100].

1.1.1 Focus on the BCR structure, generation and diversification through SHM

Igs present at the antigen (Ag) receptor are Y-shaped membrane-bound macro proteins composed of four polypeptide chains assembled by disulfide bonds: two identical heavy (H) chains and two identical light (L) chains. They are encoded in humans by the genes IGH and IGL (or IGK), respectively [61]. Each chain consists of two regions: the C-terminal stem (Constant, or C region) specifies the effector functions of the molecule, while the N-terminal prongs (Variable, or V region), composed by the variable parts of the two chains together (V_H and V_L respectively), specify the Ag-recognizing capacity [133]. There exist five different classes of Igs (IgM, IgD, IgG, IgA and IgE), which can be distinguished by their C regions. Their corresponding heavy chains are denoted by μ, δ, γ, α and ε respectively, and they confer to an antibody its effector function. More subtle differences confined to the V region account for antigen binding specificities. Mature plasma B-cells can secrete their Igs as antibodies [102].

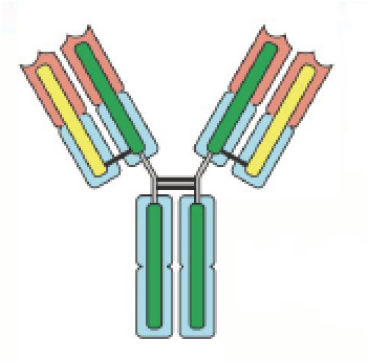


Figure 1.5: Schematic structure of antigen receptors (source [102]). The heavy chains are in green and the light chains are in yellow. Each chain has a constant part (shaded blue) and a variable part (shaded red)

The DNA encoding for the antigen-combining portion of the antibody, hence defining the initial diversity of the BCR repertoire, is the result of a

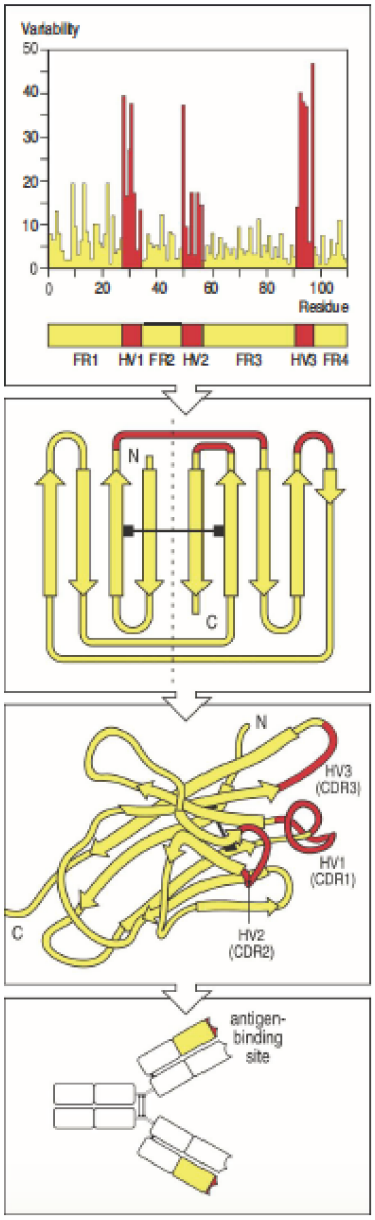


Figure 1.6: The hypervariable regions of a light chain (source [102])

somatic recombination process called V(D)J recombination. This process brings together one each of the variable (V), diversity (D), and joining (J) segments of the IGH locus to form the heavy chain Ig gene, and one each of the V and J segments of the IGL (or IGK) locus to form the light chain [61]. Moreover, additional sequence diversity is generated by random deletion or insertion of nucleotides at segment junctions, called junctional diversity. Before being released from the bone marrow into peripheral blood, the generated naive B-cells which do not contain nonproductive (*e.g.* out-of-frame) coding sequences, undergo an initial round of selection for lack of self-reactivity [102]. At this step the rearrangement process is able to generate between 10^5 and 10^6 different antibody specificities [38]. The V regions of any BCR differ from those of every other B-cell. Nevertheless the sequence variability is not distributed uniformly, but rather concentrated in three hypervariable segments (HV1, HV2, HV3), which have been identified both in the V_H and V_L. They are more commonly called Complementary Determining Regions (CDRs), because they determine the antigen-binding site and the surface they form is complementary to that of the antigen they bind. The regions between the CDRs are more conserved regions termed framework regions (FWRs): they provide structural support [102, 61].

In this context complementarity means that the amino-acids composing the antigen-binding site or paratope and the antigenic determinant or epitope are distributed in such a way to form bonds which are able to hold the antigen to

the B-cell. These bonds are all non-covalent, thus by their nature reversible. Multiple bonding between the antigen and the B-cell ensures that the antigen is bound tightly to the B-cell. The interaction between paratope and epitope can be characterized in terms of a binding affinity, that will be proportional to their complementarity. The *affinity* is the strength of the reaction between a single antigenic determinant and a single combining site on the B-cell: it summarizes the attractive and repulsive forces operating between the antigenic determinant and the combining site of the B-cell, and corresponds to the equilibrium constant that describes the antigen-B-cell reaction [1, 141, 80]. Most antigen-antibody interactions involve at least one electrostatic interaction, which occurs between charged amino-acid side chains. Moreover, interactions can also occur between electric dipoles, as in hydrogen bonds, or can involve short-range Van der Waals forces. Finally, hydrophobic interactions can also participate to the antigen-antibody binding: they occur when two hydrophobic surfaces come together to exclude water and their strength is proportional to the surface area hidden from water [102].

Noncovalent forces	Origin	
Electrostatic forces	Attraction between opposite charges	$-NH_3^{\oplus}$ $^{\ominus}OOC-$
Hydrogen bonds	Hydrogen shared between electronegative atoms (N, O)	$>N^{\delta-}-H^{\delta+}--O^{\delta-}=C<$
Van der Waals forces	Fluctuations in electron clouds around molecules polarize neighboring atoms oppositely	$\delta^+\ \delta^- \rightleftharpoons \delta^-\ \delta^+$
Hydrophobic forces	Hydrophobic groups interact unfavorably with water and tend to pack together to exclude water molecules. The attraction also involves van der Waals forces	H_2O δ^+ δ^- δ^- δ^+ $O<H_2$

Figure 1.7: The non-covalent forces which determine antigen-antibody interactions (source [102])

The surface of an antigen presents variable motifs that B-cells, through their receptors, can discriminate as distinct epitopes. If we define an epitope by its spatial contact with a BCR during binding, the number of relevant amino-acids is approximately 15, and among these amino-acids only around 5 in each epi-

tope strongly influence the binding. These strong sites may contribute about one-half of the total free energy of the reaction, while the other amino-acids influence only marginally the binding strength, or even have no detectable effect. Simultaneously, a BCR contains a variety of possible binding sites and each antibody binding site defines a paratope: about 50 variable amino-acids make up the potential binding area of a BCR. In agreement with the above, only around 15 among these 50 amino-acids physically contact a particular epitope: these define the structural paratope. Consequently, antibodies have a large number of potential paratopes as the 50 or so variable amino-acids composing the binding region define many putative groups of 15 amino-acids [80].

The V(D)J recombination, which is responsible for the initial antibody repertoire of B-cells, takes place in the bone marrow without interactions of B-cells with antigens. Even if this primary repertoire is large, it does not suffice to face all possible antigens that the immune system could encounter during an individual lifetime. Hence B-cells undergo a second phase of diversification when they get activated after the encounter with an antigen. This is achieved through SHM during the GCR. SHM incorporates point mutations in the recombined V(D)J exon of the heavy and light chain encoding genes to enhance the affinity of the antibody to specific antigens.

The genetic code is a sequence of four nucleotides, guanine (G), adenine (A) (called purines), thymine (T) and cytosine (C) (pyrimidines), joined together. They make three-letter words: the codons. Each codon corresponds to a specific amino-acid or to a stop signal, which interrupts the building of the protein during translation. Different kind of genetic mutations can affect the DNA sequence of a gene. They can be regrouped in three main categories: base substitutions, insertions and deletions. A single base substitution is a switch of a nucleotide with another. This is the simplest kind of mutation and it can turn out to be missense, nonsense or silent, once we observe the resulting new protein. We said that a mutation is missense if the result of the genetic mutation is a different amino-acid in the protein. The mutation is nonsense when the genetic mutation results in a stop codon instead of an amino-acid. Finally, a silent mutation is a mutation with no effect on the amino-acid string, *i.e.* the mutated sequence codes for an amino-acid with identical binding properties. We talk about insertion (resp. deletion) when one or more nucleotides are added (resp. removed) at some place in the DNA code.

SHM is driven by an enzyme called activation-induced cytidine deaminase (AID) which is expressed specifically in this case. AID was classified into

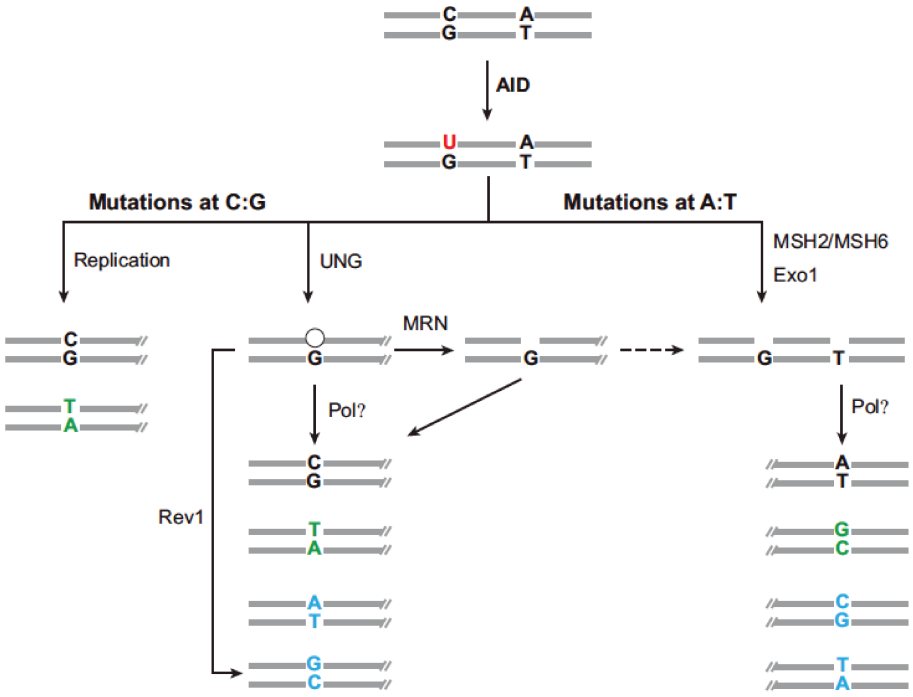


Figure 1.8: DNA deamination model of SHM (source [133])

the APOBEC family of polynucleotide cytidine deaminases, which perform hydrolytic deamination of Cytosine (C) to Uracil (U) [133]. This protein can bind to single-stranded DNA (ssDNA) only. Thus it seems to target only genes being transcribed (for which the transcription phenomenon separates temporarily double stranded DNA into small portions of two ssDNA sequences) [71]. The SHM process uses the initiating U:G lesion to activate natural DNA repair pathways for mutagenic purposes. In particular, two mechanisms tend to repair lesions in the DNA caused by these substitutions of C by U [115]:

a) either *mismatch repair* : substitution for the damaged zone by another sequence of nucleotides thanks to proteins MSH2/MSH6. The U base is read as T leading to a transition from a C:G pair to T:A.

b) or *base excision repair* : U is excised by a successive action of uracil-DNA glycolase (UNG) and apurinic/apyrimidinic endonuclease (APE1). The DNA contains then a nick, after replication, a random nucleotide is inserted in order to fill the vacant space leading to transversions and transitions.

Additionally, Polη, a Y-family polymerase, specifically targets A/T, A motifs and play a major role in generating A:T mutations [53, 147].

From \ To		Pyr T	Pyr C	Pur G	Pur A	
Pyr	T		7	4	2	13
	C	16		2	4	22
Pur	G	7	6		15	28
	A	3	12	22		37

Figure 1.9: Estimated frequency of C:G and A:T mutations (source [38])

SHM introduces single nucleotide substitutions in a stepwise manner at a frequency of around 10^{-3} per base pair per generation. All four bases can be targeted for mutations: C:G pairs and A:T pairs are targeted with approximately equal frequencies [38]. Although the process of SHM appears to be stochastic, there are clear intrinsic biases [127]. Indeed it is strongly context-dependent, since it has been observed that mutation rates vary more than 10-fold across sites [45]. In particular, substitutions occur at higher rates in hot spots motives like DGYW/WRCH where (G:C is the mutable position and D ∈ {A,G,T}, H ∈ {A,C,T}, R ∈ {A,G}, W ∈ {A,T} and Y ∈ {C,T}, and the underlined letters are the loci of mutations) [112, 63].

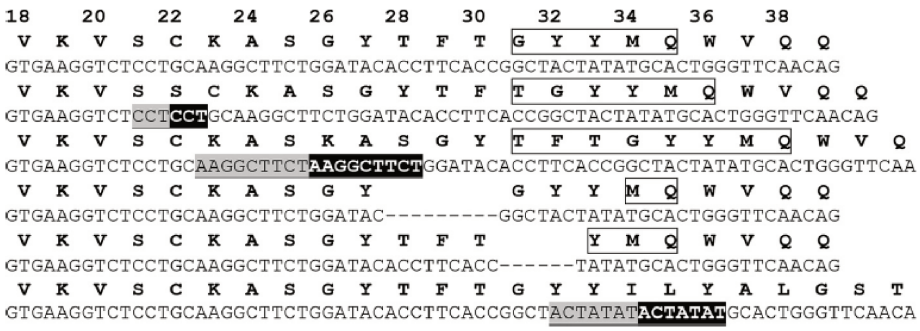


Figure 1.10: Examples of indels observed for a single antibody heavy chain. On the top the starting amino-acid chain is shown. Five in-frame indels and one indel producing a frameshift are then displayed. CDR1 regions are highlighted with a box. Deletions are indicated by dashes, while for insertions, the upstream sequence is shown in underlined light gray type, and the resulting duplicated sequence is highlighted in black (source [26]).

SHM introduces mostly single nucleotide exchanges. Nevertheless, small deletions and duplications, *i.e.* the insertion of extra copies of a portion of genetic material already present within the DNA code [63, 26, 27], have also been observed. *In vivo* studies of derived Ig genes have highlighted the importance of nucleotide insertions and deletions (indels) in affinity maturation; together with their contribute to the diversification of the antibody repertoire [26]. In-

dels generated during SHM of activated B-cells are associated to hotspots and localized predominantly in CDRs. It has been estimated that the frequency of indels mutations in circulating B-cells is up to 6.5%. The majority of *in vivo* in-frame indels mutations are short, with $\sim 90\%$ being of at most 3 amino-acids [26], and none of more than 9 amino-acids.

1.2 Mathematical modeling of AAM, an overview

GCs represent a typical example of a highly dynamic biological system, in which various coupled reaction processes occur in a spatially compartmentalized microenvironment, involving the contributions of different cell types and chemokine gradients [49]. The interactions among all such components are extremely intricate and not fully understood. One of the main goals of mathematical modeling is to identify and characterize the main mechanisms, as well as the interactions among the elementary components involved in a GCR, in order to deduce the generic macroscopic properties and features of the system [108]. Understanding the basic functional and physical principles of GC kinetics is not only important in medical science, but it also contributes to the fundamental understanding of molecular evolution [148, 103]. Indeed the immune system is faced to the challenge of producing high-affinity antigen-specific antibodies from initial low affinity precursors: its strategy is the same followed by germline evolution to produce novel proteins, which is an iterative alternation of mutation, clonal expansion and selection [103]. While germline evolution takes millions of years to be achieved, AAM needs only a few weeks to improve of ~ 100 fold the initial affinity of naive B-cells for the target antigen, representing an example of an extremely efficient and rapid evolutionary mechanism. Hence the study of GCR could also enhance our understanding of population dynamics in evolution.

As we have already underlined in Section 1.1, the key dynamics and main components of GCR are now well characterized and understood thanks to the combined effort of cellular and molecular biologists and immunologists. Nevertheless, there are still facts that remain unclear and which can not been elucidated via *in vivo* experiments. Indeed it is still very hard to follow and sequence each B-cell at any time within a single GC in order to gather precise phylogenetic data of the B-cell repertoire during a GCR. Similarly, it is really difficult to have precise spatial and temporal data about lymphocytes within the GC during an immune response, or to understand the exact dynamic of mutation and selection of B-cells while they are submitted to AAM [98, 42]. Mathematical modeling has already played an important role contributing to improve our understanding of the GC kinetics and AAM. Since it allows to capture the

global coordinated behavior of the GCR in a simplified way, it can eventually suggest how certain interactions among cells and molecules could lead to the experimentally observed results [108]. This suggestions can give rise to *in vivo* experimentation and lead to new insights. One can think, for example, to [75]: there T. B. Kepler and A. S. Perelson suggested for the first time the hypothesis of the existence of a recycling mechanism of B-cells during GCR after positive selection signals. Other examples are given by [96, 93], where M. Meyer-Hermann and coworkers predicted a dominant limiting role for Tfh cells to induce AAM. These mechanisms have now been confirmed by experiments [139, 119].

There exist many different possible approaches to conceive and study mathematical models of GCR and AAM. In [108] A. S. Perelson and G. Weinsbuch present an overview of several immunological problems which they formalize using physical concepts and mathematical methods. For instance they estimate the size of the immune repertoire and predict the size of epitopes by using probabilistic methods, or they propose a model of receptor cross-linking and affinity maturation. For the latter they have opportunely applied laws of mass action to define the concentrations of ligands, kinetic constants and Ordinary Differential Equation (ODE) systems. These are highly theoretical works with the objective of capturing some general features of the system. Similarly in [105] A. S. Perelson and M. Oprea describe the B-cell population in a typical GC as a result of dynamic interactions between mutation and selection. In particular, they develop a model of somatic mutation and B-cell expansion trying to understand from an optimal control perspective how the relatively few mutations that lead to high affinity antibodies are consistently observed. Following the dynamics of a single average GC, they propose that the optimal GCR is obtained by alternating cycles of expansion without mutation, followed by mutation and selection.

Other theoretical works investigate the problem of SHM and AAM as framed in the language of optimal control theory. For example, in [76] T. B. Kepler and A. S. Perelson have developed a single-compartment model for the process of AAM and an optimization algorithm based on the Pontryagin's maximum principle to find the optimal mutation schedule: the quantity to be maximized is the total affinity, which takes into account both the average affinity for the immunizing antigen and the number of B-cells involved in the response. Here again their results suggest that the optimal mutation schedule is one with brief bursts of high mutation rates interspersed between periods of mutation-free growth. They model mutations using a transition probability matrix over the state-space of possible Ig genotypes. In addition they overcome the highly complicated prob-

lem of specifying the binding affinity to a given antigen as a function of the Ig primary sequence, by defining affinity classes with respect to the presented antigen. This problem has further been discussed in [122], where the authors try to measure the similarities of amino-acid chains and then predict binding affinities by essentially using two tools: a similarity kernel on the set of fundamental amino-acids and a good amino-acid substitution matrix (*e.g.* BLOSUM62 [60]).

An interesting theoretical framework to study AAM, which shares some similarities with the one considered in Chapter 2, is given in [70] where S. A. Kauffman and E. D. Weinberger introduce the NK models. Amino-acid chains are represented as N length strings, and K corresponds to the number of sites whose state bears on the fitness contribution of each site. Hence the parameter K assures the richness of epistatic interactions among sites. When K increases with respect to N the affinity landscape passes from smooth and single peaked to jagged and multipeaked. They choose the hypercube vertex set as the basic structure to define the affinity landscape of BCRs. They assign to each node an affinity strength and perform adaptive random walks, biased with respect to the affinity gradient: a clone lying on a given node can jump to a neighbor node after mutation if the latter is fitter than the first one. They investigate the affinity landscape exploration trying to understand how it changes depending on the richness of epistasis.

In more recent years biologically very detailed models of GCs were proposed using, for instance, agent-based models (*e.g.* [92, 120, 94]), mostly analyzed through extensive numerical simulations. For example in [77] the authors focus on the dynamics of a single GC, investigating the impact of T cells on GC kinetics and termination. They allow for T-B-cell interactions and consider antigen consumption by LZ B-cells. Here and in [100] the major causes of GCR termination are investigated: this is still not fully understood. Two main hypotheses arise from these papers: a lack of antigen on FDCs or an increasing differentiation of B-cells into plasma and memory B-cells as a consequence of differentiation of FDCs and Tfh cells. A crucial parameter in [77] is the probability that a positive selected B-cell recycles back to the DZ. Understanding the mechanism and regulation of recycling is also considered as a key to understand AAM in [64]. Here, by comparing model predictions with experimental data, the authors propose that the selection probability of B-cells and the recycling probability of selected B-cells are not constant, but rather vary during the GCR with respect to time.

Another process affecting B-cells during a GCR which remains unclear is the

selection mechanism. This was investigated for example in [98] through a non-linear, non-local and inhomogeneous parabolic PDE, describing a population of B-cells submitted to mutation, division and selection during a GCR. Conversely to *e.g.* [76, 70, 96], in [98] the space of traits is continuous (the interval $[0, 1]$), and is directly translated into an affinity function characterizing the likelihood that a given B-cell binds to the immunizing antigen. In this framework the termination is regulated by the number of selected B-cells, since the division rate is defined as a decreasing function of the selected pool size. A substantially different approach to investigate selection mechanisms in GCs is applied in [96]. There M. Meyer-Hermann and coworkers employ an extended version of a previously described agent-based model for GCR [92, 95]: they suggest that for physiologically reasonable parameter values only clonal competition for Tfh cells help or a refractory time for B-FDCs interactions can enable AAM while generating the experimentally observed GC characteristics. They consider a very detailed model which results really hard to study mathematically, as well as in *e.g.* [94] by M. Meyer-Hermann *et al.* Indeed, they take into account different cell type populations, interactions, cell motility and diffusion of molecular signals.

In most papers GCs are considered as isolated from each others. In [148] the authors present a coarse-grained model mathematically formalized through deterministic mean field differential equations, to calculate the B-cell population development in AAM. There they study the enhancement of affinity improvement due to B-cell migration between GCs. They investigate the reasons behind optimal parameters such as the optimal mutation rate or the optimal selection strength. Their findings suggest that GCs have been optimized by evolution to generate high-affinity antibodies efficiently and in a very short timeframe. In [148] two puzzles observed in the previous works of A. S. Perelson and coworkers [76, 105] are solved. For instance these previous models did not succeed in showing the extremely high improvement of affinity ($\sim$ 100 fold) and the "all-or-none" phenomenon observed in experiments. The latter refers to the fact that the fraction of strong affinity B-cells, usually characterized by a certain key mutation or a unique piece of Ig gene sequence, is more likely to be high or low, but less likely to be intermediate.

Most of papers presented so far consider a deterministic continuum approach, where cell concentrations are described by a set of coupled ODEs changing deterministically and continuously during time. This approach has many computational advantages and has often been employed to model biological systems. Nevertheless it is not able to take into account those local inhomogeneities

related to the discrete nature of cells and stochastic fluctuations in reaction processes. M.T. Figge in [49] introduce a microscopic reaction-diffusion model for GCR on a d-dimensional lattice, performing numerical simulations within a stochastic discrete event approach. In particular, in order to simulate the correct time evolution of this complex biological system, each single reaction event is monitored in space and time. Each reaction changes the lattice configuration into another configuration with a given probability, and the reactions occur in a stochastic manner. In [45] Y. Elhanati *et al* find biological evidence for an evolutionary model of B-cells where substitution rates across sites in the Ig primary chain strictly depend on the context. In order to do so they apply probabilistic inference methods and advanced statistical techniques to quantify the process that shape B-cell repertoire diversity. In [91] the authors developed and applied modern statistical methods to investigate selection on BCRs and infer B-cell sequence evolution. They use stochastic mapping and empirical Bayes estimates, comparing the evolution of BCRs rearrangements.

The work we develop in Chapters 2-4 is inserted in this extremely varied and stimulating context. Our aim is to define a simplified mathematical model of the learning process of B-cells in GC, focusing on the most basic mechanisms: mutation, division and selection. We introduce and successively couple these fundamental processes, and we perform a rigorous mathematical analysis using probabilistic tools ranging from simple random walks to multi-types Galton Watson processes. Our simplified mathematical framework allowed to introduce and study many different mutation-division-selection processes while already bringing interesting mathematical problems.

1.3 Main modeling assumptions and results

The aim of this thesis is to contribute to the mathematical foundations of adaptive immunity by building and analyzing a simplified mathematical model of the mutation-division-selection process of B-cells in GCs leading to AAM. Mutation, division and selection correspond for us to the fundamental building blocks of the AAM process: our approach consists in studying precisely each block and progressively enriching our model with supplementary bricks. We want to understand how the different biological parameters affect the system's functionality. We are particularly interested in estimating via probabilistic methods how different mutational rules affect typical time-scales to reach a specific configuration (or a set) of the traits of B-cells, as a function of the given mutational rule, as well as in quantifying GCs' efficiency. Beyond the fundamental understanding of physiological processes and their associated pathologies, this research is

also motivated by important biotechnological applications, such as the synthetic production of specific antibodies for drugs, vaccines or cancer immunotherapy [6, 79, 122]. Indeed this production process involves the selection of high affinity peptides and requires smart methods to generate an appropriate diversity [34]. Moreover, the study of this learning process has also given rise in recent years to bio-inspired algorithms such as in [30, 107], mainly addressed to solve optimization and learning problems [25].

Chapter 2 focuses on pure mutational models, aiming to model the SHM mechanism and understanding how different mutational rules can drive the exploration of the state-space of B-cell traits, hence affect AAM. Moreover, understanding the role and functional implication of mutations is a central question in biological evolutionary theory[50, 145, 57, 47], as well as for the study of evolutionary algorithms [9, 2]. The preliminary analysis we made of SHM suggests us to pattern these mutations as random walks on graphs, whose characteristics change depending on the introduced mutational rule. Hence we focus on the variation of hitting times as a function of the underlying graphs. This allows us to relate mutation rules to the characteristic time-scales of the process. In order to simplify the problem, here and in Chapter 3 we suppose we are allowed to classify the amino-acids which determine the chemical properties of both BCRs and antigen into two classes, named 0 and 1 respectively. They may corresponds to amino-acids positively charged and negatively charged. Henceforth BCRs and antigen are represented by binary strings of same fixed length N: the BCR state-space is $\{0,1\}^N$. This simplified choice is motivated by the difficulty of modeling *e.g.* the binding affinity between BCR and antigens, as well as the effect of genetic mutations affecting the Ig primary sequence over the geometrical structure of the resulting binding region of BCRs. We consider a linear contact between BCR and antigen, *i.e.* for the sake of simplicity, we state that 0 matches with 0 and 1 with 1, and define the affinity as the number of identical bits shared by the BCR representing string and the antigen representing string. In all following Chapters the antigen representing string is denoted by $\overline{\mathbf{x}}$. Definitions and notations are clarified in Section 2.2.

We follow the evolution of the trait, hence the binding affinity, of a single B-cell for a given antigen. We suppose it is submitted to mutations in the absence of other biasing mechanisms such as division and selection. The choice of a mutation rule corresponds to the prescription of a graph structure over $\{0,1\}^N$: a mutation step is modeled as a random jump to a neighbor node of the obtained graph. In Section 2.2 we define the basic mutational rule: at each time step a randomly chosen amino-acid composing the BCR switches the class it belongs

to. Mathematically this corresponds to a Simple Random Walk (SRW) on the N-dimensional hypercube, which is denoted by $\mathcal{H}_N$. We denote by $\mathcal{P}$ the transition probability matrix corresponding to this mutational rule. Of course the SRW over $\mathcal{H}_N$ has already been studied in different contexts. After recalling some already known results about RWs on graphs applied to this specific case, we consider the evolution of the Hamming distance to $\overline{\mathbf{x}}$ during this mutational process, seen as a RW on $\{0, \dots, N\}$. Due to the perfect symmetry of the hypercube and our particular choice of the affinity (which is directly related to the Hamming distance), by studying this new RW we reduce considerably the number of vertices of the graph, passing from 2^N to $N+1$ nodes, without losing the most important properties of the corresponding transition matrix, *e.g.* its eigenvalues. By studying this RW we obtain a new explicit formula to evaluate the hitting time to cover a given initial Hamming distance for the SRW on $\mathcal{H}_N$, which is proportional to the number of vertices. Moreover in Theorem 2.2.12 we improve this result by giving an explicit formula to compute the mean hitting time to reach a sphere of radius r centered in $\overline{\mathbf{x}}$.

It is possible to modify this basic mutational rule in many different ways to define more complex mutational mechanisms. We want to understand the effects of different mutational models on the connectivity of the graph and the efficiency of state-space exploration. In Section 2.3 we introduce and study several mutation rules on $\{0,1\}^N$, their effects on the structure of the graph and, consequently, the associated RWs. In particular, using both spectral and probabilistic methods, we compute the hitting times: starting from a random initial condition, we count the time expected to reach a target node. It has a clear biological interpretation, as it represents the expected number of mutations we need to build the BCR with fittest affinity, given a particular antigen and the initial lower-affinity BCR trait. This allows us to compare the ability of different mutational rules in exploring the state-space of all possible BCRs. We especially focus on two mutation rules that are the combination of simpler ones: the class switch of 1 or 2-length strings, where the mutation rule depends on the distance to the target, and the mutation rule which allows to do more than a single mutation at each step, defined as a convex combination of $\mathcal{P}^i$ for $0 \leq i \leq k$, and k fixed at most equal to N. Therefore here k represents the amplitude of the maximal change in the affinity strength in a single time step. We estimate that at least for N big enough, the hitting time corresponding to the model of class switch of 1 or 2-length strings is halved comparing to the basic mutational model, which is confirmed by numerical simulations (Proposition 2.3.11 and Table 2.3). We define two variants of the model of multiple point mutations, whose corresponding transition probability matrices are respectively

$\frac{1}{k}\sum_{i=1}^{k}\mathcal{P}^i =: \mathcal{P}^{(k)}$ and $\mathcal{P}^k$. Since in this case the Hamming distance does not correspond to the graph distance (except if $k = 1$) we average the hitting time over all couples of nodes having an initial Hamming distance $\overline{d}$. By applying a general formula given in [85] we succeed in determining an explicit formula to evaluate these mean hitting times, which is given in Proposition 2.29. We observe that for $k > 2$ the mutational model which assures the best hitting time is given by $\mathcal{P}^k$. Table 2.2 summarizes the main results of Section 2.2 and 2.3.

In Section 2.4 we present a biologically more involved model and discuss its numerical outputs within our mathematical framework, providing as well limitations and possible extensions of our approach. In particular we deeply describe the SHM process and how a single genetic mutation affect the composition of the corresponding amino-acid chain. We take into account the possibility of inserting or deleting an amino-acid from the string as a consequence of SHM. Indeed SHM introduces mostly single nucleotide exchanges, but also small deletions and duplications, *i.e.* insertions of extra copies of a portion of genetic material already present within the DNA code [63, 26, 27]. We observe numerically how it affects the hitting time (Table 2.6). We also discuss our choice of a binary representation and how our estimations can be compared to other models with a bigger amino-acid alphabet.

In Chapter 3 we introduce the division process in the same mathematical framework set in Chapter 2. We want to understand how interactions between division and different mutational models affect the diversification of the B-cell population repertoire as a consequence of clonal expansion and SHM and in the absence of any selection mechanism. Therefore we are particularly interested in analyzing characteristic time-scales for which a certain proportion of possible traits is expressed in the population: starting from a single individual, what would be the typical time until a finite proportion of the traits are covered by the exponentially increasing population? We consider $\{0,1\}^N$ as the state-space of all possible BCRs and the mutational rules already discussed in Sections 2.2 and 2.3. A division event is always associated to mutation, meaning that the newborn particles move to neighbor nodes according to a given mutation rule. Therefore we model the division-mutation process as Branching Random Walks with constant division rate 2 (2-BRW) (except for Section 3.5.2) over the graph defined on $\{0,1\}^N$ by the prescription of a given mutational rule. By applying the theory of expander graphs on the underlying graphs, we obtain estimates for the partial cover times of the considered BRWs.

In Section 3.4 we consider a simple 2-BRW (also called COBRA walk [43, 33])

where two or more particles having the same trait coalesce into a single one. The coupling of branching mechanisms and random walks necessarily implies an important speedup in the characteristic time-scales of state-space exploration, typically passing from a time $\mathcal{O}(2^N)$ to $\mathcal{O}(N)$ for the SRW on $\mathcal{H}_N$. Of course this has a cost: considering a branching process means also to produce new individuals at each time step. Indeed, in a time $T = \mathcal{O}(N)$ we have 2^T individuals (in the case in which multiplicity is taken into account; $\leq 2^N$ otherwise), as we do not consider here neither selection nor death. Therefore we decide to estimate which is the proportion of nodes we expect to activate in a time of the order of N and depending on the mutational rule. We want to compare the ability of different mutational models in increasing the diversity of expressed BCRs after $\mathcal{O}(N)$ rounds of clonal expansion and mutation. Therefore in Section 3.4 we compare the 2-BRW referring to the mutational models underlined by $\mathcal{P}$ and $\mathcal{P}^{(k)}$ respectively. The main results of this section are collected in Table 3.1: while the basic mutational model allows to cover a small portion of the state-space in $\mathcal{O}(N)$, in a time of the same order the model corresponding to $\mathcal{P}^{(k)}$ allows to explore almost a half of the state-space. In order to obtain these results (Theorems 3.4.9 and 3.4.13) we have characterized the expansion properties of the corresponding mutational graphs.

The mathematical analysis we made of the 2-BRW-$\mathcal{P}^{(k)}$ has revealed an interesting phenomenon concerning the impact of the mutation rate on the exploration speed. Intuitively, one would suggest that increasing the number of mutations at each division would result in a BRW with a faster exploration time-scale. However, in Section 3.4.3 we show the existence of an early saturation phenomenon: when increasing from one to two mutations, the exploration is indeed faster, but allowing more than two mutations (up to N) modifies only marginally the exploration speed. This discovery is also confirmed by numerical simulations, as shown in Figure 3.4.

In Section 3.5 we propose some extensions of the model. In particular, we introduce the BRW with multiplicity and obtain the transition matrix related to the number of individuals carrying a given trait together with their limiting distribution (Lemma 3.5.3). This adds a further building block to our model. Indeed, taking into account the number of particles lying on the same vertex allows to consider the size of the effective population and not only how many different BCR configurations are expressed at a certain time. In a further step we investigate how this distribution can change by introducing a division rate, and provide comparisons between different mutation-division models. In this way, theoretical results presented in previous sections are displayed in a wider con-

text. In particular Lemma 3.5.5 shows that the addition of a division rate allows to overcome the problem of the eventual bipartite structure of the considered graph. Moreover in Section 3.5.3 we propose a model in which the division rate is dependent on the affinity. This is consistent with the experimentally observed fact that Tfh-B-cells interactions determine the number of cycles of proliferation of positive selected B-cells which recycle back to the DZ. This seems to be proportional to their affinity strength (Section 1.1). We observe through numerical simulations that this actually allows the fittest clones to have an advantage over the low-affinity population.

In Chapter 4 we introduce and study more complex models involving mutation, division and affinity-dependent selection mechanisms. In this context we refer to some more general modeling assumptions. For instance we do not need to define a specific state-space for B-cell traits, but rather we suppose that all BCRs can be classified into a certain number of affinity classes with respect to the presented antigen. They are enumerated from 0, the higher affinity class, to N, the lower one, and we assume that all B-cells belonging to the same affinity class have similar binding abilities. Affinity classes may contain all B-cells having the same Hamming distance from the target, if we suppose that B-cell traits are represented as N-length binary strings and their affinity is described using the Hamming distance, as in Chapters 2 and 3. SHM implies that a mutated clone eventually passes from the affinity class of its mother cell to another one: this is modeled through a transition probability matrix over $\{0, \dots, N\}$. Under modeling assumptions of Chapters 2 and 3, the transition probability matrix over $\{0, \dots, N\}$ describes the evolution of the Hamming distance to $\overline{\mathbf{x}}$ as a consequence of SHM.

In Section 4.2 we define the main model analyzed in Chapter 4. The process starts with z_0 naive B-cells entering the GC at time 0, eventually belonging to different affinity classes. At each time step each GC B-cell can die with rate r_d. If not it can divide with rate r_{div}, giving rise to two newborn cells with a mutated trait, according to the allowed mutational rule. Then, each cell in the population can be submitted to selection with rate r_s: a threshold is fixed for positive selection. If the B-cell submitted to selection has a worst fitness than the threshold, it dies by apoptosis, otherwise it exits the GC and enters the selected pool. Hence in this case no recycling mechanisms are taken into account.

We mathematically formalize this model in Section 4.3 by opportunely using a $(N+3)$-type Galton Watson (GW) process. This model predicts the evolu-

tion of GC population and provides useful information concerning the extinction probability of the GC, the average affinity of clones, the expected size of the GC and the expected number of selected cells. These qualitative informations are rigorously addressed in this section (Proposition 4.3.9).

What is the behavior of the expected number of selected B-cells as a function of the model parameters? In particular, is there an optimal value of the selection rate which maximizes this number? Thanks to the spectral decomposition of the matrix describing the average behavior of the introduced multi-type GW process, we determine explicitly the optimal value of the selection pressure which maximizes the expected number of selected B-cells at a given time step. This corresponds to $1/t$ independently from all other parameters and from the mutational model (Corollary 4.3.11).

The model we set can be easily modified to define *e.g.* other affinity-dependent mechanisms, which could be studied at least numerically. Indeed in Section 4.4 we propose two variants of the previous model: a positive selection model and a negative selection one. In the first case the selection mechanism acts only positively, meaning that if a B-cell submitted to selection has a trait good enough to be positive selected, then it exits the GC and reach the selected pool like in the main model. On the contrary, when its affinity is not high enough, nothing happens: it remains in the GC for the next time step. The model of negative selection acts in the opposite way: a positive selected B-cell stays in the GC for further rounds of mutation-division-selection, while a negative selected B-cell dies. This last model corresponds to the case of 100% of recycling.

Because of the peculiar structure of matrices containing the average evolution of each type cell for both models of positive selection and of negative selection, we are not able to compute explicitly their spectra. Henceforth we can not give an explicit formula for *e.g.* the extinction probability of the corresponding GCs or evaluate the optimal values of the selection rate to maximize the production of output cells as we did for the model analyzed in Section 4.3. Nevertheless we can give some estimations (Proposition 4.4.3) by using standard arguments for positive matrices such as the Perron Frobenius Theorem. Moreover we can easily perform numerical simulations illustrating our theoretical results: we give and comment some of the obtained graphics in Section 4.4.2. In particular in Figure 4.7 we compute the optimal choice of the selection rate which maximizes the expected number of selected B-cells at a given time step, for the model of positive selection. We show that from one hand the searched optimal value

depends on the relation between the initial affinity of naive B-cell clones and the fixed affinity threshold. On the other hand it seems that for t big enough the optimal r_s tends to $1/t$ as in the main model and independently from the other parameters. One has to interpret this result as the ideal optimal strength of the selection pressure to obtain a peak of the GC production of output cells at a given time step. For example, let us suppose that a time step corresponds to 1 day. The peak of the GC reaction has been measured to be close to day 12 [144]: for the kind of model we built and analyzed in this paper, a constant selection pressure of $1/12$ assures that the production of plasma and memory B-cells is maximized at the GC peak.

In Section 4.5 we discuss the modeling assumptions considered in Chapter 4 and provide possible extensions of the presented models. Indeed the mathematical tools used in Sections 4.3 and 4.4 can be applied to define and study models with different affinity-dependent selection mechanisms, as well as models in which one or more parameters vary during time, or with alternate periods of mutation-free growth. We plot in Figure 4.9 an example of the profiles we can expect letting the selection pressure increase over time. This shall take into account, for instance, the early GC phase in which simple clonal expansion of B-cells with no selection occurs [36].

Chapters 2-4 define a simple but powerful mathematical framework in which many different evolutionary processes can be formalized and studied. We demonstrate how it is possible to enrich the models by progressively adding new bricks and hypotheses. We provide as well suitable mathematical tools to study the introduced models and perform many numerical simulations which confirm our theoretical results. Of course the framework remains highly theoretical and can be improved in many different ways. In Chapter 5 we discuss some limitations and propose some possible improvements of the models described so far.

Chapter 2

Random walks on binary strings applied to the somatic hypermutation of B-cells

Summary Within the germinal center in follicles, B-cells proliferate, mutate and differentiate, while being submitted to a powerful selection: a micro-evolutionary mechanism at the heart of adaptive immunity. A new foreign pathogen is confronted to our immune system, the mutation mechanism that allows B-cells to adapt to it is called somatic hypermutation: a programmed process of mutation affecting B-cell receptors at extremely high rate. By considering random walks on graphs, we introduce and analyze a simplified mathematical model in order to understand this extremely efficient learning process. The structure of the graph reflects the choice of the mutation rule. We focus on the impact of this choice on typical time-scales of the graphs' exploration. We derive explicit formulas to evaluate the expected hitting time to cover a given Hamming distance on the graphs under consideration. This characterizes the efficiency of these processes in driving antibody affinity maturation. In a further step we present a biologically more involved model and discuss its numerical outputs within our mathematical framework. We provide as well limitations and possible extensions of our approach.

2.1 Introduction

Understanding the role and functional implication of mutations is a central question in biological evolutionary theory [50, 145, 57, 47], but also for the study of evolutionary algorithms [9, 2]. Beyond the mutation rate, which is naturally an important parameter, our aim in this Chapter is to highlight the role of various mutation rules on the exploration of the space of traits. In our mathematical framework, configurations are represented as vertices of a graph which are connected if there exists a mutation allowing to pass from one trait to another. We are mainly interested in understanding the characteristic time-scales for the exploration of the state-space as a function of the mutation rule. To this end, we relate mutation rules with specific graph topologies and build upon random walks on graphs and spectral graph theories to analyze resulting time-scales.

More precisely, beyond general theoretical results, we are particularly interested to apply our framework to the B-cell affinity maturation in Germinal Centers (GCs). The adaptive immune system is able to create a specific response against almost any kind of pathogens penetrating our organism and inflicting diseases. This task is performed by the production of high affinity antigen-specific antibodies. These proteins are produced by B-lymphocytes which are submitted to a learning process improving their affinity to recognize a particular antigen. This process is called Antibody Affinity Maturation (AAM) and takes place in GCs [102]. Even if substantial progress has been made in adaptive immunology, since somatic hypermutation was discovered by the nobel price Susumu Tonegawa [135] in 1987, there are still facts that remain unclear about the GC reaction and the exact dynamics of AAM. Indeed, it seems difficult to make exact measurements of the antigenic repertoire *in vivo* inside a single GC, following and sequencing each B-cell at any time, or to have precise spatial and temporal data about lymphocytes within the GC during an immune response, or to understand the exact dynamic of mutation and selection of B-cells while they are submitted to AAM (*e.g.* [48, 106]). Nevertheless, some refined techniques start to be available [131, 55], showing possible correlations between proliferation and mutation rates with respect to B-cells' affinity to the presented antigen. This provides further motivation for setting appropriate mathematical frameworks to describe such systems.

The affinity of a B-cell is biologically observed as a matching between the B-cell receptor (BCR) and the antigen. We aim at understanding how mutation rules allow to explore possible trait-configurations of BCRs. The mutational mechanism that B-cells undergo in GCs to improve their affinity is called So-

matic Hypermutation (SHM): it targets, at a very high rate, the DNA encoding for the specific portion of the BCR involved in the binding with the antigen, called Variable (V) region. SHM can introduce mutations at all four nucleotides, and mutation hot-spots have been identified [133, 45, 127]. The effect of these mutations on the BCR, once expressed on the outer surface of B-cells, is very complex, as the substitution of a single amino-acid can modify the geometrical structure of the BCR, creating or deleting bonds (see [1], Chapter 4, for more details about the crystal structure of BCRs and their binding with antigens).

Although mutations occur at the level of the DNA, their outcome might be expressed at the level of amino-acids composing the BCR. In the present Chapter, SHMs are taken in account this way (Section 2.4.3). However, the structure of our mathematical model can be left substantially unchanged when considering mutations at the DNA level, which leads to modify the definition of affinity and the size of the state-space.

There already exists a certain number of mathematical models about GC reaction and AAM. In particular, [75, 76] proposed deterministic population modeling of SHM and AAM, considering for instance the hypothesis of recycling mechanisms during GC reaction, later investigated by experiments [139]. In [105, 108, 52, 64], the authors introduced and discussed several immunological problems, such as the size of the repertoire, or the strength of antigen-antibody binding, or the pourcentage of recycling. They provide suitable mathematical tools, using both deterministic and probabilistic approaches, together with numerical simulations. More recently, biologically very detailed models of GCs were proposed [92, 120], using, for instance, agent-based models [94], mostly analyzed through extensive numerical simulations. Our aim here is not to build a very complex model, but rather to contribute to the theoretical foundation of adaptive immunity modeling through the mathematical analysis of generic mutation models on graphs. So far, this approach has not been developed and applied to GC reaction and AAM modeling. In particular, this framework enables the study of various mutation rules, as for instance, affinity-dependent mutations, which are currently debated in the biological literature [55]. Our mathematical framework shares some similarities with the NK models proposed by S. A. Kauffman and E. D. Weinberger in [70], for instance the choice of the hypercube vertex set as the basic structure to define the affinity landscape of BCRs. Nevertheless their approach and goals are fundamentally different from ours. Indeed, in [70] the graph which defines the mutational rule is predefined (*i.e.* they refer only to the basic mutational rule we introduced as well in Section 2.2), while the affinity function changes according to the main parameters of the

model, N and k for instance. Therefore, the random walks over these affinity landscapes, modeling the maturation of the immune response, are biased with respect to the affinity gradient. In our mathematical framework the structure of the graph reflects the mutational rule, hence it is not predefined. Moreover, since in this Chapter we only take into account mutations, the random walks over the state-space are not biased by the fitness of each trait to the target one. From our point of view the selection pressure should be taken into account as a separate operator (see below).

This research is also motivated by important biotechnological applications. The fundamental understanding of the evolutionary mechanisms involved in antibody affinity maturation have been inspiring many methods for the synthetic production of specific antibodies for drugs, vaccines or cancer immunotherapy [6, 79, 122]. Indeed, this production process involves the selection of high affinity peptides and requires smart methods to generate an appropriate diversity [34]. Beyond the biomedical motivations, the study of this learning process has also given rise in recent years to a new class of bio-inspired algorithms such as in [30, 107], mainly addressed to solve optimization and learning problems [25].

In this Chapter, we consider pure mutational models obtained as random walks on graphs given by alterations of the edge set of the N-dimensional hypercube. We focus on the variation of hitting times as a function of the underlying graphs, hence relating mutation rules to the characteristic time-scales of the process. Our intention here is not to provide biologically relevant outcomes, since the AAM involves several mechanisms (division, selection, etc) that we do not take into account in this Chapter. Instead we provide a rigorous analysis of an essential single building block: mutation. We study the structure of RWs on the hypercube and compute hitting times depending on the graph associated to the mutational rule. We prove that they are proportional to the number of vertices (see Table 2.2). Therefore our specific approach consists in observing how different mutational rules allow to explore the state-space and lead a naive B-cell to build the fittest possible trait. We are not interested here in proposing new statistical or phylogenetic strategies to infer the more realistic phylogenetic trees given a final antibodies repertoire [54, 32]. Nevertheless we define accurately the biological context since it is relevant for further steps. Clearly, other mechanisms such division and mutations provide significant biases of hitting times, our approach consists in studying precisely the differences when enriching our model with supplementary bricks. For instance, by branching we introduce a population dispatched on the vertices of the hypercube which decreases the hitting time, but at the cost of the biological maintaining of the

population (Chapter 3). This is our strategy here and in the forthcoming Chapters.

Section 2.2 contains results on random walks theory [104, 97, 111] and, more specifically, random walks on graphs [85, 4]. This is a topic of active research due to the great number of important applications in recent years, such as graph clustering [117], ranking algorithms for search-engines [19, 68], or social network modeling [72, 56, 78]. We start with the most basic mutational model which is the simple random walk on the N-dimensional hypercube [41, 58, 39, 140]. We set notations in order to define the models, then we overview various properties of random walks on graphs, and establish particular results in the case of the hypercube. In Section 2.3 we study several mutation rules and their effects on the structure of the graph and, consequently, its associated random walk. In particular we compute the hitting times: starting from a random initial condition, we count the expected time to reach a target node with the best fitness. We use both spectral and probabilistic methods. We especially focus on two mutation rules that are the combination of simpler ones: the class switch of 1 or 2-length strings, where the mutation rule depends on the distance to the target, and the mutation rule which allows to do more than a single mutation at each step. Table 2.2 in Section 2.3.2 summarizes the main results of Section 2.2 and 2.3: we display expected times to reach some position of the graph, as a function of each mutation rule. Finally, Section 2.4 is dedicated to modeling aspects and discussions about possible extensions and limitations of the proposed framework.

2.2 A basic mutational model

In this section we set the general mathematical framework, which we keep in order to pattern and study mutational mechanisms discussed in the current section and in Section 2.3. Indeed, we state a basic mutational model. The choice of this environment is motivated by the modeling of amino-acids chains and their modifications during SHM. It is for this reason that we often recall biological facts and refer to BCRs and antigens. Nevertheless, this framework is flexible and adapts to different mutational rules in a more general evolutionary context.

We assume that it is possible to classify the amino-acids into 2 classes denoted by 0 and 1 respectively (they could represent amino-acids negatively and positively charged respectively). Henceforth BCRs and antigen are represented by binary strings of same fixed length N, hence, the state-space of all possi-

ble BCR configurations is $\{0,1\}^N$. We will give some more details about these hypotheses in Section 2.4.3.

Definition 2.1. We denote by $\mathcal{H}_N$ the standard N-dimensional hypercube. BCR and antigen configurations are represented by vertices of $\mathcal{H}_N$, denoted by $\mathbf{x}_i$ with $1 \leq i \leq 2^N$, or sometimes simply by their indices. We denote the antigen target vertex by $\overline{\mathbf{x}}$: it is given at the beginning of the process and never changes.

We suppose that there is a single B-cell entering the GC reaction. The configuration of its receptors is denoted by $\mathbf{X}_0$. If $\mathbf{X}_t$ is the configuration of the BCR after t mutations, then depending on the mutational rule, one or more bits in $\mathbf{X}_t$ can change after the next mutation. This gives rise to a Random Walk (RW) on $\{0,1\}^N$, where a mutation on the BCR corresponds to a jump to a neighbor node. Of course, the definition of neighbors changes depending on the mutation rules we introduce (we specify the neighborhood set each time we discuss a new mutation rule). In a general way:

Definition 2.2. Given $\mathbf{x}_i$, $\mathbf{x}_j \in \{0,1\}^N$, we say that $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors, and denote $\mathbf{x}_i \sim \mathbf{x}_j$, if there exists at least one edge (or loop) between them.

As far as the complementarity is concerned, we have to make a further simplification. As we have already discussed in the Introduction, the tridimensional structure of the BCR is hard to model. For this reason we consider a linear contact, *i.e.* positively charged amino-acids are complementary to negatively charged ones when they are at the same position within the binary string. For the sake of simplicity, we state that 0 matches with 0 and 1 with 1 (we can suppose that the antigen representing string is given in its complementary form). Formally, we define the affinity as the number of identical bits shared by the BCR representing string and $\overline{\mathbf{x}}$. Equivalently, one can see $\overline{\mathbf{x}}$ as the optimal BCR trait, with the highest affinity for the immunizing antigen.

Definition 2.3. For all $\mathbf{x}_i \in \{0,1\}^N$, its affinity with $\overline{\mathbf{x}}$, $\mathrm{aff}(\mathbf{x}_i, \overline{\mathbf{x}})$ is given by $\mathrm{aff}(\mathbf{x}_i, \overline{\mathbf{x}}) := N - h(\mathbf{x}_i, \overline{\mathbf{x}})$, where $h(\cdot,\cdot) : (\{0,1\}^N \times \{0,1\}^N) \to \{0,\dots,N\}$ returns the Hamming distance.

Definition 2.4. For all $\mathbf{x} = (x_1,\dots,x_N)$, $\mathbf{y} = (y_1,\dots,y_N) \in \{0,1\}^N$, their Hamming distance is given by:

$$h(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N} \delta_i \qquad \text{where} \qquad \delta_i = \begin{cases} 1 & \text{if} \quad x_i \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

Other definitions of affinity are often (*e.g.* [92]) constructed as functions of the Hamming distance $\mathrm{aff}(\mathbf{x}_i, \overline{\mathbf{x}}) = F(h(\mathbf{x}_i, \overline{\mathbf{x}}))$, for instance with F given by

the Gaussian probability density function. These modeling aspects become important when considering the selection mechanism, which is not treated in the present article. Therefore, for our purpose, we can focus on the above definition of affinity.

As a first basic mutational rule, we study single switch-type mutations: at each time step a randomly chosen amino-acid within the BCR binary string switches its amino-acid class. This clearly leads us to a Simple Random Walk (SRW) on $\mathcal{H}_N$. Indeed, we formalize it as follows:

Definition 2.5. Let $\mathbf{X}_n \in \mathcal{H}_N$ be the BCR at step n. Let $i \in \{1, \dots, N\}$ be a randomly chosen index. Then $\mathbf{X}_{n+1} := (X_{n,1}, \dots, X_{n,i-1}, 1 - X_{n,i}, X_{n,i+1}, \dots, X_{n,N})$.

Remark 1. Referring to Definition 2.2 of neighborhood, as we consider here the standard N-dimensional hypercube, $\forall\, \mathbf{x}_i,\, \mathbf{x}_j \in \mathcal{H}_N$, $\mathbf{x}_i \sim \mathbf{x}_j \Leftrightarrow h(\mathbf{x}_i, \mathbf{x}_j) = 1$.

We denote the transition probability matrix of the SRW on $\mathcal{H}_N$ by $\mathcal{P}_N$ or simply by $\mathcal{P}$ if no misunderstanding is possible. For all $\mathbf{x}_i,\, \mathbf{x}_j \in \mathcal{H}_N$:

$$\mathbb{P}(\mathbf{X}_n = \mathbf{x}_j \,|\, \mathbf{X}_{n-1} = \mathbf{x}_i) =: p(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1/N & \text{if} \quad \mathbf{x}_j \sim \mathbf{x}_i, \\ 0 & \text{otherwise.} \end{cases}$$

The entries of $\mathcal{P}$ are $(p(\mathbf{x}_i, \mathbf{x}_j))_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{H}_N}$. The unique stationary distribution for $\mathcal{P}$ is the homogeneous probability distribution on $\mathcal{H}_N$, denoted by $\boldsymbol{\pi}$: $\forall\, \mathbf{x}_i \in \mathcal{H}_N$, $\pi_i := \boldsymbol{\pi}(\mathbf{x}_i) = 2^{-N}$. Indeed, $(\mathbf{X}_n)_{n \geq 0}$ is clearly reversible with respect to $\boldsymbol{\pi}$. The uniqueness follows by the Ergodic Theorem.

We also recall a property of $\mathcal{H}_N$ that we will have to deal with: the bipartiteness.

Definition 2.6. A graph $G = (V, E)$ is bipartite if there exists a partition of the vertex set $V = V_1 \sqcup V_2$, s.t. every edge connects a vertex in V_1 to a vertex in V_2.

Typically a bipartition of the hypercube can be obtained by separating the vertices with an odd number of 1's in their string from those with an even number of 1's. In Figure 2.1 we emphasize the bipartite structure of the hypercube $\mathcal{H}_3$.

A direct and elementary consequence of this property is the periodic behavior of the SRW on $\mathcal{H}_N$, which in particular causes some problems for the convergence through $\boldsymbol{\pi}$. This problem is classically overcome by adding N loops

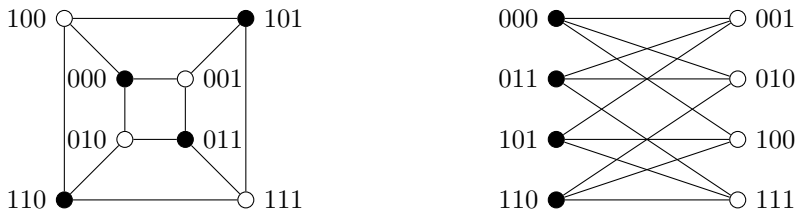


Figure 2.1: Hypercube for $N = 3$ showing its bipartite structure.

at each vertex, that makes this RW become a *lazy Markov chain* [83]. The corresponding transition probability matrix is given by $\mathcal{P}_L := (\mathcal{P} + \mathcal{I}_{2^N})/2$, where $\mathcal{I}_n$ denotes the n-dimensional identity matrix.

2.2.1 Spectral analysis

Most matrices describing the characteristics of the SRW on $\mathcal{H}_N$ can be obtained recursively, thanks to the recursive construction of the hypercube and the operation of cartesian product between two graphs.

Definition 2.7. Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, the cartesian product between G_1 and G_2, $G_1 \times G_2$, is a graph with vertex set $V = V_1 \times V_2 = \{(u, v) \,|\, u \in V_1,\, v \in V_2\}$. Two different vertices (u_1, v_1) and (u_2, v_2) are adjacent in $G_1 \times G_2$ if either $u_1 = u_2$ and $v_1 v_2 \in E_2$ or $v_1 = v_2$ and $u_1 u_2 \in E_1$.

It is a known result [58] that for $N > 1$, $\mathcal{H}_N$ is obtained from $\mathcal{H}_{N-1}$ as: $\mathcal{H}_N = \mathcal{H}_{N-1} \times \mathcal{H}_1$. This characteristic implies the recursive construction of the adjacency matrix and allows to determine the corresponding eigenvalues and eigenvectors. We denote by A_N the adjacency matrix corresponding to $\mathcal{H}_N$; by $\mathcal{I}_n$ the n-dimensional identity matrix. Then we have:

$$A_1 = \begin{matrix} 0 \\ 1 \end{matrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \quad A_2 = \begin{matrix} 00 \\ 01 \\ 10 \\ 11 \end{matrix} \left(\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right) = \left(\begin{array}{c|c} A_1 & I_2 \\ \hline I_2 & A_1 \end{array}\right)$$

Here we wrote in gray the strings corresponding to each row: in order to obtain the adjacency matrices in this form, we simply have to order vertices of $\mathcal{H}_N$ in lexicographical order.

By iteration we obtain [51]:

$$A_n = \left(\begin{array}{c|c} A_{n-1} & \mathcal{I}_{2^{n-1}} \\ \hline \mathcal{I}_{2^{n-1}} & A_{n-1} \end{array}\right)$$

This iterative construction allows also to determine recursively the spectra of A_N and, consequently, of $\mathcal{P}_N = A_N/N$ (as $\mathcal{H}_N$ is a N-regular graph, the transition probability matrix corresponds to the adjacency matrix divided by N). Here below we recall the explicit values of the eigenvalues of A_N and $\mathcal{P}_N$ respectively. An extensive proof can be found in [51].

Theorem 2.2.1. *The eigenvalues of A_N are: $N, N-2, N-4, \ldots, -N+4, -N+2, -N$. If we order the $N+1$ distinct eigenvalues of A_N as $\lambda_1^A > \lambda_2^A > \cdots > \lambda_{N+1}^A$, then the multiplicity of λ_k^A is $\binom{N}{k-1}$, $1 \leq k \leq N+1$*

Corollary 2.2.2. *The eigenvalues of $\mathcal{P}_N$ are: $1, 1-2/N, 1-4/N, \ldots, -1+4/N, -1+2/N, -1$. If we order the $N+1$ distinct eigenvalues of $\mathcal{P}$ as $\lambda_1 > \lambda_2 > \cdots > \lambda_{N+1}$, then the multiplicity of λ_k is $\binom{N}{k-1}$, $1 \leq k \leq N+1$*

Finally we recall the expression of the eigenvectors of A_N (and then also of $\mathcal{P}$), that we gather together into a matrix. The eigenvectors for A_1 are:

$$\mathbf{z}_1 = \begin{bmatrix} 1 \\ \\ 1 \end{bmatrix} \text{ for } \lambda_1^A = 1 \quad \text{and} \quad \mathbf{z}_2 = \begin{bmatrix} 1 \\ \\ -1 \end{bmatrix} \text{ for } \lambda_2^A = -1 \Rightarrow \mathcal{Z}_1 = [\mathbf{z}_1, \mathbf{z}_2]$$

Thanks to the relations between the cartesian product of two graphs and their eigenvectors, it follows by induction that [51]:

$$\mathcal{Z}_n = \left(\begin{array}{c|c} \mathcal{Z}_{n-1} & \mathcal{Z}_{n-1} \\ \hline \mathcal{Z}_{n-1} & -\mathcal{Z}_{n-1} \end{array}\right)$$

Finally, one renormalizes each vector $\mathbf{z}_i$ multiplying it by $\sqrt{2^{-N}}$. We denote by Q_N the resulting matrix, where each column is a 2^N vector $\mathbf{v}_i = \sqrt{2^{-N}}\mathbf{z}_i$.

2.2.2 Evolution of Hamming distances to a fixed node

In this section we focus on the distance process, which is the process obtained from the SRW on $\mathcal{H}_N$ by looking at the Hamming distance between the B-cell representing string at each mutation step and the antigen target representing string. More precisely, $(D_n)_{n\geq 0} := (h(\mathbf{X}_n, \overline{\mathbf{x}}))_{n\geq 0}$ is a RW on $\{0, \ldots, N\}$. From

a biological point of view this process represents the evolution of the affinity of the mutating B-cell to the presented antigen. The idea of analyzing the distance of a RW on a graph to some position, where distance means the minimal number of steps that separate two positions, is not unusual. N. Berestycki in [18] applied that to genome rearrangements, where the distance on the graph corresponds biologically to the minimal number of reversals or other mutations needed to transform one genome into the other. Due to the perfect symmetry of the graph under consideration and our particular choice of the affinity (which is directly related to the Hamming distance), by studying (D_n) we reduce considerably the number of vertices, passing from 2^N to $N+1$ nodes, without losing the most important properties of the corresponding transition matrix. However, if we consider more complicated models of mutation, it is not possible to reduce the study of the process to the distances to a fixed node. In Figure 2.2 we show explicitly how to pass from $(\mathbf{X}_n)$ to (D_n): since $\overline{\mathbf{x}}$ is fixed and known, we are able to group the vertices by their Hamming distance to $\overline{\mathbf{x}}$. Moreover we keep the original probability of going to the next distance class by considering weighted and directed edges.

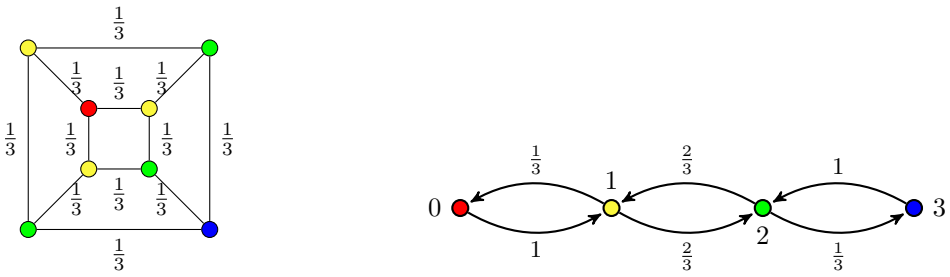


Figure 2.2: From the $(\mathbf{X}_n)$ process (on the left) to the (D_n) process (on the right) (case $N = 3$). Near each arrow the probability to travel in the corresponding direction is exhibited. The red vertex always corresponds to $\overline{\mathbf{x}}$, while we represent vertices at the same distance with the same color (yellow for $h = 1$, green for $h = 2$, and blue for $h = 3$).

The transition probability matrix for (D_n), denoted by $\mathcal{Q}$, is given by Proposition 2.2.3 below.

Proposition 2.2.3. *For all d, $d' \in \{0, \dots, N\}$:*

$$\mathbb{P}(D_n = d' \,|\, D_{n-1} = d) =: q(d,d') = \begin{cases} d/N & \text{if} \quad d' = d-1 \\ (N-d)/N & \text{if} \quad d' = d+1 \\ 0 & \text{if} \quad |d'-d| \neq 1 \end{cases} \tag{2.1}$$

$\mathcal{Q} = (q(d,d'))_{d,\,d'\in\{0,\dots,N\}}$ is a $(N+1)\times(N+1)$ tridiagonal matrix where the main diagonal consists of zeros. The stationary distribution for $\mathcal{Q}$ is the binomial probability distribution $\mathcal{B}\left(N,\frac{1}{2}\right) = \left(C_N^d \frac{1}{2^N}\right)_{d\in\{0,\dots,N\}}$, where $C_N^d = \binom{N}{d} = \frac{N!}{d!(N-d)!}$ is the binomial coefficient. It is the unique stationary distribution for $\mathcal{Q}$: a simple calculation points out the fact that $(D_n)_{n\geq 0}$ is reversible with respect to $\mathcal{B}\left(N,\frac{1}{2}\right)$, then the uniqueness follows by the Ergodic Theorem.

Anew, we have to deal with bipartiteness: the graph we are taking into account in this section is clearly bipartite, since we can separate its vertices into two subsets containing odd and even nodes respectively and no edge connects any vertices in the same subset. In order to overcome this problem we add N loops at each vertex $\mathbf{x}_i \in \mathcal{H}_N$ which means that the new transition probability matrix for the (D_n) process is, for all $d,\, d' \in \{0,\dots,N\}$:

$$\mathbb{P}(D_n = d' \,|\, D_{n-1} = d) =: q_L(d,d') = \begin{cases} 1/2 & \text{if} \quad d' = d \\ d/(2N) & \text{if} \quad d' = d-1 \\ (N-d)/(2N) & \text{if} \quad d' = d+1 \\ 0 & \text{if} \quad |d'-d| \neq 1 \end{cases} \tag{2.2}$$

We denote by $\mathcal{Q}_L := (q_L(d,d'))_{d,\,d'\in\{0,\dots,N\}}$.

Proposition 2.2.4. *$(D_n)_{n\geq 0}$ converges in law to a binomial random variable with parameters N and $1/2$. Explicitly:*

$$(\mathcal{Q}_L)_d \to \mathcal{B}\left(N,\frac{1}{2}\right)_d \quad \textit{for} \quad n\to+\infty$$

Proof. The proof follows directly observing that $\mathcal{Q}_L$ represents an irreducible and, now, aperiodic MC, with the same stationary distribution as $\mathcal{Q}$ (see [104] for a proof of the general result). □

The spectral analysis of $\mathcal{Q}$ gives the following result.

Theorem 2.2.5. *For fixed N, the spectra of the transition probability matrix $\mathcal{Q}$ corresponding to the (D_n) process is composed by the same $N+1$ distinct eigenvalues as the spectra of $\mathcal{P}$, each with multiplicity 1.*

Proof. The proof consists of a simple calculation of the eigenvalues of matrix $\mathcal{Q}$, which is easily done for $N = 1, 2$. Then we reason by iteration. We can also give the system we use for determining the eigenvectors. For fixed N let us

denote by $\lambda_{\pm k}$ the eigenvalue $\frac{\pm(N-2k)}{N}$ for $0 \leq k \leq \lfloor N/2 \rfloor$. We denote by $\mathbf{x}_{\pm k}$ the corresponding unknown eigenvector. Then we have the following matrix equation:

$$\mathcal{Q}\,\mathbf{x}_{\pm k} = \lambda_{\pm k}\mathbf{x}_{\pm k}$$

Which is:

$$\begin{cases} x_{\pm k,2} = \lambda_{\pm k} x_{\pm k,1} \\ \frac{1}{N} x_{\pm k,1} + \frac{N-1}{N} x_{\pm k,3} = \lambda_{\pm k} x_{\pm k,2} \\ \frac{2}{N} x_{\pm k,2} + \frac{N-2}{N} x_{\pm k,4} = \lambda_{\pm k} x_{\pm k,3} \\ \vdots \\ \frac{N-1}{N} x_{\pm k,N-1} + \frac{1}{N} x_{\pm k,N+1} = \lambda_{\pm k} x_{\pm k,N} \\ x_{\pm k,N} = \lambda_{\pm k} x_{\pm k,N+1} \end{cases}$$

□

Remark 2. Using the classical results of S. N. Ethier and T. G. Kurtz [46] it is possible to prove that, denoting by $x_N(t)$ the process $x_N(t) = \frac{D_{\lfloor Nt \rfloor}}{N}$, it converges in probability through $x(t)$, solution of the differential equation $\dot{x}(t) = -2x(t) + 1$ on a finite time window:

$$\forall\, \varepsilon > 0,\ \forall\, T > 0,\ \mathbb{P}\left(\sup_{t \in [0,T]} |x_N(t) - x(t)| > \varepsilon \right) \to 0 \quad \text{for } N \to \infty.$$

Remark 3. We can easily observe that $x(t)$ rapidly converges to $1/2$ for all $x_0 \in [0, 1]$. In particular if we start at $x_0 = 1/2$, we stay there for all t. That suggests that the (D_n) process, for N going to infinity, reaches a value of about $N/2$ exponentially fast, and then tends to remain there.

From an heuristic viewpoint we can explain how we derived the above equation. First of all, we take into account the following rescaled process:

$$x_n := D_n/N$$

As $(D_n) \in \{0, \dots, N\}$, $x_n \in [0,1]$. Denoting by $q_n(x) = \mathbb{P}(x_n = x)$ and using Equation (2.1), we have:

$$q_{n+1}(x) = (1-x)q_n\left(x - \frac{1}{N}\right) + xq_n\left(x + \frac{1}{N}\right)$$

Now we apply the Taylor theorem for $N \gg 1$:

$$q_{n+1}(x) = (1-x)\left(q_n(x) - \frac{1}{N}q_n'(x) + o\left(\frac{1}{N}\right)\right) + x\left(q_n(x) + \frac{1}{N}q_n'(x) + o\left(\frac{1}{N}\right)\right)$$

From which we get:

$$q_{n+1}(x) - q_n(x) = \frac{1}{N}(x - (1-x))q_n'(x) + o\left(\frac{1}{N}\right)$$

Defining the process $\tilde{q}(t,x) = q_{\lfloor Nt \rfloor}(x)$, with $t = \frac{n}{N}$, we obtain:

$$\partial_t \tilde{q}(t,x) = (2x-1)\partial_x \tilde{q}(t,x) + o\left(\frac{1}{N}\right)$$

And consequently, the corresponding transport equation is:

$$\partial_t q(t,x) = (2x-1)\partial_x q(t,x) \tag{2.3}$$

The differential equation associated with Equation (2.3) (its characteristic equation) is:

$$\dot{x}(t) = -2x(t) + 1$$

which has solution:

$$x(t) = \frac{1}{2} + \left(x_0 - \frac{1}{2}\right)e^{-2t}$$

It is also possible to derive a diffusion approximation by expanding the generator at second order.

2.2.3 Hitting times

In this section we give explicit formulas to compute the hitting time from node $\mathbf{x}_i$ to $\mathbf{x}_j$: the expected number of steps before $\mathbf{x}_j$ is visited, starting from $\mathbf{x}_i$. More precisely, we define by $\tau_{\{\mathbf{x}_j\}} := \inf\{n \geq 0 \,|\, \mathbf{X}_n = \mathbf{x}_j\}$: we are interested in studying its expectation, $\mathbb{E}_{\mathbf{x}_i}[\tau_{\{\mathbf{x}_j\}}]$. The formula we gave in Section 2.2.3 is directly obtained from the more general one given by L. Lovász in [85]: we recall it simply because we will need it later. On the other hand, the formula given in Section 2.2.3 is obtained from the (D_n) process and the procedure is inspired by those used in [82].

Analysis of $\mathbb{E}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}]$ using the spectrum of $\mathcal{P}$.

Definition 2.8. Let H be the $2^N \times 2^N$ symmetric matrix having as $(i,j)^{\text{th}}$-entry: $(H)_{ij} = H(i,j) = \mathbb{E}_{\mathbf{x}_i}[\tau_{\{\mathbf{x}_j\}}]$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{H}_N$. Clearly $H(i,i) = 0$ for all i.

The N-regularity of the graph implies that:

$$H(i,j) = 1 + \sum_{\{k|h(i,k)=1\}} \mathcal{P}_{ik} H(k,j) = 1 + \frac{1}{N} \sum_{\{k|h(i,k)=1\}} H(k,j) \quad \text{for } i \neq j \tag{2.4}$$

To relate the hitting time with the spectrum, we first define $F := \mathcal{J}_{2^N} + \mathcal{P}H - H$, where $\mathcal{J}_{2^N}$ is a $2^N \times 2^N$ matrix whose entries are all 1. From Equation (2.4), it follows that F is a diagonal matrix, as $(H)_{ij} = (\mathcal{J}_{2^N})_{ij} + (\mathcal{P}H)_{ij}$ for $i \neq j$. Moreover $F'\boldsymbol{\pi} = \mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)'$, since

$$F'\boldsymbol{\pi} = \left(\mathcal{J}_{2^N} + (\mathcal{P} - \mathcal{I}_{2^N})H\right)' \boldsymbol{\pi} = \mathcal{J}_{2^N}\boldsymbol{\pi} + H'(\mathcal{P} - \mathcal{I}_{2^N})'\boldsymbol{\pi} = \mathcal{J}_{2^N}\boldsymbol{\pi} + H'(\mathcal{P}'\boldsymbol{\pi} - \boldsymbol{\pi}) = \mathcal{J}_{2^N}\boldsymbol{\pi} = \mathbf{1}$$

Therefore, we deduce that $F = 2^N \mathcal{I}_{2^N}$ and H is solution of

$$(\mathcal{I}_{2^N} - \mathcal{P})H = \mathcal{J}_{2^N} - 2^N \mathcal{I}_{2^N} \tag{2.5}$$

Theorem 2.2.6. *Given a SRW on $\mathcal{H}_N$, the hitting time from vertex i to j is given by:*

$$H(i,j) = 2^N \sum_{k=2}^{2^N} \frac{1}{1-\lambda_k}(v_{kj}^2 - v_{ki}v_{kj}), \tag{2.6}$$

where λ_k is the k^{th}-eigenvalue of $\mathcal{P}$ and v_{ki} corresponds to the i^{th}-component of the k^{th}-eigenvector of $\mathcal{P}$, as given in Section 2.2.1.

Proof. We can not directly solve equation (2.5), since matrix $(\mathcal{I}_{2^N} - \mathcal{P})$ is singular. The spectral decomposition theorem insures that $\mathbb{R}^{2^N} = \oplus_{i=1}^{2^N} \text{Span}\{\mathbf{v}_i\}$. On the subspace $\oplus_{i=2}^{2^N} \text{Span}\{\mathbf{v}_i\}$, $(\mathcal{I}_{2^N} - \mathcal{P})$ is invertible. At the same time, the right hand side in (2.5) reduces to a constant times the identity matrix when restricted to this same subspace. Thus a possible candidate solving (2.5) is:

$$\tilde{H} = -2^N \sum_{i=2}^{2^N} (1-\lambda_i)^{-1} \mathbf{v}_i \mathbf{v}_i'$$

Nevertheless, for every vector $\mathbf{w} \in \mathbb{R}^{2^N}$, $\tilde{H} + \mathbf{1}\mathbf{w}'$ is a solution of (2.5) as well. Thus H can be unambiguously determined by imposing the condition over its main diagonal: $H(i,i) = 0$ for all $i \in \{0, \ldots, 2^N\}$. □

Analysis of $\mathbb{E}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}]$ from the D_n viewpoint.

For the sake of simplicity, we denote $H(D_0) := \mathbb{E}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}]$ as it depends only on the initial Hamming distance of $\mathbf{X}_0$ to $\overline{\mathbf{x}}$, D_0.

Remark 4. Due to (2.1), starting at point $\mathbf{x}_0$ with $D_0 = \overline{d}$, we have:

$$\begin{cases} \mathbb{P}(D_1 = \overline{d}+1 \,|\, D_0 = \overline{d}) =: q(\overline{d}, \overline{d}+1) = (N-\overline{d})/N \\ \mathbb{P}(D_1 = \overline{d}-1 \,|\, D_0 = \overline{d}) =: q(\overline{d}, \overline{d}-1) = \overline{d}/N \end{cases}$$

We are now able to define a new recursive formula for (2.4), which will be more convenient if evaluated explicitly:

$$H(\overline{d}) = 1 + \frac{N-\overline{d}}{N} H(\overline{d}+1) + \frac{\overline{d}}{N} H(\overline{d}-1) \tag{2.7}$$

with boundary conditions:

$$H(0) = 0 \text{ and } H(1) = 2^N - 1 = \sum_{j=0}^{N} C_N^j - 1 \tag{2.8}$$

Taking the difference $\Delta(\overline{d}) := H(\overline{d}) - H(\overline{d}-1)$, we obtain:

$$\Delta(\overline{d}+1) = H(\overline{d}+1) - H(\overline{d}) = \frac{\overline{d}}{N}\left(\Delta(\overline{d}+1) + \Delta(\overline{d})\right) - 1$$

And finally:

$$\Delta(\overline{d}+1) = \frac{\overline{d}}{N-\overline{d}}\Delta(\overline{d}) - \frac{N}{N-\overline{d}} \quad \text{with } \Delta(1) = H(1) \tag{2.9}$$

Then we can prove by iteration the following result:

Theorem 2.2.7. *Given a SRW on $\mathcal{H}_N$, the hitting time to cover a Hamming distance equal to $\overline{d}$, $H(\overline{d})$ with $0 \leq \overline{d} \leq N$ is obtained as:*

$$H(\overline{d}) = \sum_{d=0}^{\overline{d}-1} \frac{\sum_{j=1}^{N-1-d} C_N^{d+j} + 1}{C_{N-1}^d} \tag{2.10}$$

Proof. One have to prove that:

$$\Delta(d+1) = \frac{\sum_{j=1}^{N-1-d} C_N^{d+j} + 1}{C_{N-1}^d} \tag{2.11}$$

$$\begin{aligned}\Delta(d+1) &= \frac{d\cdot\Delta(d)}{N-d} - \frac{N}{N-d} = \frac{d}{N-d}\left(\frac{(d-1)\cdot\Delta(d-1)}{N-(d-1)} - \frac{N}{N-(d-1)}\right) - \frac{N}{N-d} \\ &= \frac{d(d-1)\cdot\Delta(d-1)}{(N-d)(N-(d-1))} - N\left(\frac{d}{(N-d)(N-(d-1))} + \frac{1}{N-d}\right) \qquad (2.12)\end{aligned}$$

Proceeding by iteration we obtain two terms, where the first one multiplies $\Delta(1)$. From Equation (2.9) we know that $\Delta(1) = H(1) = \sum_{j=0}^{N} C_N^j - 1$. A convenient use of the properties of the factorial operator allows us to reach the following expression:

$$\begin{aligned}(2.12) &= \frac{d!(N-1-d)!}{(N-1)!}\left(\sum_{j=0}^{N} C_N^j - 1\right) - N\left(\frac{d!(N-1-d)!}{(N-1)!} + \frac{d!(N-1-d)!}{2!(N-2)!} + \cdots\right. \\ &\quad \left. + \frac{d!(N-1-d)!}{(d-1)!(N-(d-1))!} + \frac{d!(N-1-d)!}{d!(N-d)!}\right) = \\ &= \frac{d!(N-1-d)!}{(N-1)!}\left(1 + \sum_{j=1}^{N-1-d} \frac{N!}{(d+j)!(N-(d+j))!}\right) = \frac{\sum_{j=1}^{N-1-d} C_N^{d+j} + 1}{C_{N-1}^d}\end{aligned}$$

By using again (2.9), we can now easily express $H(\overline{d})$ in the following way

$$H(\overline{d}) = \sum_{d=0}^{\overline{d}-1} \Delta(d+1) = \sum_{d=0}^{\overline{d}-1} \frac{\sum_{j=1}^{N-1-d} C_N^{d+j} + 1}{C_{N-1}^d}$$

which can be evaluated for reasonable values of N. □

We can immediately observe that $H(\overline{d})$ is a monotonically increasing function. Moreover, H is concave. Indeed, thanks to Proposition 2.2.7 we can prove that $\forall\, d \in \{1, \ldots, N-1\}$:

$$H(d) - H(d-1) \geq H(d+1) - H(d) \iff \Delta(d) \geq \Delta(d+1)$$

Furthermore, we can evaluate the following limit:

$$\lim_{N\to\infty} \frac{H(\alpha N)}{2^N} \quad \text{for } \alpha \in]0, 1]. \qquad (2.13)$$

Remark 5. The case $\alpha = 0$ is trivial: if $\alpha = 0$ this limit is equal to 0 since $H(0) = 0$.

Remark 6. Proposition 2.2.8 below, which evaluates (2.13), confirms the statement made in Remark 3: as N goes to infinity, (D_n) goes quickly to $N/2$ and then $H(d)$ is always of order $\sim 2^N$ irrespective of $d \neq 0$.

Proposition 2.2.8. *For all* $\alpha \in]0,1]$*:*

$$\lim_{N\to\infty} \frac{H(\alpha N)}{2^N} = 1$$

Proof. Since H is an increasing function and by using Equation (2.10) we have:

$$2^N - 1 = H(1) \leq H(\alpha N) \leq H(N) = \sum_{d=0}^{N-1} \frac{1}{C_{N-1}^d} + \sum_{d=0}^{N-1} \sum_{j=1}^{N-1-d} \frac{C_N^{d+j}}{C_{N-1}^d} =: S_1 + S_2$$

We examine the two terms of the last member separately.

$$S_1 \leq 2 + \frac{2}{N-1} + (N-4)\frac{2}{(N-1)(N-2)} \tag{2.14}$$

We can prove it just by looking at Pascal's triangle.

Now, if we consider S_2, we see that there is no contribution for $d = N-1$, as the internal sum is zero valued. Moreover we have:

$$\sum_{j=1}^{N-1-d} C_N^{d+j} \leq \sum_{j=0}^{N} C_N^j = 2^N$$

And so:

$$S_2 \leq 2^N \sum_{d=0}^{N-2} \frac{1}{C_{N-1}^d} \overset{(2.14)}{\leq} 2^N \left(1 + \frac{2}{N-1} + (N-4)\frac{2}{(N-1)(N-2)}\right)$$

By putting together all these inequalities and dividing by factor 2^N we get that:

$$1 - \frac{1}{2^N} \leq \frac{H(\alpha N)}{2^N} \leq 1 + \frac{2}{N-1} + \frac{2(N-4)}{(N-1)(N-2)} + \frac{1}{2^N}\left(2 + \frac{2}{N-1} + \frac{2(N-4)}{(N-1)(N-2)}\right)$$

The result comes directly by applying the squeeze theorem. □

This result can be extended to a SRW on a generic state-space $\mathcal{S}^N$, with $|\mathcal{S}| = s$. More precisely, one can prove in a similar way as we did for $\mathcal{H}_N$ the following result:

Proposition 2.2.9. *The order of magnitude of the hitting time for a switch-type mutational model on the state-space* $\mathcal{S}^N$*, with* $|\mathcal{S}| = s$*, is* s^N*, for* N *big enough.*

This is the consequence of Theorem 2.2.10 and Proposition 2.2.11 below.

Theorem 2.2.10. *Given a SRW on S^N, the hitting time to cover a Hamming distance equal to $\overline{d}$, $H^s(\overline{d})$ with $0 \leq \overline{d} \leq N$ is obtained as:*

$$H^s(\overline{d}) = \sum_{d=0}^{\overline{d}-1} \frac{\sum_{j=d+1}^{N} C_N^j (s-1)^j}{C_{N-1}^d (s-1)^d} \tag{2.15}$$

Proposition 2.2.11. *For all $\alpha \in]0,1]$:*

$$\lim_{N\to\infty} \frac{H^s(\alpha N)}{s^N} = 1$$

Remark 7. In the current Section and in Section 2.3 we evaluate the expected hitting time to reach a specific vertex of $\mathcal{H}_N$. From a biological viewpoint this means to reach the optimal B-cell trait against the presented antigen. The single-peak landscape assumption has already been discussed in other mathematical models of GC reaction [121, 70, 69]. Looking for a perfect complementarity of the whole BCR to the target profile might not be really biologically significant: the matching of entire strings means designing a receptor for each possible antigen, this is not reasonable considering repertoire sizes. Therefore, we evaluate the hitting time of a set of vertices instead. This implies, of course, a speed-up of the time-scales (see Table 2.1 for instance). Let $A_r := \{\mathbf{x}_i \in \mathcal{H}_N \,|\, h(\mathbf{x}_i, \overline{\mathbf{x}}) \leq r\}$ be the sphere of radius r in the graph metric, centered in the target vertex $\overline{\mathbf{x}}$, and considering $\mathcal{P}$ as transition probability matrix. We are interested in explicitly evaluate the mean hitting time to enter A_r. We consider the distances process defined in Section 2.2.2, hence the graph underlined by matrix $\mathcal{Q}$ (Proposition 2.2.3). The sphere A_r can be characterized as:

$$A_r := \{j \in \{0, \dots, N\} \,|\, j \leq r\}$$

We denote by $H_i(r)$ the expected time to reach A_r starting from initial Hamming distance i. By using Equation (2.1), we obtain:

$$\begin{cases} H_i(r) = 0 & \text{if } i \leq r \\ H_i(r) = 1 + \dfrac{i}{N} H_{i-1}(r) + \dfrac{N-i}{N} H_{i+1}(r) & \text{if } i > r \end{cases} \tag{2.16}$$

Let us define $\Delta_r(i)$ as the difference between $H_i(r)$ and $H_{i-1}(r)$:

$$\Delta_r(i) := H_i(r) - H_{i-1}(r)$$

Therefore:

$$\begin{aligned}\Delta_r(i) &= 1 + \frac{i}{N} H_{i-1}(r) + \frac{N-i}{N} H_{i+1}(r) - H_{i-1}(r) \\ &= 1 + \frac{N-i}{N} \left(H_{i+1}(r) - H_{i-1}(r) \right) \\ &= 1 + \frac{N-i}{N} \left(\Delta_r(i+1) + \Delta_r(i) \right)\end{aligned}$$

And finally:

$$\Delta_r(i) = \frac{N-i}{i} \Delta_r(i+1) + \frac{N}{i} \tag{2.17}$$

With the condition:

$$\Delta_r(N) := H_N(r) - H_{N-1}(r) = 1 + H_{N-1}(r) - H_{N-1}(r) = 1 \tag{2.18}$$

Theorem 2.2.12. *For all $i > r \geq 0$ the mean hitting time to reach A_r starting from initial Hamming distance i from $\overline{\mathbf{x}}$ is given by:*

$$H_i(r) = \sum_{s=r+1}^{i} \frac{\sum_{j=0}^{N-s} C_N^j}{C_{N-1}^{N-s}} \tag{2.19}$$

Table 2.1: Average expected times to reach the sphere A_r of radius r centered in $\overline{\mathbf{x}}$, for different values of r. Simulations correspond to $N = 10$ and an initial Hamming distance $h(\mathbf{X}_0, \overline{\mathbf{x}}) = 10$. Table 2.1 shows results obtained over 20480 simulations. We denote by $|A_r|$ the number of vertices of $\mathcal{H}_N$ included in A_r. $H_{10}(r)$ corresponds to the theoretical value obtained by Equation (2.19). We denote by $\widehat{\tau_{\{\overline{\mathbf{x}}\}}}_n$ the average value obtained over $n = 20480$ simulations and by $\widehat{\sigma}_n$ its corresponding estimated standard deviation.

r	$\lvert A_r \rvert$	$H_{10}(r)$	$\widehat{\tau_{\{\overline{\mathbf{x}}\}}}_n$	$\frac{\widehat{\sigma}_n}{\sqrt{n}}$
0	1	1186.540	1184.499	8.1736
1	11	163.540	163.747	1.064
2	56	50.984	51.729	0.298
3	176	24.095	24.118	0.116

Remark 8. One can demonstrate that $H_i(0) = H(i)$ as defined by Equation (2.10).

Proof. Considering Equations (2.17) and (2.18) we can demonstrate by iteration

that $\forall\, k \in \{0, \ldots, N-1\}$:

$$\Delta_r(N-k) = \frac{1}{C_{N-1}^k} \sum_{j=0}^{k} C_N^j \tag{2.20}$$

The result follows by observing:

$$H_i(r) = \sum_{s=r+1}^{i} \Delta_r(s) = \sum_{s=r+1}^{i} \Delta_r(N-(N-s)) \tag{2.21}$$

□

We simulate the average expected time to reach a sphere of radius r centered in the vertex $\overline{\mathbf{x}}$, for different values of r. Table 2.1 shows the results obtained over more than 20000 simulations. We clearly see that the average hitting time decreases significantly if we consider bigger radius r, as expected.

2.3 More mutational models: how does the structure of the hypercube change?

In this section, we explore other mutation rules, which change the internal graph structure of the hypercube, therefore the dynamics of the RW and the characteristic time-scales of the exploration of the state-space.

2.3.1 Study of various mutation rules

In this section, we study four mutation rules:

- a model of permutation of two bits;
- a model of switch of k-length strings;
- a model of switch of 1 or 2-length strings depending on the Hamming distance to a fixed node representing the antigen target cell;
- multiple point mutations models.

The exchange mutation model.

We consider a model where given an initial B-cell representing string, each mutation step consists in permuting two randomly chosen bits.

Definition 2.9. Let $\mathbf{X}_n \in \{0,1\}^N$ be the BCR at step n. Let $i \in \{1,\dots,N\}$, $j \in \{1,\dots,N\} \setminus \{i\}$ two randomly chosen indexes. We can suppose, without loss of generality, that $j > i$:

$$\mathbf{X}_{n+1} = (X_{n,1},\dots,X_{n,i-1},X_{n,j},X_{n,i+1},\dots,X_{n,j-1},X_{n,i},X_{n,j+1},\dots,X_{n,N})$$

With this mutation rule, we loose a very important property: the connectivity of the graph. We denote by $\mathcal{H}_{(s)} \subset \{0,1\}^N$ the set containing the C_N^s vertices having s 1 in their strings. The state-space $\{0,1\}^N$ is divided into $N+1$ connected components: $\mathcal{H}_{(s)}$, $0 \le s \le N$.

Proposition 2.3.1. *There are exactly* $\frac{N(N-1)}{2}$ *(non-oriented) edges ending at each vertex counting the possible loops. Each node* $\mathbf{x} \in \mathcal{H}_{(s)}$ *has exactly* $\frac{(N-s)^2-(N-s^2)}{2}$ *loops.*

Corollary 2.3.2. $\mathbb{P}(\mathbf{X}_n = \mathbf{x}_j | \mathbf{X}_{n-1} = \mathbf{x}_j) = \frac{(N-s)^2-(N-s^2)}{N(N-1)}$. *In particular, the probability of remaining on the same node is 1 if* $s = 0$ *or* $s = N$.

Proof. (Proposition 2.3.1) The first statement is obtained by simple combinatory arguments. Let us consider $\mathbf{x} \in \mathcal{H}_{(s)}$ with $0 \le s \le N$: it is composed by exactly s ones and $N-s$ zeros. For the sake of clarity let us consider that $\{0,\dots,N\} = I \sqcup J$ so that $|I| = s$, $|J| = N-s$ and $x_i = 1\ \forall\, i \in I$, $x_j = 0$ $\forall\, j \in J$. We obtain a loop each time we choose both random indices either in I (C_s^2 possibilities) or in J (C_{N-s}^2 possibilities). Then the total number of loops is obtained by the sum of these two cases, *i.e.* $\frac{(N-s)^2-(N-s^2)}{2}$. □

We can also describe qualitatively the behavior of the (D_n) process referring to this current model. As a general principle, we have that $D_n = D_{n-1} + i$, $i \in \{0,\pm 2\}$. Therefore, clearly $\mathbb{P}(D_n = d' | D_{n-1} = d) = 0$ if $|d'-d| > 2$ or $|d'-d| = 1$. Moreover, we have maximal and minimal values of D_n depending on s_0 and $\overline{s}$ so that $\mathbf{X}_0 \in \mathcal{H}_{(s_0)}$ and $\overline{\mathbf{x}} \in \mathcal{H}_{(\overline{s})}$. Indeed:

Proposition 2.3.3. *Given* $\overline{\mathbf{x}} \in \mathcal{H}_{(\overline{s})}$ *and* $\mathbf{X}_0 \in \mathcal{H}_{(s_0)}$, *then* $\forall\, n \ge 0$*:*

$$\begin{cases} |\overline{s} - s_0| \le D_n \le \overline{s} + s_0 & \text{if } \overline{s} + s_0 \le N \\ |\overline{s} - s_0| \le D_n \le (N-\overline{s}) + (N - s_0) & \text{if } \overline{s} + s_0 > N \end{cases}$$

Proof. The proof follows immediately by counting how many possibilities there are to arrange s ones and $N-s$ zeros in a N-length string. □

Remark 9. From Proposition 2.3.3 one can see that if $\overline{s} = s_0 =: s$ and $2s \neq N$ then:

$$0 \leq D_n < N$$

Class switch of k-length strings.

Let $\mathbf{X}_0 = (X_{0,1}, \ldots, X_{0,N}) \in \{0,1\}^N$ be the B-cell entering the somatic hypermutation process. At each mutation step we switch the class of k consecutive amino-acids.

Definition 2.10. Let $\mathbf{X}_n \in \{0,1\}^N$ be the BCR at step n. Let $i \in \{1, \ldots, N-(k-1)\}$ be a randomly chosen index. Then $\mathbf{X}_{n+1} := (X_{n,1}, \ldots, X_{n,i-1}, 1 - X_{n,i}, \ldots, 1 - X_{n,i+k-1}, X_{n,i+k}, \ldots, X_{n,N})$.

Remark 10. If $k = 1$ we are in the case of a SRW on $\mathcal{H}_N$.
If $k = N$ we stay on a 2-length cycle. Indeed we have that $\mathbf{X}_l = \mathbf{X}_0$ for l even and $X_l = \mathbf{1} - \mathbf{X}_0$ for l odd. For this reason the case $k = N$ does not appear interesting neither from a mathematical nor from a biological point of view.

Here below we give some basic properties of this RW, that one can easily prove by simple combinatory arguments.

Proposition 2.3.4. *Each vertex has exactly $N-(k-1)$ neighbors and no loops. Therefore, for all $\mathbf{x}_i$, $\mathbf{x}_j$ in $\{0,1\}^N$:*

$$\mathbb{P}(\mathbf{X}_n = \mathbf{x}_j | \mathbf{X}_{n-1} = \mathbf{x}_i) =: p_k(i,j) = \begin{cases} \dfrac{1}{N-(k-1)} & \text{if } \mathbf{x}_j \sim \mathbf{x}_i \\ \\ 0 & \text{otherwise} \end{cases}$$

Remark 11. As regards to this current model, given $\mathbf{x}_i$, $\mathbf{x}_j \in \{0,1\}^N$, we have: $\mathbf{x}_i \sim \mathbf{x}_j \Leftrightarrow h(\mathbf{x}_i, \mathbf{x}_j) = k$ and the k different elements have consecutive indexes.

Thus, $\mathcal{P}_k = (p_k(\mathbf{x}_i, \mathbf{x}_j))_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{H}_k}$ is the $2^N \times 2^N$ transition probability matrix.

For fixed $k \in \{1, \ldots, N\}$ the graph underlying the RW corresponding to the model of class switch of k-length strings has exactly 2^{k-1} connected components, each one composed of $2^{N-(k-1)}$ elements.
Because of the non connectivity of the graph, we can focus on the connected component to which $\mathbf{X}_0$ belongs and find out the properties of our RW on it. For fixed N and k and dealing with each connected component separately, we are describing a SRW on a $(N-(k-1))$-hypercube. Henceforth we obtain 2^{k-1} distinct hypercube-type structures of the same size.

We can limit our study to the connected component containing $\mathbf{X}_0$, which is, up to a change of variables, a $(N-(k-1))$-dimensional hypercube. Let $\overline{\mathcal{P}}_k$ be the restriction of $\mathcal{P}_k$ to this connected component. If we conveniently order the $2^{N-(k-1)}$ distinct vertices, than $\overline{\mathcal{P}}_k = \mathcal{P}_{N-(k-1)}$. At this stage, it is possible to translate all classical results we know about the SRW on $\mathcal{H}_n$, for $n = N-(k-1)$, on each connected component of this current graph, remembering the definition of neighborhood given in Remark 11.

Class switch of 1 or 2-length strings depending on the Hamming distance to $\overline{\mathbf{x}}$.

The exchange mutation model and the model of class switch of k-length strings present an important limitation: the underlying graphs are non-connected. Due to our choice of affinity, a model which does not enable to explore the whole state-space is not very relevant. Indeed, if the graph is non-connected and the target chain does not belong to the connected component containing the B-cell which first enters the somatic hypermutation process, then we never reach the target configuration. From a biological viewpoint, it may be more relevant to consider a smoother affinity model, in which the BCR representing string reaches the target when most, but not all, bits are similar. In this case, considering a non-connected graph, is not necessarily a problem.

Another way to overcome the problem of non-connectivity is to consider a model which allows to vary the length of the strings submitted to switch-type mutations. Moreover, it is biologically credible that during the GC process B-cells can modify their mutation rate, making it somehow proportional to their affinity to the antigen [22, 17, 55]. Indeed, B-cells compete for different rescue signals (from Helper T-cells or FDCs), and that determines their fate: undergo further mutations or differentiate into plasma cells or memory cells ([1], Chapter 7). Here we suppose that the mutational rate is inversely proportional to the affinity: the greater the affinity, the lower is the mutational rate. We found the hypothesis that the regulation of the hypermutation process is dependent on receptor affinity also in other works, as [30, 2], where the authors proposed computational implementations of the clonal selection principle to design genetic optimization algorithms, taking into account AAM during an adaptive immune response. In terms of our mathematical model, we can translate it by making the size k of the strings which can mutate to be directly proportional to the Hamming distance to $\overline{\mathbf{x}}$ at each mutation step:

$k_n = f(D_n)$, with $f : \{0, \dots, N\} \to \{0, \dots, N\}$ being an increasing function.

Despite many choices of the function f are possible, hereinafter we consider a very elementary case, where f is a step function on two intervals.

Definition 2.11. Let $\mathbf{X}_n \in \{0,1\}^N$ be the BCR at step n. We denote by k_n:

$$k_n := f(D_n) = \begin{cases} 1 & \text{if } D_n \leq 1 \\ 2 & \text{if } D_n > 1 \end{cases}$$

Let $i \in \{1, \ldots, N - (k_n - 1)\}$ be a randomly chosen index. Then:
$\mathbf{X}_{n+1} := (X_{n,1}, \ldots, X_{n,i-1}, 1 - X_{n,i}, \ldots, 1 - X_{n,i+k_n-1}, X_{n,i+k_n}, \ldots, X_{n,N})$.

This model is an interesting and simple way to generalize the basic mutational model without losing the property of connectivity of the graph. The addition of this flexibility was not only motivated by biological reasons, but we also expect that this modification decreases the hitting time to a fixed node. This is actually true: the hitting time is halved compared to the basic model (at least for N big enough). We will also show that the stationary distribution is concentrated on a half part of the hypercube, the one to whom $\overline{\mathbf{x}}$ belongs.

Remark 12. For fixed N and $k = 2$ the graph is divided into two connected components composed of 2^{N-1} vertices. Two nodes belonging to the same connected component have a Hamming distance of $2t$ with $0 \leq t \leq \lfloor N/2 \rfloor$. On the other hand, two vertices belonging to different connected components have a Hamming distance of $(2t+1)$ with $0 \leq t \leq \lfloor (N-1)/2 \rfloor$.

In order to analyze this process, we have to distinguish two cases. For fixed N and $\overline{\mathbf{x}}$, the process we obtain:

case 1: $D_0 = 2t$, $t > 0$. $\mathbf{X}_0$ belongs to the same connected component as $\overline{\mathbf{x}}$, so we are working on a $(N-1)$-dimensional hypercube, following the model of class switch of 2-length strings. we stay in this connected component all over the process till we arrive at $\overline{\mathbf{x}}$, as it is impossible to obtain a Hamming distance equal to 1.

case 2: $D_0 = 2t + 1$, $t > 0$. We necessarily take $k = 2$ and Remark 12 implies that $\mathbf{X}_0$ belongs to a different connected component than $\overline{\mathbf{x}}$. In order to reach the connected component containing $\overline{\mathbf{x}}$, we have to visit a node $\mathbf{x}^*$ so that $h(\mathbf{x}^*, \overline{\mathbf{x}}) = 1$, and $|\{\mathbf{x}^* \,|\, h(\mathbf{x}^*, \overline{\mathbf{x}}) = 1\}| = N$. Then, if $D_0 = 1$ we are allowed to change only one element of the B-cell representing string. With probability $1/N$ we arrive directly at $\overline{\mathbf{x}}$ and with probability $(N-1)/N$ we obtain $D_1 = 2$. Then we go back to case 1.

Proposition 2.3.5. *The graph corresponding to the current model is divided into two connected components: $\mathcal{H}_N^{(1-2)}$ and its complementary $\overline{\mathcal{H}_N}^{(1-2)}$, s.t.*

$\overline{\mathbf{x}} \in \overline{\mathcal{H}_N}^{(1-2)}$. $\overline{\mathcal{H}_N}^{(1-2)}$ *is accessible from* $\mathcal{H}_N^{(1-2)}$*, but not conversely. Vertices belonging to* $\overline{\mathcal{H}_N}^{(1-2)}$ *are positive recurrent and vertices belonging to* $\mathcal{H}_N^{(1-2)}$ *are transient.*

Proof. The existence of two connected components depends on the use of the model of switch of 2-length strings. Indeed the structure of the graph we are considering here essentially corresponds to that of the graph underlying the model of switch of 2-length strings, up to the addition of some oriented edges from $\mathcal{H}_N^{(1-2)}$ to $\overline{\mathcal{H}_N}^{(1-2)}$. As long as we stay in $\overline{\mathcal{H}_N}^{(1-2)}$ or $\mathcal{H}_N^{(1-2)}$ we are just allowed to switch 2-length strings. Moreover, we have already observed that when we are in $\overline{\mathcal{H}_N}^{(1-2)}$ we can't exit, while when we are in $\mathcal{H}_N^{(1-2)}$ we can reach $\overline{\mathcal{H}_N}^{(1-2)}$ by visiting one among the N nodes having Hamming distance 1 from $\overline{\mathbf{x}}$, and that happens in a finite number of steps. Therefore:

$$\begin{cases} \mathbb{P}(\tau_{\mathbf{x}_i} < \infty) = 1 \quad \text{for all } \mathbf{x}_i \in \overline{\mathcal{H}_N}^{(1-2)} \quad \Rightarrow \quad \mathbf{x}_i \text{ is recurrent} \\ \\ \mathbb{P}(\tau_{\mathbf{x}_i} < \infty) < 1 \quad \text{for all } \mathbf{x}_i \in \mathcal{H}_N^{(1-2)} \quad \Rightarrow \quad \mathbf{x}_i \text{ is transient} \end{cases}$$

In particular, vertices belonging to $\overline{\mathcal{H}_N}^{(1-2)}$ are positive recurrent as the chain is irreducible on $\overline{\mathcal{H}_N}^{(1-2)}$ and $|\overline{\mathcal{H}_N}^{(1-2)}| < \infty$. □

The following known result about stochastic processes, justifies Corollary 2.3.7 below.

Theorem 2.3.6. *Let* $(\mathbf{X}_n)_{n\geq 0}$ *be a Markov chain on a state-space* $\mathcal{S}$ *and* $\mathbf{x}_i \in \mathcal{S}$ *be positive recurrent. Let* m_i *be the mean return time:* $m_i = \mathbb{E}(\tau_{\{\mathbf{x}_i\}} \,|\, \mathbf{X}_0 = \mathbf{x}_i)$*. Denoting by* $\mathcal{S}_r \subseteq \mathcal{S}$ *the positive recurrent connected component to which* $\mathbf{x}_i$ *belongs, then a stationary distribution* $\overline{\boldsymbol{\pi}}$ *is given by:*

$$\overline{\pi}_i = m_i \quad \forall \mathbf{x}_i \in \mathcal{S}_r$$

$$\overline{\pi}_i = 0 \quad \forall \mathbf{x}_i \in \mathcal{S} \setminus \mathcal{S}_r$$

Theorem 2.3.6 is proven by considering the relations among recurrent and transient classes, stationary distributions and return time (see [104] for some more details).

Corollary 2.3.7. *The stationary distribution for the RW we describe in the*

present section, $\overline{\boldsymbol{\pi}}$*, is given by:*

$$\overline{\pi_i} = \begin{cases} \dfrac{1}{2^{N-1}} & \text{if } \mathbf{x}_i \in \overline{\mathcal{H}_N}^{(1-2)} \\ \\ 0 & \text{if } \mathbf{x}_i \in \mathcal{H}_N^{(1-2)} \end{cases} \tag{2.22}$$

Corollary 2.3.7 is a consequence of Theorem 2.3.6 and the study of the SRW on an N-dimensional hypercube.

Allowing 1 to k mutations

In this section we analyze how the N-dimensional hypercube changes if we allow 1 to k independent switch-type mutations at each step, with k fixed, $k \leq N$.

Definition 2.12. Let $\mathbf{X}_n \in \{0,1\}^N$ be the BCR at step n. Let k be an integer, $1 \leq k \leq N$ and $\forall\, i$, $1 \leq i \leq k$, $a_i := \mathbb{P}(i \text{ independent switch-type mutations})$. Then with probability a_i, $\mathbf{X}_{n+1}$ is obtained from $\mathbf{X}_n$ by repeating i times, independently, the process described by Definition 2.5.

By definition, the corresponding transition probability matrix is a convex combination of $\mathcal{P}^i$, for $1 \leq i \leq k$ ($\mathcal{P}^i$ is the transition probability matrix corresponding to i iterations of the process of a single bit mutation):

$$\sum_{i=1}^{k} a_i \mathcal{P}^i, \quad \text{with} \sum_{i=1}^{k} a_i = 1. \tag{2.23}$$

Definition 2.13. Let us fix $a_i = 1/k \;\forall\, i$. We denote by $\mathcal{P}^{(k)} := 1/k \sum_{i=1}^{k} \mathcal{P}^i$. Accordingly, we denote the graph underlying this RW $\mathcal{H}_N^{(k)}$.

Remark 13. Since the mutations are assumed to be independent, then k represents the maximum Hamming distance the process can cover in a single mutation step. Thanks to the independence of each single mutation, two or more mutations may nullify their respective action: in particular for $k \geq 2$ there is a non-zero probability of remaining at the same position. From a biological point of view, this behavior can be interpreted as the possibility of doing mutations which have no effect on the BCR structure.

We can now evaluate the eigenvalues of $\mathcal{P}^{(k)}$, $\lambda_j^{(k)}$ by using the eigenvalues λ_j of $\mathcal{P}$ (Section 2.2.1). Due to the fact that all $\mathcal{P}^i$ commute with each other, the eigenvalues are given by:

$$\lambda_j^{(k)} = \frac{1}{k} \sum_{i=1}^{k} \lambda_j^i \tag{2.24}$$

and $\mathcal{P}^{(k)}$ and $\mathcal{P}$ have the same eigenvectors. We give explicitly the expression of all $\lambda_i^{(k)}$ and concentrate on the second largest eigenvalue, $\lambda_2^{(k)}$.

Proposition 2.3.8. *The $N+1$ distinct eigenvalues of matrix $\mathcal{P}^{(k)}$ are:*

- $\lambda_1^{(k)} = 1$;
- $\lambda_j^{(k)} = \frac{\lambda_j}{k} \cdot \frac{1-\lambda_j^k}{1-\lambda_j}$ *for* $2 \leq j \leq N$;
- $\lambda_{N+1}^{(k)} = \frac{1}{2k}\left((-1)^k - 1\right) = \begin{cases} 0 & \text{if } k \text{ is even} \\ -1/k & \text{if } k \text{ is odd} \end{cases}$

The multiplicity of $\lambda_j^{(k)}$ is $\binom{N}{j-1}$, $1 \leq j \leq N+1$

Proof. This result comes directly from the evaluation of Equation (2.24), for the already known values of all λ_j (Corollary 2.2.2). □

Then, in particular, the second largest eigenvalue of $\mathcal{P}^{(k)}$ is:

$$\lambda_2^{(k)} = \frac{N-2}{2k}\left(1 - \left(1 - \frac{2}{N}\right)^k\right) \tag{2.25}$$

Remark 14. For all $k \geq 2$, $\lambda_2 > \lambda_2^{(k)}$. First of all, we can observe that $\lambda_2^{(k)}$ decreases for increasing k. Therefore:

$$\lambda_2 - \lambda_2^{(k)} \geq \lambda_2 - \lambda_2^{(2)} = \frac{N-2}{4N^2}(4N - N^2 + (N-2)^2) = \frac{N-2}{N^2} > 0$$

For $N \gg 1$, the series expansion of $\lambda_2^{(k)}$ gives us:

$$\begin{aligned} \lambda_2^{(k)} &= \frac{N-2}{2k}\left(1 - \left(1 - \frac{2k}{N} + \frac{2k(k-1)}{N^2} + \mathcal{O}\left(\frac{1}{N^3}\right)\right)\right) \\ &= \frac{N-2}{N} - \frac{(N-2)(k-1)}{N^2} + \mathcal{O}\left(\frac{1}{N^2}\right) \end{aligned}$$

We can observe how the spectral gap changes. If we consider the series expansion of $\left(1 - \frac{2}{N}\right)^k$ for $N \to \infty$, we get:

$$\lambda_1^{(k)} - \lambda_2^{(k)} = \frac{2}{N} + \frac{(N-2)(k-1)}{N^2} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

It can be interesting to choose k as a function of N. Let us consider, for example, $k = \alpha N$, with $0 < \alpha \leq 1$. In this case, we have:

$$\begin{aligned}
\lambda_2^{(\alpha N)} &= \frac{N-2}{2\alpha N}\left(1-\left(1-\frac{2}{N}\right)^{\alpha N}\right) \\
&\overset{\text{for } N\to\infty}{=} \frac{N-2}{2\alpha N}\left(1-\left(e^{-2\alpha}+\mathcal{O}\left(\frac{1}{N}\right)\right)\right) \\
&= \frac{(N-2)\left(1-e^{-2\alpha}\right)}{2\alpha N}+\mathcal{O}\left(\frac{1}{N}\right) \to \frac{1-e^{-2\alpha}}{2\alpha} \text{ for } N\to\infty
\end{aligned}$$

We can observe that $\frac{1-e^{-2\alpha}}{2\alpha} =: \overline{\lambda}_2^{(\alpha N)}$ decreases when α increases. Moreover:

- $\overline{\lambda}_2^{(\alpha N)} \to 1$ for $\alpha \to 0$, which means that the spectral gap, $1-\lambda_2^{(\alpha N)}$ converges to zero for $N\to\infty$ and $\alpha\to 0$;
- If $\alpha = 1$ then $\overline{\lambda}_2^{(N)} = \frac{1}{2}-\frac{1}{2e^2}$. Therefore, the spectral gap is $\frac{1}{2}+\frac{1}{2e^2}$

The spectral gap indicates how quickly a RW converges to its stationary distribution. As expected, if $\alpha \to 0$ then the spectral gap gets close to 0. On the other hand for all $\alpha > 0$ the spectral gap tends to a strictly positive quantity, while the spectral gap corresponding to the case of the basic model converges to zero for $N\to\infty$. In particular, when $\alpha = 1$ (*i.e.* we are considering the optimal case, in which we are allowed to do among 1 and N mutations at each mutation step), the spectral gap, $\frac{1}{2}+\frac{1}{2e^2}$, is significantly bigger than the one obtained for the basic model, $2/N$.

2.3.2 Comparison of hitting times

In this section we compare hitting times referring to some relevant mutational models we have already presented. We do not consider models that entail non-connected graphs (the exchange mutation model and the model of class switch of k-length strings). Indeed, as we have already discussed in Section 2.3.1, the loss of graph connectivity implies a great lack of the model due to our choice of affinity. In Table 2.2 we collect most important characteristics of these RWs on $\{0,1\}^N$: the hitting time and its approximation for big N, that we will discuss in this current section, the stationary distribution and the value of the second larger eigenvalue when known.

Class switch of 1 or 2-length strings depending on the Hamming distance to $\overline{\mathbf{x}}$.

We use results obtained in Section 2.2 for the (D_n) process concerning the SRW on the N-dimensional hypercube and we apply them to this model. Here we shall introduce another definition of the distance, which is adapted to a

Table 2.2: Table 2.2 summarizes the main characteristics of most random processes we introduce and analyze in Sections 2.2 and 2.3.

Model	**Hitting time**	**Stationary distribution**	**Second biggest eigenvalue**
Basic model	$H(\overline{d}) = \sum_{d=0}^{\overline{d}-1} \frac{\sum_{j=1}^{N-1-d} C_N^{d+j}+1}{C_{N-1}^d} \sim 2^N$	$\boldsymbol{\pi}$	$1 - \frac{2}{N}$
Switch 1-2	$\sim 2^{N-1}$	$\boldsymbol{\pi}\big\vert_{\overline{\mathcal{H}_N}^{(1-2)}}$	-
Allowing 1 to k mutations	$\overline{T}_N^{(k)}(\overline{d}) = \sum_{l=2}^{2^N} \mu_l^{(k)} - \frac{1}{2^N C_N^{\overline{d}}} \sum_{l=2}^{2^N} \mu_l^{(k)} R_N(l, \overline{d})$	$\boldsymbol{\pi}$	$\frac{N-2}{2k}\left(1 - \left(\frac{N-2}{N}\right)^k\right)$

connected component $\mathcal{H}_{N,2} \subset \{0,1\}^N$, where $\mathcal{H}_{N,2}$ denotes one of the two parts in which $\{0,1\}^N$ is divided applying the model of class switch of 2-length strings. We recall that $\mathcal{H}_{N,2}$ is a $(N-1)$-dimensional hypercube, and that the graph underlying the model of class switch of 1 or 2-length strings corresponds essentially to the graph obtained with the model of switch of 2-length strings, up to the addition of some oriented edges from $\mathcal{H}_N^{(1-2)}$ to $\overline{\mathcal{H}_N}^{(1-2)}$.

Definition 2.14. For all $\mathbf{x}_i$, $\mathbf{x}_j \in \mathcal{H}_{N,2}$ we denote by $h^{(2)}(\mathbf{x}_i, \mathbf{x}_j)$ the number of edges in a shortest path connecting them. Simultaneously we denote by $D_n^{(2)} = h^{(2)}(\mathbf{X}_n, \overline{\mathbf{x}})$, $D_n^{(2)} \in \{0, \dots, N-1\}$ $\forall\, n \geq 0$.

Considering the process $(D_n^{(2)})_{n\geq 0}$, all results stated in Section 2.2 hold true. Furthermore, let us denote by $\mathbb{E}_{\mathbf{x}_i}^{(2)}[\tau_A]$ the expected number of steps before set $A \in \mathcal{H}_{N,2}$ is visited starting at $\mathbf{x}_i \in \mathcal{H}_{N,2}$ and following the model of switch of 2-length strings. Then, we also denote by $H_{N-1}^{(2)}(d) = \mathbb{E}_{\mathbf{x}}^{(2)}[\tau_{\{\overline{\mathbf{x}}\}}]$ where $d = h^{(2)}(\mathbf{x}, \overline{\mathbf{x}})$.

Remark 15. Clearly if $D_0 = 2t$ and $t > 0$, which means that $\mathbf{X}_0$ and $\overline{\mathbf{x}}$ belong to the same connected component in the model of class switch of 2-length strings, then the mean hitting time for the current model will be of the order of a half the mean hitting time for the basic model. Indeed, we are considering here a $(N-1)$-dimensional hypercube instead of a N-dimensional one.

The result below, which is an immediate application of the Ergodic Theorem, will help us understand better the general behavior of this mean hitting time:

Proposition 2.3.9. *Let $(\mathbf{X}_n)_{n\geq 0}$ be a SRW on $\mathcal{H}_N$. We denote by $T_d^+ :=$*

$\inf\{n \geq 1 \,|\, D_n = d\}$ *and* $T_d := \inf\{n \geq 0 \,|\, D_n = d\}$. *Then:*

$$\mathbb{E}_{D_0=d}[T_d^+] = \frac{2^N}{C_N^d} \tag{2.26}$$

Proof. The proof is obtained by applying the Ergodic Theorem to the (D_n) process and its stationary distribution, the binomial probability distribution. □

For the discussion we made in Section 2.2.2 and, in particular, Remark 3 we can conclude that for $N \gg 1$ the order of magnitude of the time we spend to reach the N nodes at Hamming distance 1 from $\overline{\mathbf{x}}$ is:

$$\mathbb{E}_{D_0=d}[T_1] \sim \frac{2^N}{N} \tag{2.27}$$

Then we can claim the following result, which comes directly from Equation (2.27):

Proposition 2.3.10. *Let us suppose that* $D_0 = 2t^* + 1$ *with* $0 < t^* \leq \lfloor (N-1)/2 \rfloor$. *Then for* $N \gg 1$ *we have:*

$$\mathbb{E}^{(2)}_{D_0=d}[T_1] \sim \frac{2^{N-1}}{N}$$

Finally:

Proposition 2.3.11. *We denote by* $\mathbb{E}^{(1-2)}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}]$ *the mean hitting time to reach* $\overline{\mathbf{x}}$ *starting from* $\mathbf{x}_0$ *and referring to the mutation model of class switch of 1 or 2 length strings. Then, for* $N \gg 1$ *we have:*

$$\mathbb{E}^{(1-2)}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}] \sim \frac{1}{2}\mathbb{E}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}] \quad \textit{with} \quad \mathbb{E}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}] \sim 2^N,$$

where $\mathbb{E}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}]$ *is the hitting time from* $\mathbf{x}_0$ *to* $\overline{\mathbf{x}}$ *according to the basic model, as defined in Section 2.2.3.*

Proof. First of all we observe that the last statement is a direct consequence of Proposition 2.2.8. As far as the first statement is concerned, we observe that according to the model we are analyzing here and due to Proposition 2.3.10, for $N \gg 1$ the order of magnitude of $\mathbb{E}^{(1-2)}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}]$ is:

$$\mathbb{E}^{(1-2)}_{\mathbf{x}_0}[\tau_{\{\overline{\mathbf{x}}\}}] \sim \frac{1}{2}\left(\frac{2^{N-1}}{N} + 2^{N-1}\right) + \frac{1}{2}2^{N-1}$$

where the first term corresponds to the case $\mathbf{x}_0 \notin \overline{\mathcal{H}_N}^{(1-2)}$ and the second one corresponds to the opposite case (as we choose randomly the first vertex, $\mathbf{x}_0$,

we have probability 1/2 that it belongs to each part of the hypercube). For the last term we used again Proposition 2.2.8 applied to a $(N-1)$-dimensional hypercube and according to the $(D_n^{(2)})$ process and the corresponding hitting time $H_{N-1}^{(2)}(d)$. The result follows. □

Table 2.3: Average expected times from $[0,\ldots,0]$ to $[1,\ldots,1]$, comparing the basic mutational model and the model of class switch of 1 or 2 length strings. Here we denote by $\widehat{\tau_{\{\overline{\mathbf{x}}\}}}_n$ the average value obtained over n simulations and by $\widehat{\sigma}_n$ its corresponding estimated standard deviation.

Mutational model	N	n	$\widehat{\tau_{\{\overline{\mathbf{x}}\}}}_n$	$\frac{\widehat{\sigma}_n}{\sqrt{n}}$
Basic	10	5000	1188.7996	16.2930
	11	5000	2312.5648	32.1073
Switch 1-2	10	5000	602.8124	8.4773
	11	5000	1181.5174	16.9023

Remark 16. We simulated the basic mutational model and the model of class switch of 1 or 2 length strings in order to compare the hitting times from $\mathbf{x}_0 := [0,\ldots,0]$ to $\overline{\mathbf{x}} := [1,\ldots,1]$ for both mutational models. We consider the case $N = 10$ and $N = 11$ in order to have an example in which the process starts from $\overline{\mathcal{H}_N}^{(1-2)}$ and from $\mathcal{H}_N^{(1-2)}$ respectively. Indeed, if $N = 10$ the process starts from the connected component to which $\overline{\mathbf{x}}$ belongs, while when $N = 11$ we have to reach one of the N nodes having distance 1 from $\overline{\mathbf{x}}$ to reach the connected component containing $\overline{\mathbf{x}}$. The average resulting hitting times are summarized in Table 2.3.

Allowing 1 to k mutations.

In this section we study the mean hitting time to cover a fixed Hamming distance d. First of all, we give the expression of the hitting time from node i to node j using the spectra. This formula is deduced by the more general one given in [85], in the case of regular graphs (the graph obtained by a convex combination of matrices $\mathcal{P}^i$ is a regular multigraph). We refer to the notations given in Section 2.2 for the eigenvectors of matrix $\mathcal{P}$: $\mathbf{v}_s = (v_{s1},\ldots,v_{s2^N})$ is the normalized eigenvector of $\mathcal{P}$ corresponding to λ_s. These eigenvectors are the columns of matrix Q_N (Section 2.2.1), and each component v_{si} corresponds to node i, as they were organized while constructing the adjacency matrix. Denoting by $T(i,j)$ the hitting time from node i to node j in $\mathcal{H}_N^{(k)}$, we obtain the following

expression:

$$T(i,j) = 2^N \sum_{l=2}^{2^N} \frac{1}{1-\lambda_l^{(k)}} (v_{lj}^2 - v_{li}v_{lj}),$$

which can be written using column vectors of $\mathcal{Z}_N$.

$$T(i,j) = \sum_{l=2}^{2^N} \frac{1}{1-\lambda_l^{(k)}} (z_{lj}^2 - z_{li}z_{lj})$$

We are interested in studying the equation below:

$$\overline{T}_N^{(k)}(d) := \frac{1}{2^N C_N^d} \sum_{h(i,j)=d} T(i,j) = \frac{1}{2^N C_N^d} \sum_{l=2}^{2^N} \frac{1}{1-\lambda_l^{(k)}} \sum_{h(i,j)=d} (z_{lj}^2 - z_{li}z_{lj}), \tag{2.28}$$

where $2^N C_N^d$ corresponds to the number of couples of nodes of $\{0,1\}^N$ having Hamming distance d.

First of all we can observe that for all l and for all j, $z_{lj}^2 = 1$. Moreover, in order to simplify notations, we denote $\mu_l^{(k)} := (1-\lambda_l^{(k)})^{-1}$. Also, we denote $R_N(l,d) := \sum_{h(i,j)=d} z_{li}z_{lj}$. Finally we obtain:

Proposition 2.3.12.

$$\overline{T}_N^{(k)}(d) = \sum_{l=2}^{2^N} \mu_l^{(k)} - \frac{1}{2^N C_N^d} \sum_{l=2}^{2^N} \mu_l^{(k)} R_N(l,d) \tag{2.29}$$

All the elements of this equation are known, except $R_N(l,d)$. Let us consider the $2^N \times (N+1)$ matrix $\mathcal{R}_N = (R_N(l,d))$, with $1 \leq l \leq 2^N$ and $0 \leq d \leq N$. One can prove by iteration:

Proposition 2.3.13.

$$\mathcal{R}_N = \mathcal{Z}_N \cdot \mathcal{L}_N \tag{2.30}$$

where $\mathcal{Z}_N := (\mathbf{z}_1, \ldots, \mathbf{z}_{2^N})$ *is recursively obtained from* $\mathcal{Z}_{N-1}$ *(Section 2.2.1), and*

$$\begin{cases} \mathcal{L}_1 = 2\mathcal{I}_2, \ \mathcal{I}_n \text{ being the } n\text{-dimensional identity matrix} \\ \\ \mathcal{L}_N = \begin{pmatrix} 2 \cdot \mathcal{L}_{N-1} & \mathbf{0}_{2^{N-1}} \\ \mathbf{0}_{2^{N-1}} & 2 \cdot \mathcal{L}_{N-1} \end{pmatrix}, \ \mathbf{0}_n \text{ being the } n\text{-length zero column vector} \end{cases}$$

Numerical simulations

In Figure 2.3 we plot some examples of the dependence of $\overline{T}_N^{(k)}(d)$ on d and k for different values of N.

Figure 2.3 (a) shows that for increasing k, $\overline{T}_N^{(k)}(d)$ varies on a smaller interval: $[1023, 1186.5]$ for $k = 1$, $[1028.1, 1068.6]$ for $k = 5$ and $[1025.6, 1044.8]$ for $k = 10$. It is intuitive to understand this fact: the hitting time depends less from the initial Hamming distance if we allow more mutations at the same mutational step. Indeed, we can actually visit more distant nodes since the first steps, so the initial Hamming distance has a smaller influence on the result. Figures 2.3 (b) and 2.3 (c) show the dependence of $\overline{T}_N^{(k)}(d)$ on k. We obtain the best result for the biggest k, except in the case $d = 1$ (as already shown by Figure 2.3 (a)). Curves corresponding to the case $d = 5$ and $d = 10$ are really close: we can evaluate their minimal and maximal values, which are respectively 1043.25 and 1177.60 for $d = 5$; 1044.82 and 1186.54 for $d = 10$. This fact highlights once again that if $d > 1$, the initial Hamming distance poorly influences the value of the hitting time. The case $d = 1$ shows surprisingly that the hitting time is not necessarily a monotone function of k. Figure 2.3 (c) allows us to focus to this behavior and better understand its causes. Indeed, as N is quite small, this figure shows more clearly the oscillating behavior of $\overline{T}_N^{(k)}(d)$ while studying its dependence on k: for even values of k, $\overline{T}_5^{(k)}(1)$ increases, while for odd values of k it decreases. Intuitively, as the distance we want to cover is $d = 1$, if we allow to do 2 mutations instead of simply one, then we have a high probability to go further since the beginning of the process. Let us now look to Equation (2.28) and, in particular to the factor: $\sum_{l=2}^{2^N}(1 - \lambda_l^{(k)})^{-1}$. We can understand the phenomenon plotted in Figure 2.3 (c) by looking at Proposition 2.3.8. If k is odd and little enough then the last eigenvalue, which is negative (equal to $-1/k$), has an important negative influence over the value of $\overline{T}_N^{(k)}(d)$. Clearly, this fact has a substantial effect only if N and k are little enough, otherwise it will be compensated by the effect of all other eigenvalues.

One may wonder what would be the best choice for the coefficients a_i (Definition 2.12), $1 \leq i \leq k$, so that $\overline{T}_N^{(k)}(d)$ is minimized for a fixed k. We have to minimize the convex combination $\sum_{i=1}^{k} a_i \lambda_l^i$. The answer is quite evident: if $k > 2$ the minimum is obtained by taking all $a_i = 0$ and $a_{k^*} = 1$, where $k^* = 2\lfloor (k+1)/2 \rfloor - 1$. Consequently, the best choice for the transition probability matrix is $\mathcal{P}^{k^*}$. The fact that we need to consider the greater odd component has also another explanation, which is more intuitive. Indeed if we consider the RW given by $\mathcal{P}^{2t}$, we will be trapped in one of the connected-components of

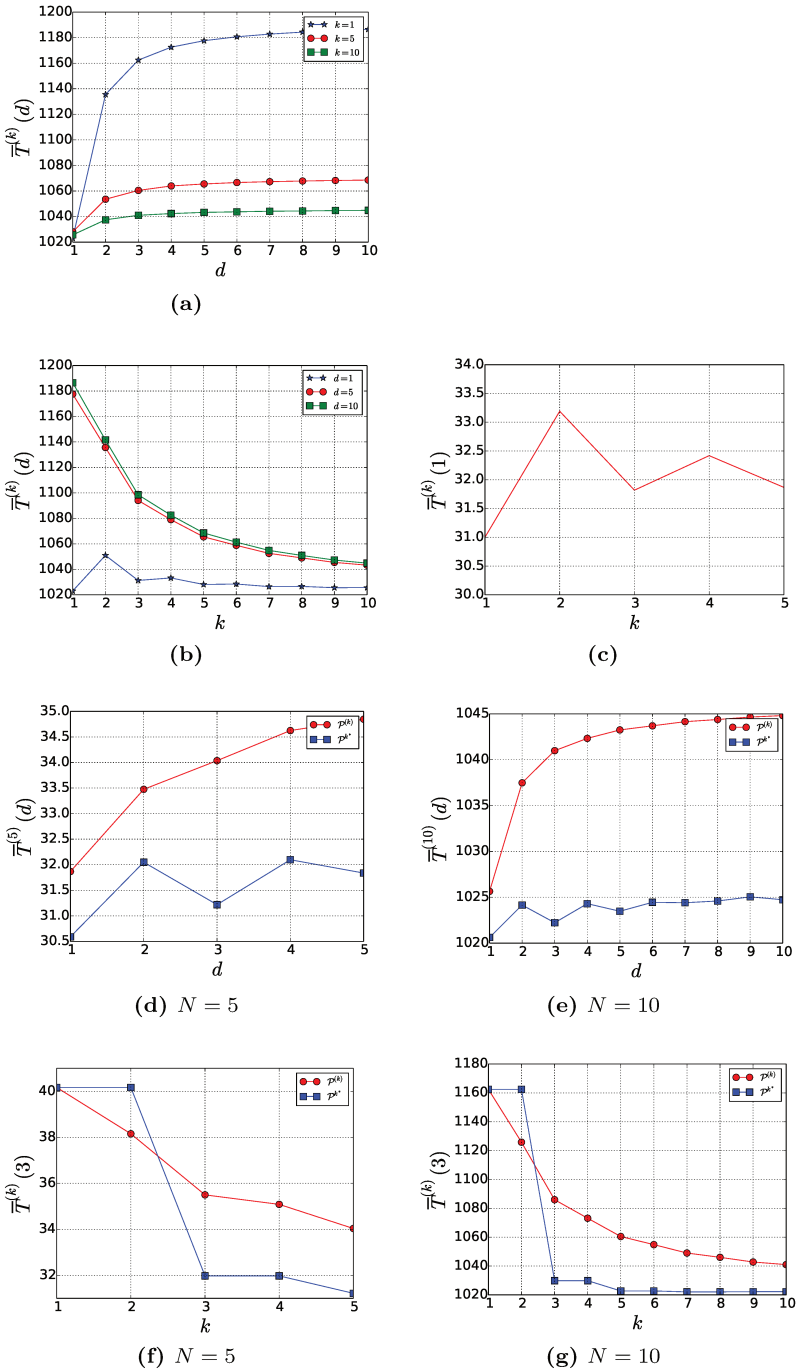


Figure 2.3: (a) Dependence of $\overline{T}_N^{(k)}(d)$ on d for $N = 10$ and $k = 1$, 5 or 10. (b) Dependence of $\overline{T}_N^{(k)}(d)$ on k for $N = 10$ and different values of d. (c) Dependence of $\overline{T}_5^{(k)}(1)$ on k. (d, e) Dependence of $\overline{T}_N^{(k)}(d)$ on d for different values of both N and k. Values obtained by using as transition probability matrices $\mathcal{P}^{(k)}$ and $\mathcal{P}^{k^*}$ respectively are compared. (f, g) Dependence of $\overline{T}_N^{(k)}(d)$ on k for different values of both N and d. Again, cases corresponding to $\mathcal{P}^{(k)}$ and $\mathcal{P}^{k^*}$ are compared.

the graph, due to the bipartite structure of the hypercube. Indeed, the graph underlined by $\mathcal{P}^{2t}$ is non-connected $\forall t > 0$. Therefore, we will not be able to reach those nodes having a different parity of 1s in their string, referring to $\mathbf{X}_0$.

In Figures 2.3 (d), 2.3 (e), 2.3 (f) and 2.3 (g) we plotted together the values of hitting times to cover a Hamming distance d for different values of N, k, and d, comparing the process given by $\mathcal{P}^{(k)}$ and the one corresponding to $\mathcal{P}^{k^*}$. This gives more evidence of the fact that the second one is the optimal one. It is interesting to look at the case in which d is fixed and we let k vary. For $k = 1$ both processes gave the same result as $\mathcal{P}^{1^*} = \mathcal{P} = \mathcal{P}^{(1)}$. Moreover, for $k = 2$ the process $\mathcal{P}^{(2)}$ is clearly the faster one: we recall that defining $\mathcal{P}^{k^*}$ we consider the greater odd k, and then $\mathcal{P}^{2^*} = \mathcal{P}$, while the process $\mathcal{P}^{(2)}$ allows to do 1 or 2 mutations at each mutation step. Then $\mathcal{P}^{k^*}$ is actually the best choice among all possible convex combinations of $\mathcal{P}^i$ iff $k > 2$. In Figures 2.3 (d) and 2.3 (e) we observe the oscillating behavior of $\overline{T}_N^{k^*}(d)$. That depends on the structure of $\mathcal{R}_N$, considering that $\sum_{l=2}^{2^N-1} R_N(l,d) = 0$ for d odd and $\sum_{l=2}^{2^N-1} R_N(l,d) = -2(2^N C_N^d)$ for d even. One can get convinced of this fact by explicitly computing $\overline{T}_N^{k^*}(d)$ for $N = 3$. Moreover simulations show that this behavior is softened for increasing d, and that $\overline{T}_N^{k^*}(N-1) > \overline{T}_N^{k^*}(N)$. This fact is confirmed by simulations on the real process. Finally, Figures 2.3 (f) and 2.3 (g) clearly show that for $k = 2$ the process given by $\mathcal{P}^{(k)}$ allows to cover quickly a fixed Hamming distance. As expected, the best hitting time is obtained for $k = N$, and for increasing N and k the value of this hitting time has a smaller variation.

Table 2.4: An example of comparison between the theoretical and experimental values of $\overline{T}_5^{(5)}(4)$ for $\mathcal{P}^{(5)}$. $\widehat{\overline{T}_5^{(5)}(4)}_n$ denotes the average value obtained over n simulations and $\widehat{\sigma}_n$ its corresponding estimated standard deviation.

Transition probability matrix	N	d	k	n	$\overline{T}_5^{(5)}(4)$	$\widehat{\overline{T}_5^{(5)}(4)}_n$	$\frac{\widehat{\sigma}_n}{\sqrt{n}}$
$\mathcal{P}^{(k)}$	5	4	5	480000	34.62	34.67	0.05

We can test all these observations by simulating the real process for both transition probability matrices, $\mathcal{P}^{k^*}$ and $\mathcal{P}^{(k)}$. Results obtained are consistent with our theoretical analysis. In order to give an idea of experimental values obtained by testing the process, in Table 2.4 we compare the theoretical value of $\overline{T}_N^{(k)}(d)$ corresponding to $\mathcal{P}^{(k)}$, and the experimental value with its precision, for $N = 5$, $k = 5$ and $d = 4$.

2.4 Modeling issues

The mathematical framework described in previous sections can be used to model mutations characteristic of SHM. In Sections 2.4.1 and 2.4.2 we give some more details about GCs and the binding between B-cells and antigens. Therefore, in Section 2.4.3 we set the modeling assumptions which justify to mathematically describe SHMs as RWs on binary strings. Of course, this is a not exhaustive approximation. Hence, some limitations are discussed in Section 2.4.4 and some propositions for further developments are given as well.

2.4.1 The germinal center reaction

Antigen-activated B-cells, together with their associated T cells, move into a primary lymphoid follicle, where they proliferate and ultimately form a GC. GCs are composed mainly of B-cells, but antigen specific T-cells, which have also been activated and migrated to the lymphoid follicle, make up about 10% of GC lymphocytes and provide indispensable help to B-cells [110, 124, 102]. Indeed, when B-cells start to proliferate in GC, they need to receive proper survival signals, or they die by apoptosis. The number of B-cells within a germinal center grows at high pace: it can double every 6-8 hours [55, 36]. After about 3 days of strong proliferation, B-cells start undergoing SHM, in order to diversify the variable region of their BCRs, and those cells that express newly generated BCRs are selected for enhanced antigen binding. The fast proliferation rate of B-cells is required for the generation of a large number of modified BCRs within a short frame time (one cell gives 10^4 blasts in 72 hours). Some B-cells positively selected in the light zone differentiate into memory B-cells or plasma cells. The GC reaches its maximal size within approximately two weeks, after which the structure slowly involutes and disappears within several weeks [136]. During the GC process B-cells are subjected to powerful selection mechanisms that facilitate the generation of high affinity antibodies: a B-cell that express a newly generated BCR needs to be tested for enhanced antigen binding. This process is mediated by FDCs and follicular helper T-cells. BCR stimulation through antigen binding coupled with co-stimulatory signals transmitted by GC T-cells, provides survival signals to the cell. By contrast, failure of the BCR to bind antigen and receive proper rescue signals causes cell death by apoptosis [36]. The final differentiation of a GC B-cell into a plasma cell or a long-lived memory B-cell is driven by the acquisition of a high-affinity BCR. For short-lived memory B-cells, the differentiation process seems to be stochastic, as throughout GC reaction B-cells are constantly selected to enter the memory pool [102, 126].

2.4.2 B-cell receptors and antigen-antibody binding

Immunoglobulins (Ig) present at the antigen receptor are Y-shaped macro proteins composed of four polypeptide chains assembled by disulfide bonds: two identical heavy (H) chains and two identical light (L) chains. Each chain consists of two regions: a constant (C) region, which has an effector function, and a variable (V) region composed by the variable parts of the two chains together. During GC reaction the only one involved in SHMs is the V region, which also determines the antigen binding site ([102], Chapter 1). We call *antigen binding site* or *paratope* the specialized portion of the BCR V region used for identifying other molecules, while the regions on any molecule that paratopes can recognize are called *epitopes*. B-cells are able to bind ligands whose surfaces are 'complementary' to that of their antigen binding site, where complementarity means that the amino-acids composing the paratope and the epitope are distributed in such a way to form bonds which hold the antigen to the B-cell. In this case these bonds are all non-covalent (as hydrogen bonds, electrostatic bonds, van der Waals forces and hydrophobic bonds), which are by their nature reversible. Multiple bonding between the antigen and the B-cell ensures that the antigen is bound tightly to the B-cell. The interaction between paratope and epitope can be characterized in terms of a binding affinity, proportional to their complementarity. The *affinity* is the strength of the reaction between a single antigenic determinant and a single combining site on the B-cell: it summarizes the attractive and repulsive forces operating between the antigenic determinant and the combining site of the B-cell, and corresponds to the equilibrium constant that describes the antigen-B-cell reaction [1, 141, 80].

Each antigen typically has several epitopes, so that the surface of an antigen presents variable motifs that B-cells, through their receptors, can discriminate as distinct epitopes. If we define an epitope by its spatial contact with a BCR during binding, the number of relevant amino-acids is approximately 15, and among these amino-acids only around 5 in each epitope strongly influence the binding. These strong sites may contribute about one-half of the total free energy of the reaction, while the other amino-acids influence in binding constant by up to one order of magnitude or even have no detectable effect. Simultaneously, a BCR contains a variety of possible binding sites and each antibody binding site defines a paratope: about 50 variable amino-acids make up the potential binding area of a BCR. In agreement with the above, only around 15 among these 50 amino-acids physically contact a particular epitope: these define the structural paratope. Consequently, antibodies have a large number of potential paratopes as the 50 or so variable amino-acids composing the binding

region define many putative groups of 15 amino-acids [80].

Substitutions both in and away from the binding site can change the spatial conformation of the binding region and affect the binding reaction. The consequence of mutation at a particular site depends on the original amino-acid and the amino-acid used for substitution ([1], Chapter 4).

2.4.3 From DNA to amino-acids: choosing the best viewpoint

Mutations observed on the binding site of B-cells during the GC process are the result of genetic mutations produced by SHM on the portion of DNA encoding for the BCR V region. In the current section we discuss a model of genetic mutations and its effects on the amino-acid string, under the assumption of having two amino-acid classes. We show that the framework we set up in previous sections can adapt to model the effects of SHM over BCRs and study the variation of the affinity with the presented antigen.

The genetic code is a sequence of four nucleotides, guanine (G), adenine (A) (called purines), thymine (T) and cytosine (C) (pyrimidines), joined together. They make three-letter words: the codons. Each codon corresponds to a specific amino-acid or to a stop signal, which interrupts the building of the protein during translation. As the number of possible combinations of 4 nucleotides in 3-length words is 64, and there exists 20 amino-acids in naturally derived proteins, more than a single codon codes for the same amino-acid [125]. Table 2.5 shows the correspondence between codons and amino-acids.

Different kind of genetic mutations can affect the DNA sequence of a gene. They can be regrouped in three main categories: base substitutions, insertions and deletions. A single base substitution is a switch of a nucleotide with another. This is the simplest kind of mutation and it can turn out to be missense, nonsense or silent, once we observe the resulting new protein. We said that a mutation is missense if the result of the genetic mutation is a different amino-acid in the protein. The mutation is nonsense when the genetic mutation results in a stop codon instead of an amino-acid. Finally, a silent mutation is a mutation with no effect on the amino-acid string, *i.e.* the mutated sequence codes for an amino-acid with identical binding properties. We talk about insertion (resp. deletion) when one or more nucleotides are added (resp. removed) at some place in the DNA code. These last kinds of mutations can both be frameshift mutations, which are given by the insertion or deletion of a number of bases that

Table 2.5: The correlation between codons and amino-acids: most of the amino-acids derives from more than a single codon.

	T		**C**		**A**		**G**		
T	TTT	Phe (F)	TCT	Ser (S)	TAT	Tyr (Y)	TGT	Cys (C)	**T**
	TTC	Phe (F)	TCC	Ser (S)	TAC	Tyr (Y)	TGC	Cys (C)	**C**
	TTA	Leu (L)	TCA	Ser (S)	TAA	Stop	TGA	Stop	**A**
	TTG	Leu (L)	TCG	Ser (S)	TAG	Stop	TGG	Trp (W)	**G**
C	CTT	Leu (L)	CCT	Pro (P)	CAT	His (H)	CGT	Arg (R)	**T**
	CTC	Leu (L)	CCC	Pro (P)	CAC	His (H)	CGC	Arg (R)	**C**
	CTA	Leu (L)	CCA	Pro (P)	CAA	Gln (Q)	CGA	Arg (R)	**A**
	CTG	Leu (L)	CCG	Pro (P)	CAG	Gln (Q)	CGG	Arg (R)	**G**
A	ATT	Ile (I)	ACT	Thr (T)	AAT	Asn (N)	AGT	Ser (S)	**T**
	ATC	Ile (I)	ACC	Thr (T)	AAC	Asn (N)	AGC	Ser (S)	**C**
	ATA	Ile (I)	ACA	Thr (T)	AAA	Lys (K)	AGA	Arg (R)	**A**
	ATG	Met (M)	ACG	Thr (T)	AAG	Lys (K)	AGG	Arg (R)	**G**
G	GTT	Val (V)	GCT	Ala (A)	GAT	Asp (D)	GGT	Gly (G)	**T**
	GTC	Val (V)	GCC	Ala (A)	GAC	Asp (D)	GGC	Gly (G)	**C**
	GTA	Val (V)	GCA	Ala (A)	GAA	Glu (E)	GGA	Gly (G)	**A**
	GTG	Val (V)	GCG	Ala (A)	GAG	Glu (E)	GGG	Gly (G)	**G**

is not a multiple of 3, altering the reading frame of the gene. SHM introduces mostly single nucleotide exchanges, together with small deletions and duplications, *i.e.* the insertion of extra copies of a portion of genetic material already present within the DNA code [63, 26, 27]. Among these point mutations, transitions (*i.e.* substitution of a purine nucleotide with another purine one, or a pyrimidine with a pyrimidine) dominate over transversions (substitution of a purine with a pyrimidine or conversely). About half of the mutations (53%) have been estimated to be silent, about 28% nonsense, and only about 19% of all mutations have been estimated to be missense and then have an effect on affinity, which can either be of an improving nature, or of worsening and even lead to the formation of autoreactive clones [64].

The 20 existing amino-acids are typically classified in charged amino-acids, polar (non-charged) amino-acids and hydrophobic amino-acids, depending on their chemical characteristics. As we have already discussed in Section 2.4.2 the bonding between BCR and antigen is made thanks to non-covalent bonds, in

particular ionic bonds and hydrogen bonds. Ionic bonds are the result of interactions between two amino-acids oppositely charged: arginine (R) and lysine (K) are positively charged, while aspartic acid (D) and glutamic acid (E) are negatively charged. As long as hydrogen bonds are concerned, also polar amino-acids can participate. In particular arginine (R), lysine (K) and tryptophan (W) have hydrogen donor atoms in their side chains; aspartic acid (D) and glutamic acid (E) have hydrogen acceptor atoms in their side chain while asparagine (N), glutamine (Q), histidine (H), serine (S), threonine (T) and tyrosine (Y) have both hydrogen donor and acceptor atoms in their side chains.

Stop codons also have an important role. Indeed, during translation (the last step necessary to build a protein starting from the DNA molecule) amino-acids continue to be added until a stop codon is reached. There exists two types of mutations involving stop codons, named nonsense and nonstop respectively. The first one corresponds to the substitution of an amino-acid with a stop codon, while the second one is the opposite case. In both cases the resulting protein has an abnormal length, which often causes a loss of function. Moreover, errors given by both nonsense and nonstop mutations are linked to over 10% of human genetic diseases [24].

Concerning mutation in activated B-cells, SHM is driven by an enzyme called activation-induced cytidine deaminase (AID) which is expressed specifically in this case. This protein can bind to single-stranded DNA only. Thus it seems to target only genes being transcribed (for which the transcription phenomenon separates temporarily double stranded DNA into small portions of two single stranded DNA sequences) [71]. AID converts Cytosine (C) in Uracil (U) by deamination. This substitution occurs at higher rates in hot spots motives like $D\underline{G}YW/WR\underline{C}H$ where ($G : C$ is the mutable position and $D \in \{A, G, T\}$, $H \in \{A, C, T\}$, $R \in \{A, G\}$, $W \in \{A, T\}$ and $Y \in \{C, T\}$, and the underlined letters are the loci of mutations) [112, 63]. Then, two mechanisms tend to repair lesions in the DNA caused by these substitutions of C by U [115]:

a) either *mismatch repair:* substitution for the damaged zone by another sequence of nucleotides thanks to proteins MSH 2/6. The U base is read as T leading to a transition from a $C : G$ pair to $T : A$.

b) or *base excision repair:* U is excised by a successive action of uracil-DNA glycolase (UNG) and apurinic/apyrimidinic endonuclease (APE1). The DNA contains then a nick, after replication, a random nucleotide is inserted in order to fill the vacant space leading to transversions and transitions.

From a mathematical point of view this is equivalent to define the switch with a

random nucleotide depending on the motives present in the chain. The probability concerning the choice of this nucleotide to be inserted shall not be uniform due to the presence of mismatch and excision repairs [37, 115]. This is not taken into account in the model we developed.

We can therefore make the following three main assumptions to model the SHM process acting on the BCR V region:

Modeling assumption 1. SHM introduces only single point mutations in the DNA strand, missense or silent. Therefore we do not take into account nonsense mutations, in order to avoid an interruption of the mutation process due to the introduction of a stop codon. The choice of the base used for substitution is made randomly, without considering that we have mostly $A \leftrightarrow T$ and $G \leftrightarrow C$ substitutions.

Modeling assumption 2. We consider only electrostatic and hydrogen bonds as responsible for the bonding between BCR and antigen. We suppose we have two amino-acid classes represented as 0 and 1 respectively: we denote by 1 those amino-acids which have hydrogen donor atoms in their side chains (or which are positively charged) and by 0 those amino-acids which have hydrogen acceptor atoms in their side chains (or which are negatively charged). We arbitrary chose to assign 0 or 1 to amino-acids which can act as an acid or a base in hydrogen bonds. As an exemple, as serine can form hydrogen bonds with arginine and threonine, one can assign 0 to serine and 1 to threonine (arginine is represented by 1 as it is positively charged). While translating the amino-acid chain into a binary chain, we omit all hydrophobic amino-acids, as they do not participate in electrostatic or hydrogen bonds. Their position corresponds to an empty case, which does not contribute to the affinity between B-cell and antigen. This is clearly an important simplification. We will further discuss this choice in Section 2.4.4.

Modeling assumption 3. We consider a linear contact between two amino-acid strings, without taking into account the geometrical configuration of both the BCR and the antigen.

The process starts from a DNA chain coding for a BCR, $\mathbf{X}_0^{dna}$; from which we can obtain the corresponding amino-acid chain, $\mathbf{X}_0^{aa}$ (Table 2.5) and, consequently, its binary expression, $\mathbf{X}_0^{bin}$.

Example 1.

- $\mathbf{X}_0^{dna}$ = (GTT, GAG, CTA, GTG, GAA, AGT, GGA, GCC, GAA, GTA, AAA, AAG, CCA, GGT, AGT, AGT, GTT, AAA, GTC, AGT, TGT, AAA, GCA)

- $\mathbf{X}_0^{aa} = (\text{V, Q, L, V, E, S, G, A, E, V, K, K, P, G, S, S, V, K, V, S, C, K, A})$

- $\mathbf{X}_0^{bin} = (-, 1, -, -, 0, 0, -, -, 0, -, 1, 1, -, -, 0, 0, -, 1, -, 0, 0, 1, -)$

Notation 1. Given a vector $\mathbf{X}$, we denote by $|\mathbf{X}|$ its length (counting also the empty cases, if there are some). Equivalently, given a set $\mathcal{S}$, we denote by $|\mathcal{S}|$ its size

We can formalize the translation of the nucleotides chain into the amino-acids chain as follows.

Definition 2.15. Let $\mathcal{N}$ and $\mathcal{A}$ be two sets of letters with size respectively $|\mathcal{N}| = k_1$ and $|\mathcal{A}| = k_2$. Let l be an integer positive number so that $k_1^l \geq k_2$. Then we define $f_{k_1,k_2,l} : \mathcal{N}^l \to \mathcal{A}$, which associate at least an l-length sequence of letters belonging to $\mathcal{N}$ to a letter in $\mathcal{A}$.

In our specific case, following definition 2.15, $\overline{\mathcal{N}} := \{\text{G, A, T, C}\}$ is the set of nucleotides, while $\overline{\mathcal{A}}$ is the set containing all possible amino-acids, together with the stop signal. Therefore $\overline{k_1} = 4$ and $\overline{k_2} = 21$. Moreover we know that $\bar{l} = 3$ and the function $\overline{f}_{4,21,3}$ is detailed in Table 2.5.

Remark 17. We can easily observe that $\bar{l} = \min\left\{n \in \mathbb{N} \,|\, \overline{k_1}^n \geq \overline{k_2}\right\}$. Indeed, having 4 nucleotides available to build a DNA strand, we need to read them at least by 3-length blocks in order to be able to synthesize all 20 amino-acids. Moreover, choosing this value for the parameter l avoids to have too many sequences of nucleotides coding for the same amino-acid.

At the beginning of the process, the antigen string in its three representations is given as well: $\overline{\mathbf{x}}^{dna}$, $\overline{\mathbf{x}}^{aa}$ and $\overline{\mathbf{x}}^{bin}$, with $|\mathbf{X}^{dna}| = |\overline{\mathbf{x}}^{dna}| =: 3N$. Antigen representing strings remain unchanged. Assumptions 1-3 imply that for all $t \geq 0$, $|\mathbf{X}_t^{bin}| = |\overline{\mathbf{x}}^{bin}| = N$. At each time step a single point mutation (missense or silent) is introduced in the DNA chain coding for the BCR. So, if $\mathbf{X}_t^{dna}$ is the DNA code at time t, we randomly choose an index $i \in \{1, \ldots, 3N\}$, a letter $a \in \overline{\mathcal{N}}$ and we place $(X_{t+1}^{dna})_i := a$. If the new codon is a stop codon, then we choose $a' \in \overline{\mathcal{N}} \setminus \{a\}$ and we put $(X_{t+1}^{dna})_i := a'$, and so on.
In order to test the affinity, we consider the binary expression of both the BCR and the antigen, which we take in its complementary form, *i.e.* $\overline{\mathbf{x}}'^{bin} := (1 - \overline{x}_1^{bin}, \ldots, 1 - \overline{x}_N^{bin})$. This leads us back to the definition of affinity we made in Section 2.2: 0 matches with 0 and 1 with 1.
As we consider a linear contact between $\mathbf{X}_t^{bin}$ and $\overline{\mathbf{x}}'^{bin}$, at the positions where either $\mathbf{X}_t^{bin}$ or $\overline{\mathbf{x}}'^{bin}$ has an hydrophobic amino-acid, we suppose that no match

is possible. Therefore we can extend Definition 2.4 of the Hamming distance in a very natural way to this more general case:

Definition 2.16. We denote by $Hy(\mathbf{X}_t^{bin})$ (resp. $Hy(\overline{\mathbf{x}}'^{bin})$) the set of the indices corresponding to hydrophobic amino-acids in $\mathbf{X}_t^{bin}$ (resp. in $\overline{\mathbf{x}}'^{bin}$). Therefore the Hamming distance between $\mathbf{X}_t^{bin}$ and $\overline{\mathbf{x}}'^{bin}$ is given by:

$$h(\mathbf{X}_t^{bin}, \overline{\mathbf{x}}'^{bin}) = \sum_{\substack{i\in\{1,\dots,N\} \\ i\notin Hy(\mathbf{X}_t^{bin})\cup Hy(\overline{\mathbf{x}}'^{bin})}} \delta_i + |Hy(\mathbf{X}_t^{bin}) \cup Hy(\overline{\mathbf{x}}'^{bin})|$$

$$\text{where } \delta_i = \begin{cases} 1 & \text{if} \quad (X_t^{bin})_i \neq (\overline{x}'^{bin})_i \\ 0 & \text{otherwise} \end{cases}$$

Then, for all $t \geq 0$:

$$|Hy(\mathbf{X}_t^{bin}) \cup Hy(\overline{\mathbf{x}}'^{bin})| \leq h\left(\mathbf{X}_t^{bin}, \overline{\mathbf{x}}'^{bin}\right) \leq N$$

We consider that the optimal clone is reached when:

$$\text{aff}\left(\mathbf{X}_t^{bin}, \overline{\mathbf{x}}'^{bin}\right) := N - |Hy(\overline{\mathbf{x}}'^{bin})|$$

The effects of nucleotides exchanges on the binary expression of BCRs can be multiple:

No detectable effect : this is the result of either a silent mutation or a missense mutation which substitutes an amino-acid with another one belonging to the same amino-acid class.

Class-switch , derived from a missense mutation which leads to the substitution of an amino-acid with another one belonging to the other amino-acid class.

We can further complexify this model by replacing Assumption 1 with the following one:

Modeling assumption 4. SHM introduces mostly single point mutations in the DNA, missense or silent. With weak probability, deletions or insertions can occur. For the sake of simplicity, we suppose that a deletion (resp. an insertion) consist in the elimination (resp. the addition) of a non-stop codon. Moreover, in order to avoid the problem of a variation in the length of the BCR representing string, when a deletion occur, those bits situated on the right of the deleted one shift to the left, and a random extra codon is added at the right bottom. Conversely, if an insertion occurs, the right bottom bit is deleted.

Even if these mutational events are rare, they have remarkable effects over the structure of the underlying graph. Indeed a deletion or an insertion entails a great jump in the affinity function by producing a shift of a portion of the BCR representing string. This is not the case if we consider only single point mutations. Therefore, under Assumption 4 the graph we obtain is much more complex and allows random long range connections.

Numerical simulations

In order to evaluate how deletions and insertions affect the mean number of mutation steps to reach the desired B-cell trait, we make some numerical simulations. We compare a model in which only single point mutations are allowed to another one in which also deletions and insertions can occur. We refer to Assumption 4 to define these mutational events.

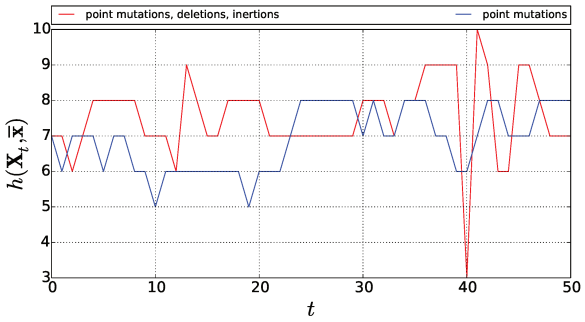


Figure 2.4: Variation of the Hamming distance to $\overline{\mathbf{x}}'^{bin}$, comparing the model of single point mutations to the one which includes also deletions and insertions (50% of all mutation events). In both cases $N = 10$. Deletions and insertions lead to a quick change in the Hamming distance. Between time 30 and 50, we can observe the effect of indels mutations.

Figure 2.4 shows the effects of deletions and insertions over the affinity. In order to do these simulations, we arbitrary fixe a BCR and an antigen with given affinity. We do not consider those base substitutions leading to no detectable effect, *i.e.* at each time step we can observe a variation of the affinity function. In Figure 2.4 we can clearly locate at what time an insertion or a deletion has occurred, because this coincides with a jump of the Hamming distance between BCR and antigen.

One can ask how these random long range connections affect the average time to reach the antigen target string. Simulations show that one needs a more long time to reach $\overline{\mathbf{x}}'^{bin}$ if the probability of making such mutations in-

creases. The results obtained through 10000 simulations are collected in Table 2.6.

Table 2.6: Average number of mutations needed to reach $\overline{\mathbf{x}}'^{bin}$, for $N = 10$ and starting from Hamming distance 7. In $\overline{\mathbf{x}}'^{bin}$, only 2 amino-acids are hydrophobic, so by Definition 2.16, the optimal affinity one can reach is 8. We compare three models: in the first one no deletions nor insertions are allowed. In the second model 10% of all mutations are deletions or insertions, 50% in the last one. We denote by $\widehat{\tau_{\{\overline{\mathbf{x}}'^{bin}\}_n}}$ the average value obtained over n simulations and by $\widehat{\sigma}_n$ its corresponding estimated standard deviation. Simulations show that $\widehat{\tau_{\{\overline{\mathbf{x}}'^{bin}\}_n}}$ increases when the pourcentage of deletions or insertions grows, and so does the corresponding variation.

% **deletions/insertions**	$\lvert\overline{\mathbf{x}}'^{\boldsymbol{bin}}\rvert$	$\boldsymbol{h(\mathbf{X}_0^{bin}, \overline{\mathbf{x}}'^{bin})}$	$\boldsymbol{n}$	$\widehat{\boldsymbol{\tau_{\{\overline{\mathbf{x}}'^{bin}\}_n}}}$	$\frac{\widehat{\sigma}_n}{\sqrt{n}}$
0	10	7	10000	8824.93	86.80
10	10	7	10000	9091.12	92.01
50	10	7	10000	10075.89	100.59

We can discuss which viewpoint is the most suitable to study mutations and their effects over the interactions between BCR and antigen. It is really hard to define a clear correspondence between genetic mutations and the evolution of the affinity, even while considering a simple linear contact between molecules (hence without observing the changes in the geometrical structure of the protein). Indeed, in order to test the affinity between BCR and antigen we constantly need to project the DNA string on the smaller state-space containing the binary representations of B-cell traits. If we directly consider mutations on binary strings, then the resulting process is faster, as we do not observe missense mutations, and the evaluation of the affinity is immediate.

The comprehension of the nature of genetic mutations and their consequences on the new generated protein, suggested us to make Assumptions 1-3 to formalize the model. In particular, we found reasonable to look directly to amino-acid chains and their binary representation: this allows to study the affinity between BCR and antigen using the Hamming distance. Therefore, under these hypotheses the general mathematical framework described in Section 2.2 can be applied to study how different kinds of missense mutations affect the dynamics of AAM. As we show in Sections 2.2-2.3, this already brings interesting and complexes mathematical problems.

2.4.4 Limitations and extensions

In this Chapter we propose and study mutational processes on N-length binary strings, which can be variously applied to evolutionary contexts. As far as the application to the SHM process is concerned, we can make some remarks about our assumptions, which can bring us to enrich and complexify the model through a more coherent representation of the true biological process.

First of all we have decided to consider only two amino-acid classes. From one side this assumption is justified as charged and polar amino-acids are effectively the most responsible in creating bonds which determine the antigen-antibody interaction. Therefore they strongly influence the affinity between BCR and antigen. Nevertheless, by making this simplification we omit all hydrophobic amino-acids from the string, and that is not without consequences. The elimination of hydrophobic amino-acids from the string significantly changes the structure of the chain, therefore the ability for charged and polar amino-acids to be in contact with each-others. Moreover, the effects of genetic mutations on the new generated protein could be even more complex than the ones we have considered in this Chapter. Finally, by taking into account also hydrophobic amino-acids, we would be able to consider hydrophobic bonds, which also influences the antigen-antibody interaction. Therefore it seems more appropriate to consider three, or more, amino-acids classes (*e.g.* [108, 101]).

As far as the nature of mutations is concerned, we have essentially described mutational processes given by combinations of single point mutation mechanisms. During SHM nucleotide exchanges are the most frequent among all possible mutations. Despite this, also some deletions and insertions occur. This has two main consequences. Firstly it means that the length of the BCR representing string could change during the process, while we consider it as fixed and equal to the length of the antigen. We can maybe overcome this problem by saying that the chain represented in our model corresponds to the portion of BCR in contact with the antigen, and this is almost fixed (Section 2.4.2). Moreover these mutations can imply substantial changes into the amino-acid chain, hence they can bring a great jump of the affinity to the presented antigen. Therefore, even if these are rare mutational events, they may have an important effect in AAM. Consequently it could be interesting to take also insertions and deletions into account. All these observations lead interesting mathematical questions.

Of course we can also envisage developments in other directions. For example by considering the creation of bonds among amino-acids of the BCR (resp.

the antigen) itself, which determines the geometrical structure of the protein and consequently the portion of the BCR and the antigen that can actually be in contact. Another interesting possibility is to consider that mutations at one site are influenced by other amino acids composing the string. This assumption was firstly proposed by S. A. Kauffman and E. D. Weinberger in [70], where they introduced the NK models. In this context the parameter K assures the richness of epistatic interactions among sites. More recently Y. Elhanati *et al* in [45] find biological evidence for an evolutionary model where substitution rates strictly depend on the context.

We propose some numerical simulations to evaluate the consequences over the hitting time of both the addiction of extra amino-acid classes and the possibility of having a BCR string longer than the antigen one.

A. S. Perelson and G. Weisbuch in [108] proposed a model with 3 amino-acid classes: hydrophobic, hydrophilic positively charged and hydrophilic negatively charged. Hydrophobic amino-acids match with hydrophobic and hydrophilic positively charged with hydrophilic negatively charged. We simulated the expected time to reach a given configuration comparing the model with 2 amino-acid classes and the one with 3 amino-acid classes, and considering single switch-type mutations. We take two random 10-length strings having maximal distance between each-others. We extend Definition 2.4 of Hamming distance to the state-space $\{0,1,2\}^N$ in a natural way, keeping the same notation: $\forall\ \mathbf{x} = (x_1, \dots, x_N)$, $\mathbf{y} = (y_1, \dots, y_N) \in \{0,1,2\}^N$, their Hamming distance is given by:

$$h(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N} \delta_i \qquad \text{where} \qquad \delta_i = \begin{cases} 1 & \text{if} \quad x_i \neq y_i \\ 0 & \text{otherwise} \end{cases} \tag{2.31}$$

Therefore the affinity is defined as in Definition 2.3. We simulated for both cases a single switch-type mutational model (Definition 2.5 for 2 amino-acid classes and Definition 2.17 below for 3 amino-acid classes), testing the time we need to reach the target vertex.

Definition 2.17. Let $\mathbf{X}_n \in \{0,1,2\}^N$ be the BCR at step n. Let $i \in \{1, \dots, N\}$ be a randomly chosen index, and $a \in \{0,1,2\} \setminus \{X_{n,i}\}$ a randomly chosen number. Then $\mathbf{X}_{n+1} := (X_{n,1}, \dots, X_{n,i-1}, a, X_{n,i+1}, \dots, X_{n,N})$.

Table 2.7 shows the results we obtained over 10000 simulations.

We already knew from theoretical analysis that the order of magnitude for the hitting time of the basic mutational model is 2^N for N big enough. Sim-

Table 2.7: Average expected times to cover a Hamming distance $h(\mathbf{X}_0, \overline{\mathbf{x}}) = 10 = N$, comparing the model with 2 amino-acid classes and the one with 3 amino-acid classes. Here we denote by $\widehat{\tau_{\{\overline{\mathbf{x}}\}}}_n$ the average value obtained over n simulations and by $\widehat{\sigma}_n$ its corresponding estimated standard deviation.

Amino-acid classes	N	$h(\mathbf{X}_0, \overline{\mathbf{x}})$	n	$\widehat{\tau_{\{\overline{\mathbf{x}}\}}}_n$	$\frac{\widehat{\sigma}_n}{\sqrt{n}}$
2	10	10	10000	1213.2108	12.0138
3	10	10	10000	62160.8263	635.0458

ulations clearly show that when we consider 3 amino-acid classes, the order of magnitude of the hitting time of a single switch-type mutational model significantly increases, and is of the order of 3^N, as proved by Proposition 2.2.9. Moreover we observe that the variance corresponding to the second model is significantly bigger as well.

It is clear that if we consider more amino-acid classes, it takes much longer to reach a precise element of the new state-space. Nevertheless, one can understand that if we keep the same distance function as defined in Equation (2.31), than we are asking for a higher degree of precision while building the B-cell trait. Therefore, we can not directly compare hitting times corresponding to a model with a greater number of amino-acid classes and keeping the same affinity function as the one used with only two amino-acid classes. If one want to obtain a comparable result by using more than two amino-acid classes, one has to use a weaker definition of affinity.

Definition 2.18. Let $\mathcal{S}$ be a set of letters, $|\mathcal{S}| = s > 2$. Let us partition $\mathcal{S}$ into two subsets: $\mathcal{S} := \mathcal{S}_1 \sqcup \mathcal{S}_2$. $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{S}^N$, their distance is given by:

$$h_{\mathcal{S}_1,\mathcal{S}_2}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \delta_i \quad \text{where} \quad \delta_i = \begin{cases} 1 & \text{if } x_i \in \mathcal{S}_1,\, y_i \in \mathcal{S}_2 \text{ or conversely} \\ 0 & \text{otherwise} \end{cases}$$

Consequently, their affinity is given by:

$$\text{aff}(\mathbf{x}, \mathbf{y}) = N - h_{\mathcal{S}_1,\mathcal{S}_2}(\mathbf{x}, \mathbf{y})$$

By using this new affinity function we can compare the hitting times and the order of magnitude is clearly the same.

Let us now go back to Assumption 2 and to the structure of the string

given in Section 2.4.3 (in particular, hydrophobic amino-acids are represented by empty cases). Contrary to what stated by Assumption 4, we suppose that the BCR length can be modified by insertions and deletions. Consequently, also a modification of the distance function is needed. We arbitrarily fixe a BCR and an antigen with given affinity. We do not consider those base substitutions leading to no detectable effect, *i.e.* at each time step we can observe a variation of the affinity function. We suppose that 90% of all mutation events are single point mutations, 10% deletions or insertions. If we are in this case and $|\mathbf{X}_t^{bin}| > |\overline{\mathbf{x}}'^{bin}|$, then with probability 1/2 a deletion occurs and with probability 1/2 an insertion occur. Otherwise, it will be necessarily an insertion (this is to avoid to obtain $|\mathbf{X}_t^{bin}| = 0$). As long as the affinity is concerned, if $|\mathbf{X}_t^{bin}| > |\overline{\mathbf{x}}'^{bin}|$, $|\mathbf{X}_t^{bin}| := n_1$, $|\overline{\mathbf{x}}'^{bin}| := n_2$, then their distance is the smaller possible one, *i.e.*:

$$h(\mathbf{X}_t^{bin}, \overline{\mathbf{x}}'^{bin}) = \min_{1 \leq i \leq n_1 - n_2 + 1} \left\{ h(\mathbf{X}_i, \overline{\mathbf{x}}'^{bin}) \,|\, \mathbf{X}_i := \left(X_{t,i}^{bin}, X_{t,i+1}^{bin}, \ldots, X_{t,i+n_2-1}^{bin}\right) \right\},$$

h as in Definition 2.16.

Table 2.8: Average number of mutations needed to reach $\overline{\mathbf{x}}'^{bin}$, for $N = 7$ and starting from a Hamming distance 5. In $\overline{\mathbf{x}}'^{bin}$, only 2 amino-acids are hydrophobic, so by Definition 2.16, the optimal Hamming distance one can reach is 2. We compare a model in which no deletions nor insertions are allowed and a model in which 10% of all mutations are deletions or insertions. We denote by $\widehat{\tau_{\{\overline{\mathbf{x}}'^{bin}\}_n}}$ the average value obtained over n simulations and by $\widehat{\sigma}_n$ its corresponding estimated standard deviation.

% **deletions/insertions**	$\|\overline{\mathbf{x}}'^{\boldsymbol{bin}}\|$	$\boldsymbol{h(\mathbf{X}_0^{bin}, \overline{\mathbf{x}}'^{bin})}$	$\boldsymbol{n}$	$\boldsymbol{\widehat{\tau_{\{\overline{\mathbf{x}}'^{bin}\}_n}}}$	$\boldsymbol{\frac{\widehat{\sigma}_n}{\sqrt{n}}}$
0	7	5	5000	374.28	5.38
10	7	5	5000	251.48	3.54

In this case, and thanks to the definition of Hamming distance as the minimal one, we clearly have more chances to obtain a good B-cell trait. This is confirmed by results collected in Table 2.8. When deletions and insertions can occur, even with very weak probability, and if we allowed the BCR length to be greater than the antigen one, then the expected number of mutations needed to built the optimal BCR is more than 30% smaller.

2.5 Conclusion

In this Chapter, we have introduced a mathematical framework to study the impact of various mutation rules on the exploration of the space of traits in an evolutionary model. In particular, we have connected mutation rules to characteristic time-scales, such as hitting-times, through the study of associated graph structures. As a leading example, which was the original motivation for this study, we have considered applications of these results to the modeling of somatic hypermutations in the germinal center. The models considered so far do not include division and selection, which would lead to studying branching random walks on graphs, a topic investigated in next Chapters.

Chapter 3

Branching random walks on binary strings for evolutionary processes

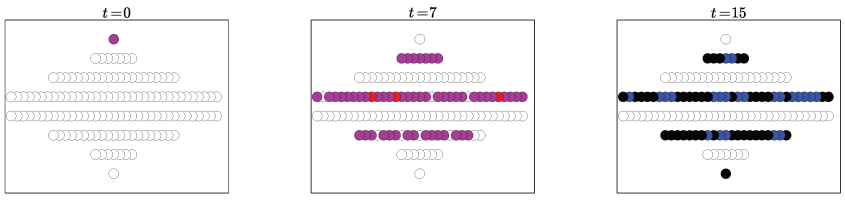
Summary In this Chapter, we study branching random walks on graphs modeling division-mutation processes inspired by adaptive immunity. We apply the theory of expander graphs on mutation rules in evolutionary processes and obtain estimates for partial cover times of branching random walks. This analysis reveals an unexpected saturation phenomenon: increasing the mutation rate above a certain threshold does not enhance the speed of state-space exploration.

3.1 Introduction

The aim of this Chapter is to understand interactions between mutation and division in evolutionary processes. In particular, we are interested in analyzing characteristic time-scales for which a certain proportion of possible traits is expressed in the population: starting from a single individual, what would be the typical time until a finite proportion of the traits are covered by the exponentially increasing population? In the models we consider, traits are represented as vertices of the N-dimensional hypercube, and the choice of a mutation rule corresponds to the prescription of a graph structure. The division-mutation process is then modeled as a Branching Random Walk (BRW) on this graph. A division event is always associated to mutation, meaning that the newborn particles move to neighboring nodes according to a given mutation rule. We consider two kinds of branching processes: a simple BRW (also called COBRA walk [43, 33]) where two or more particles having the same trait coalesce into

a single one, and the BRW with multiplicity, where we also take into account the number of individuals sharing the same trait within the population. The choice of a mutation rule influences the graph structure, and we show that the theory of expander graphs leads to new results on the typical time-scales of the state-space exploration. To our knowledge, the link between expander graphs theory and evolutionary process is new.

The motivation behind this work is the study of Antibody Affinity Maturation (AAM). This is a key process of the adaptive immune system, which allows to create specific high-affinity antibodies against pathogens that threaten a given organism. Antibodies are proteins which are secreted by B-cells, special lymphocytes trained to recognize the presented antigen [126]. This process takes place in Germinal Centers (GCs) [102, 136], where activated B-cells proliferate, mutate and differentiate. B-cells recognize the antigen thanks to transmembrane proteins called B-cell Receptors (BCRs) [36, 80]. The mutational mechanism that B-cells undergo during a GC reaction is called Somatic Hypermutation (SHM): it affects, at a very high rate, the DNA encoding for the specific portion of the BCR involved in the binding with the antigen, called Variable (V) region [133].



(a) **(b)** **(c)**

Figure 3.1: Simulation of the exploration of the state-space of all possible traits, considering division and simple switch-type mutations, *i.e.* a mutation consists in the switch of a randomly chosen bit. Traits are represented by 7-length binary strings and we give a picture of the evolution of the process at the beginning (a) after 7 time steps (b) and after 15 time steps (c). The process starts with a single individual whose trait corresponds to $\mathbf{x}_0 = [0, 0, 0, 0, 0, 0, 0]$, represented by the top circle. Starting from the top, on each line we arrange all nodes having an increasing distance from $\mathbf{x}_0$, so that, for instance, the bottom circle corresponds to the node $[1, 1, 1, 1, 1, 1, 1]$. Different colors denote a different number of individuals sharing the same trait. In particular in magenta we plotted nodes with at most 4 individuals lying on them, in red between 5 and 9, in blue between 300 and 499 and in dark more than 500 individuals.

A certain number of mathematical models and results about GC reaction and

AAM already exists. In particular, T. B. Kepler and A. S. Perelson in [75, 76] proposed deterministic population dynamics models for SHM and AAM, considering for the first time the hypothesis of the existence of a recycling mechanism of B-cells during GC reaction. This mechanism has now been confirmed by experiments [139]. In [105, 108, 52, 64] the authors introduced and discussed several immunological problems, such as the size of the repertoire, or the strength of antigen-antibody binding, while providing as well suitable mathematical tools. More recently, other articles have focused on biologically detailed models of the GC reaction (*e.g.* [92]), in particular with an agent-based modeling framework ([94], mostly analyzed through extensive numerical simulations). In 2015 the journal *Philosophical Transactions of the Royal Society B* has entirely dedicated an issue to the dynamics of antibody repertoires. For instance, in [45, 91, 32] the authors developed and applied modern statistical methods to investigate selection on BCRs and infer B-cell sequence evolution. We are interested in studying from an analytical point of view evolutionary pathways of BCRs during SHM. Here and in Chapters 2 and 4, we provide some significant building blocks in this direction and study their mathematical features.

Besides the biological motivations, the class of models studied in this Chapter is interesting from a mathematical point of view, as it is a discrete-time BRW on graphs, a type of branching process which has not been deeply investigated so far to our knowledge, despite its growing number of applications. Since the first articles about branching processes in the 50's and 60's [73, 16, 65, 66, 67], this class of stochastic processes has been used in various situations to model biological, genetic, physical, chemical or technological processes. For example branching processes can model the dynamics of population in genetics [116], or the spread of a piece of data, a rumor or a virus [13]. Most of the works that have been published so far are not interested in studying these processes on graphs. Nevertheless, in some recent papers [20, 21] the authors considers BRWs on multigraphs and mostly focus on weak and strong survival conditions. Dutta C. *et al* in [43] exhibits bounds on cover times for COBRA walks on trees, grids, and expander graphs (useful later in our analysis) in the context of gossip propagation. Results on expander graphs have been improved in [33] using a new duality relation between the COBRA walk and a discrete epidemic process. Another field of recent interest is the study of BRWs in random environnement. We refer, for example, to [3], where the authors study local and total particle populations or to [86] where conditions for recurrence and transience (almost surely wrt the random environment) are found, for the discrete-time BRW on a rooted tree with random environment. Branching annihilating RWs have been extensively studied in last years due to their applications in biological, chemi-

cal, physical and economical systems [90, 28, 29]. In [128] the authors consider these processes on random regular graphs, which they study using Monte Carlo simulations and generalized mean-field analysis.

In this Chapter, we focus on BRWs on $\{0,1\}^N$ with constant division rate 2 (except for Section 3.5.2), inspired by cellular division. The coupling of branching mechanism and random walk necessarily implies an important speedup in the characteristic time-scales of state-space exploration. Typically, for the simple random walk on the N-dimensional hypercube, the addition of a branching process enables a speedup from a time $\mathcal{O}(2^N)$ to $\mathcal{O}(N)$ (Section 3.4.2). Of course this has a cost: considering a branching process means also to produce new individuals at each time step. Indeed, in a time $T = \mathcal{O}(N)$ we have 2^T individuals (in the case in which multiplicity is taken into account; $\leq 2^N$ otherwise), as we do not consider here neither selection nor death. The mutation rule, which defines the structure of the graph, also determines the ability of the BRW in covering the vertices of the graph. In particular, using expansion properties, in Section 3.4 we prove that the best result we can obtain in a time $\mathcal{O}(N)$ for finite connected expander graphs over the state-space $\{0,1\}^N$, is to cover a half of the graph.

Moreover, our mathematical analysis of the partial cover times has revealed an interesting phenomenon concerning the impact of the mutation rate on the exploration speed. Intuitively, one would suggest that increasing the number of mutations at each division would result in a BRW with a faster exploration time-scale. However, we show in Section 3.4.3 the existence of an early saturation phenomenon: when increasing from one to two mutations, the exploration is indeed faster, but allowing more than two mutations (up to N) modifies only marginally the exploration speed.

In Section 3.2 we state the main definitions and notations setting up a general mathematical framework. Section 3.3 contains preliminary results concerning generic BRWs on graphs and their possible bipartite structure. Bipartiteness influences the dynamic of the branching process. In Section 3.4, we establish quantitative results concerning the portion of the state-space invaded in $\mathcal{O}(N)$ for two different kinds of BRWs (Theorems 3.4.9 and 3.4.13). In order to do so, we need to determine some characteristics of the graphs, in particular their expansion properties. These results provide quantitative estimates of the typical time-scale for state-space exploration resulting from the interaction between division and mutation. Then, in Section 3.5, we propose some extensions of the model. In particular, we introduce the BRW with multiplicity and obtain

the transition matrix related to the number of individuals carrying a given trait together with their limiting distribution. We investigate as well how this distribution can change by introducing a division rate, and provide comparisons between different mutation/division models. In this way, theoretical results presented in previous sections are displayed in a wider context. Finally, in Section 3.6 we conclude with a brief summary of this work and discuss the biological setting in which it is embedded justifying our hypotheses. We present as well consequences of our results and discuss possible improvements in order to cover the state-space faster in time, or to drive the covering to main interest areas of the graph.

3.2 Definitions and Notations

We start this section with some definitions and notations, establishing an elementary mathematical framework for the modeling of antibody affinity maturation in the germinal center.

We first assume that it is possible to classify the amino acids, which determine the chemical properties of both epitope and paratope, into 2 classes, typically positively charged and negatively charged. Henceforth BCRs and antigen are represented by binary strings of a same length N, hence, the state-space of all possible BCR configurations is $\{0,1\}^N$ (we refer to Chapter 2 for more details).

Definition 3.1. We denote by $\mathcal{H}_N$ the standard N-dimensional hypercube. BCR and antigen configurations are represented by vertices of $\mathcal{H}_N$, denoted by $\mathbf{x}_i$ with $1 \leq i \leq 2^N$, or sometimes simply by their indices.

In this Chapter we introduce and discuss models including mutation and division. Mathematically, this gives rise to BRWs on $\{0,1\}^N$. The structure of the graph depends then on the mutation rule we consider.

We suppose that there is a single B-cell entering the GC reaction. At each time step, each B-cell divides and mutates according to a given mutational rule. A mutation corresponds to a jump on a neighbor node.

Definition 3.2. Given $\mathbf{x}_i$, $\mathbf{x}_j \in \{0,1\}^N$, we say that $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors, and denote $\mathbf{x}_i \sim \mathbf{x}_j$, if there exists at least one edge (or loop) between them.

We are mostly interested in studying the variation of the number of expressed traits within the population, as a result of the interaction between division and

mutation. In this Chapter we refer to two different kinds of BRWs: the simple c-BRW (also called coalescing BRW [43]) and the c-BRW with multiplicity.

Definition 3.3 (Simple **c**-BRW)**.** The process starts at an arbitrary node (representing the BCR of a B-cell entering the process of division and mutation during the GC reaction), labelled as active. If at time t node $\mathbf{x}_i$ is active (*i.e.* the trait $\mathbf{x}_i$ is expressed in the GC population at time t), then at time $t+1$ it chooses c of its neighbors, independently and with replacement, to become active, while $\mathbf{x}_i$ becomes inactive again (unless, of course, another active node at time t chooses it). In this model, the number of times a node is chosen to become active is not taken into account. We suppose $c > 1$, otherwise the BRW simply becomes a RW.

Definition 3.4 (**c**-BRW with multiplicity)**.** The process starts with a B-cell entering the process of mutation and division, lying on an arbitrary node which corresponds to its trait. At each time step a particle lying on a certain node $\mathbf{x}_i$ of $\{0,1\}^N$ gives rise to c daughter cells, with $c > 1$, and die. Each one of the c newborn particles choses a neighbor node, independently and with replacement, and move on it. More than one particle can lie on the same vertex of $\mathcal{H}_N$, and each one divides at each time step.

Notation 2. Let $S \subseteq V$ be a subset of vertices of a graph $G = (V, E)$. Then we denote by $\mathcal{N}(S)$ the set of the neighbors of all vertices in S. We denote by $|S|$ and $|\mathcal{N}(S)|$ the number of vertices in S and in $\mathcal{N}(S)$ respectively. $\mathcal{N}(S)$ may include also some vertices of S.

Notation 3. Given a simple c-BRW on a generic graph G, for all $t \geq 0$ we note by S_t the set of all active nodes at time t and by $\mathcal{N}(S_t)$ the set of all the neighbors of the vertex set S_t.

The structure of the graph, and consequently the dynamics of the BRW on it, depends on the introduced mutation rule, which is defined thanks to the transition probability matrix.

Definition 3.5. Let $\mathcal{M}$ be the transition probability matrix of a graph G. We denote the BRW referring to $\mathcal{M}$ and with constant division rate c by c-BRW-$\mathcal{M}$.

In particular, we refer to two mutational rules (see Chapter 2 for more details). Here below we give the definitions of the corresponding transition probability matrices.

Definition 3.6. For all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{H}_N$:

$$\mathbb{P}(\mathbf{X}_n = \mathbf{x}_j \,|\, \mathbf{X}_{n-1} = \mathbf{x}_i) =: p(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1/N & \text{if} \quad \mathbf{x}_j \sim \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases}$$

Matrix $\mathcal{P} := (p(\mathbf{x}_i, \mathbf{x}_j))_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{H}_N}$ gives to $\{0,1\}^N$ the structure of a standard N-dimensional hypercube.

We further introduce another transition matrix, which models a mutation rule in which up to k symbols of the string are independently mutated at each division:

Definition 3.7. Let $k \in \{1, \dots, N\}$, $\mathcal{P}^{(k)} := \dfrac{1}{k} \displaystyle\sum_{i=1}^{k} \mathcal{P}^i$, $\mathcal{P}$ given by Definition 3.6.

We finally recall the definition of Hamming distance, which measures, in our model, the affinity between two traits (Chapter 2):

Definition 3.8. For all $\mathbf{x} = (x_1, \dots, x_N)$, $\mathbf{y} = (y_1, \dots, y_N) \in \{0,1\}^N$, their Hamming distance is given by:

$$h(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \delta_i \qquad \text{where} \qquad \delta_i = \begin{cases} 1 & \text{if} \quad x_i \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

Definition 3.9. For all $\mathbf{x}_i \in \{0,1\}^N$, its affinity with a given vertex $\overline{\mathbf{x}}$, $\text{aff}(\mathbf{x}_i, \overline{\mathbf{x}})$ is given by $\text{aff}(\mathbf{x}_i, \overline{\mathbf{x}}) := N - h(\mathbf{x}_i, \overline{\mathbf{x}})$, where $h(\cdot, \cdot) : (\{0,1\}^N \times \{0,1\}^N) \to \{0, \dots, N\}$ returns the Hamming distance.

3.3 c-BRW on graphs and bipartiteness

The bipartiteness deeply influences the characteristics of the BRW and, in particular, its possibility of covering all nodes of the graph simultaneously at a certain time.

Definition 3.10. A graph $G = (V, E)$ is bipartite if there exists a partition of the vertex set $V = V_1 \sqcup V_2$, s.t. every edge connects a vertex in V_1 to a vertex in V_2.

We emphasize the relations between a generic c-BRW on a given graph $G = (V, E)$, with $c \geq 2$, and the eventual bipartite structure of the above-mentioned graph.

3.3.1 c-BRW on bipartite graphs

Let us consider a simple c-BRW on a generic bipartite graph $G_b(V_1 \sqcup V_2, E)$. Instead of considering a single random active node at the beginning, we suppose that the process starts with a given initial distribution $\mathbf{p}$ of the active set. The results presented in this section do not change if we consider a c-BRW with multiplicity instead of a simple c-BRW. The fact that the trials are made with replacement does not have any consequences either.

Proposition 3.3.1. *If the initial distribution* $\mathbf{p}$ *is concentrated on* V_1 *or on* V_2 *then* $|S_t| \leq \max_{i=1,2}(|V_i|)$ *for all* $t \geq 0$*, otherwise* $S_t = V_1 \sqcup V_2$ *for some* $t > 0$ *with positive probability.*

Proof. The proof is a direct consequence of the bipartite structure of G_b. Let us suppose, without loss of generality, that $\mathbf{p}$ is concentrated on V_1. Then, due to the bipartite structure of G_b after an even number of steps we have necessarily $S_{2t} \subseteq V_1$, while after an odd number of steps we have $S_{2t+1} \subseteq V_2$, and so the first statement is proven.
If, on the contrary, $\mathbf{p}$ is not concentrated on V_1 nor on V_2, then for all $t \geq 0$ we have a positive probability that $S_t = S_{t,1} \sqcup S_{t,2}$ with $S_{t,1} \subseteq V_1$ and $S_{t,2} \subseteq V_2$, and consequently, w.p.p. we have $S_t = V_1 \sqcup V_2$ for some $t > 0$. □

Remark 18. This qualitative result does not change if we take into account the number of times a node is chosen to become active for the next time step or if we decide to make trials without replacement: these choices only have effects on the speed of the covering.

3.3.2 c-BRW on non-bipartite connected graphs

Let us now consider a non-bipartite connected graph $G = (V, E)$. We recall a classical result about bipartite graphs [114], which will be useful later:

Proposition 3.3.2. *A graph is bipartite if and only if it has no odd cycles.*

We shall prove the following statement:

Theorem 3.3.3. *Given a c-BRW on a finite non-bipartite connected graph, then, w.p.p., there exists a time* $t > 0$ *such that* $S_t = V$*.*

The proof of this theorem is based on the following three lemmas:

Lemma 3.3.4. *If* $G = (V, E)$ *is a finite connected graph, then, independently from the initial distribution,* $\forall\, \mathbf{x}_i \in V$ *there exists a time* $t < \infty$ *s.t.* $\mathbf{x}_i \in S_t$*.*

In other words, if the graph is finite and connected, then each node will be activated by the BRW at least once in a finite time interval.

Proof. The hitting time of the c-BRW to reach any node of a finite connected graph, starting from every possible initial distribution is finite, thanks to the connectivity of the graph and the fact that it has a finite set of nodes. (Note that this is still true if $c = 1$, *i.e.* for a SRW on a finite connected graph). □

Lemma 3.3.5. *If there exists a time $t \geq 0$ such that $\exists\ \mathbf{x}_1,\ \mathbf{x}_2,\ \mathbf{x}_1 \sim \mathbf{x}_2$ and $\{\mathbf{x}_1, \mathbf{x}_2\} \in S_t$, then w.p.p. there exists a time $T > t$ s.t. $S_T = V$ (independently from the initial distribution).*

This means that if at a given time t we have two neighbor nodes both active, then we have a positive probability to reach $S_T = V$ later.

Proof. Let us suppose that $\mathbf{x}_1$, $\mathbf{x}_2$ are two neighbor nodes and $S_{t^*} = \{\mathbf{x}_1, \mathbf{x}_2\}$ (we suppose that all other nodes are non-active). Then we are able to show that w.p.p., for all $t \geq t^*$, $S_t \subset \mathcal{N}(S_t)$ and $S_t = \mathcal{N}(S_t) \Leftrightarrow S_t = V$, where we recall that $\mathcal{N}(S_t)$ is the set of all neighbors of S_t. This implies that w.p.p. the active set can always grow until we reach $S_t = V$. This result is quite intuitive, indeed if $\mathbf{x}_1 \sim \mathbf{x}_2$ and $S_{t^*} = \{\mathbf{x}_1, \mathbf{x}_2\}$, then necessarily $S_{t^*} \subset \mathcal{N}(S_{t^*})$ and, consequently, there is a positive probability that $S_{t^*} \subset S_{t^*+1}$. That means that w.p.p. S_{t^*+1} contains x_1, x_2 and at most $c-1$ distinct neighbors of $\mathbf{x}_1$ and $c-1$ distinct neighbors of $\mathbf{x}_2$. Then we can repeat the same argument with all the couples of neighbors active at time t^*+1 (w.p.p. all nodes in S_{t^*+1} are neighbors two by two). Thanks to the connectivity of the graph and the fact that it is a finite graph, w.p.p. we can go on with this procedure until we reach $S_t = V$. □

Lemma 3.3.6. *If there exists at least an odd cycle on $G = (V, E)$, then, independently from the initial distribution, w.p.p. for a time $t \geq 0$ there exist two nodes $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_1 \sim \mathbf{x}_2$ and $\mathbf{x}_1$, $\mathbf{x}_2 \in S_t$.*

Proof. Let us suppose that in graph G there exists an odd cycle of length $2n+1$: $C = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{(2n+1)}, \mathbf{x}_1)$. Lemma 3.3.4 implies the existence of a time $T < \infty$ s.t. $\mathbf{x}_1 \in S_T$. Then w.p.p. we have that $\{\mathbf{x}_2, \mathbf{x}_{(2n-1)}\} \subseteq S_{T+1}$ (we recall that $c \geq 2$). We make another step and w.p.p. we have that $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_{2n}\} \subseteq S_{T+2}$. After n steps, w.p.p. we have $\{\mathbf{x}_{(n+1)}, \mathbf{x}_{(n+2)}\} \subseteq S_{T+n}$, and $\mathbf{x}_{(n+1)} \sim \mathbf{x}_{(n+2)}$, which proves Lemma 3.3.6. □

Proof. (Theorem 3.3.3) Let $G = (V, E)$ be a finite non-bipartite connected graph and $C = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{(2n+1)}, \mathbf{x}_1)$ an odd cycle of G (it necessarily exists as G is non-bipartite (Proposition 3.3.2)). Lemma 3.3.4 assures that there exists a finite time t_1 s.t. $\mathbf{x}_1 \in S_{t_1}$. Then, thanks to Lemma 3.3.6, w.p.p. there exists a time $t_2 > t_1$ s.t. $\{\mathbf{x}_{(n+1)}, \mathbf{x}_{(n+2)}\} \subseteq S_{t_2}$, and $\mathbf{x}_{(n+1)} \sim \mathbf{x}_{(n+2)}$. The proof of Theorem 3.3.3 can now be achieved by applying Lemma 3.3.5. □

3.4 Portion of $\mathcal{H}_{\mathbf{N}}$ covered in $\mathcal{O}(\mathbf{N})$ for the simple 2-BRW-$\mathcal{P}$ and the simple 2-BRW-$\mathcal{P}^{(\mathbf{k})}$

In this Section our aim is to estimate the size of the active nodes set in a time of the order of N. It clearly depends on the mutational model allowed on the state-space. We can interpret it as the number of possible BCR configurations expressed in our population after $\mathcal{O}(N)$ mutation steps. In Table 3.1 we summarize the main results of the current section. In Sections 3.4.2 and 3.4.3 we estimate the size of the active set in $\mathcal{O}(N)$ for the simple 2-BRW referring to $\mathcal{P}$ and $\mathcal{P}^{(k)}$ (Definitions 3.6 and 3.7). We prove that the 2-BRW-$\mathcal{P}$ covers a small portion of $\mathcal{H}_N$, while a half of the state-space will be covered if we take into account $\mathcal{P}^{(k)}$ as transition probability matrix, at least for N big enough.

Table 3.1: Summary of the main results of Sections 3.4.2 and 3.4.3.

Model	$\mathbf{\lvert S_T\rvert}$ **in** $\mathbf{T} = \mathcal{O}(\mathbf{N})$
$\mathcal{P}$	$\lvert S_T\rvert \geq 2^{N-r}$, $r > \dfrac{N^2e^{-2}+N-2}{Ne^{-2}+N-2}$
$\mathcal{P}^{(\mathbf{k})}$	$\lvert S_T\rvert \geq \delta 2^N$, $\delta \leq 1/2$

In order to estimate these quantities, we apply a method used in [43] to determine the partial cover time for expander graphs. The partial cover time corresponds to the expected time required to visit at least a certain portion of the state-space. We need to evaluate the expansion properties of graphs described by $\mathcal{P}$ and $\mathcal{P}^{(k)}$ respectively. For this reason in Section 3.4.1 we recall some definitions and results about expander graphs. For a more complete overview about this subject see *e.g.* [62].

3.4.1 Expander graphs

Informally, an expander graph is a graph $G = (V, E)$ which has strong connectivity properties (quantified using vertex, edge or spectral expansion). We give some mathematical characterization of this property.

Unless stated otherwise, throughout this section a graph $G = (V, E)$ is a connected undirected d-regular graph with $|V| = n$.

Definition 3.11 ((α, δ)-expander graph)**.** G is said to be an (α, δ)-expander graph, with $\delta \leq 1/2$, if: $\forall\, S \subseteq V$ s.t. $|S| \leq \delta n \;\Rightarrow\; |\mathcal{N}(S)| \geq \alpha|S|$.

In other words, an (α, δ)-expander graph is a graph where the set of all neighbors of each subset S with at most δn nodes, has at least $\alpha|S|$ vertices.

Spectrum and expansion

Let us denote by A_G the adjacency matrix of G and by $\mathcal{P}_G$ its transition probability matrix. As G is a d-regular graph, then $\mathcal{P}_G = \frac{1}{d}A_G$. We denote by $d = \lambda_1^A \geq \lambda_2^A \geq \cdots \geq \lambda_n^A$ and $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ the eigenvalues of A_G and $\mathcal{P}_G$ respectively.

Definition 3.12. We say that G is a λ eigenvalue expander, with $\lambda < d$, if $\lambda_2^A \leq \lambda$. It is a λ absolute eigenvalue expander if $|\lambda_2^A|, |\lambda_n^A| \leq \lambda$.

Remark 19. All d-regular connected graphs are λ_2^A eigenvalue expanders. Indeed under these hypotheses, the first largest eigenvalue of the adjacency matrix corresponds to d and $d > \lambda_2^A \geq \lambda_i^A$ for all $i \geq 2$.

Then we have the following known result (first proved by R. M. Tanner in [129]):

Theorem 3.4.1 (Vertex expansion)**.** *Let G be a λ eigenvalue expander. Let $S \subseteq V$ s.t. $|S| \leq n/2$. $\mathcal{N}(S)$ is large, in particular:*

$$|\mathcal{N}(S)| \geq \frac{|S|}{\frac{\lambda^2}{d^2} + \left(1 - \frac{\lambda^2}{d^2}\right)\frac{|S|}{n}}$$

Remark 20. One easily notices that $\left(\frac{\lambda^2}{d^2} + \left(1 - \frac{\lambda^2}{d^2}\right)\frac{|S|}{n}\right)^{-1} \to 1$ for $\lambda \to d$, is decreasing wrt λ.

We also give another characterization of d-regular expander graphs with respect to their eigenvalues.

Definition 3.13. We say that G is an ε-expander graph, with $\varepsilon < 1$, if the eigenvalues of its adjacency matrix are such that $|\lambda_i^A| \leq \varepsilon d$ for $i \geq 2$.

Then in particular, we have the following proposition.

Proposition 3.4.2. *Let G be not bipartite. G is a λ_2-expander graph.*

Proof. As $\mathcal{P}_G = \frac{1}{d}A_G$, we have that: $|\lambda_i^A| = |\lambda_i| \cdot d \leq \lambda_2 \cdot d$, $\forall\, i \geq 2$. This is not true for bipartite graphs as their spectrum is symmetric with respect to zero. Therefore $|\lambda_n^A| = d > \lambda_2 \cdot d$. □

As an immediate consequence of Theorem 3.4.1 applied to ε-expander graphs, we have:

Proposition 3.4.3. *Let G be a ε-expander graph. For all $S \subseteq V$ s.t. $|S| \leq \delta n$, $\delta \leq 1/2$:*

$$|\mathcal{N}(S)| \geq \frac{|S|}{\varepsilon^2(1-\delta)+\delta}$$

Finally, let us underline the clear relation existing between Definitions 3.13 and 3.11 of ε-expander graphs and (α, δ)-expander graphs respectively:

Corollary 3.4.4. *Let G be not bipartite with second largest eigenvalue λ_2, $\delta \leq 1/2$. G is a (α, δ)-expander graph with:*

$$\alpha = \frac{1}{\lambda_2^2(1-\delta)+\delta}$$

Proof. First of all, Proposition 3.4.2 tells us that G is a λ_2-expander graph. Then the condition on α is given by Proposition 3.4.3. □

3.4.2 Simple 2-BRW-$\mathcal{P}$

A simple 2-BRW-$\mathcal{P}$ on $\mathcal{H}_N$ is a generalization of a Simple RW on $\mathcal{H}_N$ (Chapter 2). We want to estimate the size of the active set in $\mathcal{O}(N)$ using $\mathcal{P}$ as transition probability matrix. In order to do so, we use an application of a more general method used in [43] to evaluate partial cover times. We show that the partial cover time for the simple 2-BRW-$\mathcal{P}$ is linear in N, while we already know that for the SRW on $\mathcal{H}_N$ it is exponential in N [8]. This highlights how the branching process gives an important speedup in exploring the hypercube. This speedup in covering is not without a cost. Indeed, for a time t large enough, the size of the population will be of the order of the maximal possible size of S_t, which is 2^{N-1} in this case (as $\mathcal{H}_N$ is bipartite) and 2^N in the case of the simple 2-BRW-$\mathcal{P}^{(k)}$.

Let us start with a preliminary result about the standard N-dimensional hypercube, $\mathcal{H}_N$.

Proposition 3.4.5. *For any $N \geq 1$, $\mathcal{H}_N$ is a N-regular $(r, 2^{-r})$-expander graph, where $r \in \{1, \dots, N\}$,* i.e.*:*

$$\forall r \in \{1,\dots,N\},\ \forall S \subset \{0,1\}^N \text{ s.t. } |S| \leq 2^{N-r} \quad \Rightarrow \quad |\mathcal{N}(S)| \geq r|S|$$

Before giving the proof of Proposition 3.4.5, let us observe the maximal number of common neighbors among two or more nodes in $\mathcal{H}_N$.

Remark 21. Two distinct vertices $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{H}_N$ cannot share more than two common neighbors. More generally, s distinct vertices in $\mathcal{H}_N$, $\{\mathbf{x}_i\}_{1 \leq i \leq s \leq 2^N}$ cannot share more than s common neighbors.

Let A_N be the standard representation of the transition probability matrix of $\mathcal{H}_N$, obtained recursively as follows [51]:

$$A_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \; A_N = \left(\begin{array}{c|c} A_{N-1} & \mathcal{I}_{2^{N-1}} \\ \hline \mathcal{I}_{2^{N-1}} & A_{N-1} \end{array}\right),$$

where $\mathcal{I}_{2^{N-1}}$ is the 2^{N-1}-identity matrix. The result is obvious since the main diagonal of A_{N-1} is composed by zeros and that A_{N-1} is a symmetric matrix.

Proof. (Proposition 3.4.5) We prove Proposition 3.4.5 by double induction on N and on r.

First of all, the statement is true for $N = 1$ and $r = 1$, and for $N = 2$ and $r \in \{1, 2\}$. We suppose the statement true up to dimension $N - 1$ and for all $r \in \{1, \ldots, N-1\}$, and we prove it for dimension N and for all $r \in \{1, \ldots, N\}$.

If $r = N$ it is true, as $\mathcal{H}_N$ is a N-regular graph. Let $r = N - 1$. Then we want to show that $\forall\, S \subset \{0,1\}^N$ s.t. $|S| \leq 2 \;\Rightarrow\; |\mathcal{N}(S)| \geq (N-1)|S|$. If $|S| = 1$, for the N-regularity we have necessarily: $|\mathcal{N}(S)| = N > N - 1$.

We suppose $|S| = 2$, and we consider the graph underlined by A_N. If we choose both vertices $\mathbf{x}_i$ with $0 \leq i \leq 2^{N-1}$, then we know, for the induction hypothesis on N and observing that the top right block of A_N is an identity matrix, that: $|\mathcal{N}(S)| \geq (N-2)|S| + |S| = (N-1)|S|$.

Let us consider two vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ s.t. $i \in \{1, \ldots, 2^{N-1}\}$ and $j \in \{2^{N-1}+1, \ldots, 2^N\}$. If we do not want to increase considerably $|\mathcal{N}(S)|$, once $\mathbf{x}_i$ is fixed, we need to choose $\mathbf{x}_j$ so that $\mathbf{x}_i$ and $\mathbf{x}_j$ share two common neighbors (Remark 21). Then, at least we have $|\mathcal{N}(S)| \geq 2N - 2 = (N-1)2 = (N-1)|S|$.

We suppose that the statement is true for dimension N and for all $r \in \{t+1, \ldots, N\}$. We prove that it's also true for $r = t$, *i.e.*:

$$\forall\, S \subset \{0,1\}^N \text{ s.t. } |S| \leq 2^{N-t} \;\Rightarrow\; |\mathcal{N}(S)| \geq t|S|$$

If $|S| \leq 2^{N-(t+1)} < 2^{N-t}$, for the induction hypothesis on r, we have:

$$|\mathcal{N}(S)| \geq (t+1)|S| > t|S|$$

Let us suppose $2^{N-(t+1)} < |S| \leq 2^{N-t}$. Again, if we choose all vertices $\mathbf{x}_i$ so

that $i \in \{1, \dots, 2^{N-1}\}$, for the induction hypothesis on N and as $r < N-1$, we have: $|\mathcal{N}(S)| \geq (t-1)|S| + |S| = t|S|$.

Let $S = \{\mathbf{x}_i\}_{1 \leq i \leq 2^{N-t}}$ so that:

$$S = S_1 \sqcup S_2 \quad S_1 = \{\mathbf{x}_{i_1}\}_{1 \leq i_1 \leq 2^{N-1}} \quad \text{and} \quad S_2 = \{\mathbf{x}_{i_2}\}_{2^{N-1}+1 \leq i_2 \leq 2^N}$$

Furthermore, we suppose: $|S_1| \leq 2^{N-(t+1)}$ and $|S_2| \leq 2^{N-(t+1)}$, as the other cases are less favorable, if our purpose is to minimize $|\mathcal{N}(S)|$. From the induction hypothesis on N, together with Remark 21:

$$|\mathcal{N}(S)| = |\mathcal{N}(S_1 \sqcup S_2)| \geq t|S_1| + |S_1| + t|S_2| + |S_2| - |S| = t|S|.$$

□

Remark 22. Considering a simple c-BRW-$\mathcal{P}$ starting from a single node, we have that $\mathcal{N}(S_t) \cap S_t = \emptyset$ because of the bipartite structure of the graph. This is not true for generic non-bipartite graphs (see Section 3.3).

We start by demonstrating the following lemma:

Lemma 3.4.6. *Given a simple 2-BRW-$\mathcal{P}$:*

$$\forall\, t \geq 0 \text{ s.t. } |S_t| \leq 2^{N-r} \quad \Rightarrow \quad \mathbb{E}[|S_{t+1}|] \geq (1+\nu)|S_t|$$

for some constant $\nu > 0$ and for $r > \frac{N^2 e^{-2} + N - 2}{N e^{-2} + N - 2}$.

Before demonstrating Lemma 3.4.6, we prove an elementary result, that we will need later:

Lemma 3.4.7. *Let $c > 0$ and $a, b > 1$ such that $a \leq b$. Then:*

$$e^{-ca} + e^{-cb} < e^{-c(a-1)} + e^{-c(b+1)}$$

Proof.

$$\begin{aligned} e^{-ca} + e^{-cb} - \left(e^{-c(a-1)} + e^{-c(b+1)}\right) &= e^{-ca}\left(1 - e^{c}\right) + e^{-c(b+1)}\left(e^{c} - 1\right) \\ &= \left(1 - e^{c}\right)\left(e^{-ca} - e^{-c(b+1)}\right) \\ &< 0 \end{aligned}$$

since $c > 0$ and $a < b + 1$ □

Proof. (Lemma 3.4.6) Let $t \geq 0$ so that $|S_t| \leq 2^{N-r}$, for a certain $r \in$

$\{1, \ldots, N\}$ that we will discuss later. The claim is proved if we show:

$$\mathbb{E}[|\mathcal{N}(S_t) - S_{t+1}|] \leq |\mathcal{N}(S_t)| - (1+\nu)|S_t| \tag{3.1}$$

For all vertices $v \in \mathcal{N}(S_t)$, let X_v be the indicator variable:

$$X_v = \begin{cases} 1 & \text{if } v \notin S_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

Then we have: $\mathbb{P}[X_v = 1] = \left(1 - \frac{1}{N}\right)^{2d_v} =: p$, where d_v represents the number of edges connecting v to S_t ($1 \leq d_v \leq N$).
Clearly $\mathbb{E}[X_v] = p$. Now we have:

$$\begin{aligned} \mathbb{E}[|\mathcal{N}(S_t) - S_{t+1}|] &\leq \mathbb{E}\left[\sum_{v \in \mathcal{N}(S_t)} X_v\right] = \sum_{v \in \mathcal{N}(S_t)} \left(1 - \frac{1}{N}\right)^{2d_v} \\ &\leq \sum_{v \in \mathcal{N}(S_t)} e^{-\frac{2d_v}{N}} \end{aligned}$$

Thanks to Lemma 3.4.7, we can claim that this expression is maximized if for any v (except possibly for one) d_v is either 1 or N. In particular let us suppose that all d_v are equal to 1 or to N and let us denote:

$$R_1 = |\{v \in \mathcal{N}(S_t) \,|\, d_v = 1\}| \quad \text{and} \quad R_N = |\{v \in \mathcal{N}(S_t) \,|\, d_v = N\}|$$

If we are able to demonstrate the result in this particular case, then it will be true for all possible distributions of d_v in $\mathcal{N}(S_t)$. Observing that $\sum_{v \in \mathcal{N}(S_t)} d_v = N|S_t|$ thanks to the N regularity, we have:

$$\begin{cases} R_1 + R_N = |\mathcal{N}(S_t)| \\ R_1 + NR_N = N|S_t| \end{cases} \Rightarrow \begin{cases} R_1 = \dfrac{N}{N-1}(|\mathcal{N}(S_t)| - |S_t|) \\ \\ R_N = \dfrac{1}{N-1}(N|S_t| - |\mathcal{N}(S_t)|) \end{cases}$$

Then:

$$\begin{aligned} \mathbb{E}[|\mathcal{N}(S_t) - S_{t+1}|] &\leq R_1 e^{-\frac{2}{N}} + R_N e^{-2} \\ &= \frac{N}{N-1}(|\mathcal{N}(S_t)| - |S_t|)e^{-\frac{2}{N}} + \frac{1}{N-1}(N|S_t| - |\mathcal{N}(S_t)|)e^{-2} \end{aligned}$$

In order to obtain (3.1), we have to impose that:

$$\frac{N}{N-1}(|\mathcal{N}(S_t)|-|S_t|)e^{-\frac{2}{N}}+\frac{1}{N-1}(N|S_t|-|\mathcal{N}(S_t)|)e^{-2}\leq|\mathcal{N}(S_t)|-(1+\nu)|S_t|$$

This is equivalent to:

$$|\mathcal{N}(S_t)|\left(1-\frac{N}{N-1}e^{-\frac{2}{N}}+\frac{1}{N-1}e^{-2}\right)+|S_t|\left(\frac{N}{N-1}e^{-\frac{2}{N}}-\frac{N}{N-1}e^{-2}-1\right)\geq\nu|S_t|$$

By hypothesis $|S_t|\leq 2^{N-r}$, which implies $|\mathcal{N}(S_t)|\geq r|S_t|$ (Proposition 3.4.5). Since $1-\frac{N}{N-1}e^{-\frac{2}{N}}+\frac{1}{N-1}e^{-2}>0$, the last inequality is true if:

$$r\left(1-\frac{N}{N-1}e^{-\frac{2}{N}}+\frac{1}{N-1}e^{-2}\right)+\left(\frac{N}{N-1}e^{-\frac{2}{N}}-\frac{N}{N-1}e^{-2}-1\right)\geq\nu \quad (3.2)$$

Therefore, our aim is to find $r(N)$ s.t. for all $r>r(N)$:

$$r\left(1-\frac{N}{N-1}e^{-\frac{2}{N}}+\frac{1}{N-1}e^{-2}\right)+\left(\frac{N}{N-1}e^{-\frac{2}{N}}-\frac{N}{N-1}e^{-2}-1\right)>0 \quad (3.3)$$

This is true iff:

$$r>\frac{Ne^{-2}+N-1-Ne^{-\frac{2}{N}}}{e^{-2}+N-1-Ne^{-\frac{2}{N}}}=:r(N) \quad (3.4)$$

We rearrange (3.3) writing:

$$(r-1)\left(1-\frac{N}{N-1}e^{-\frac{2}{N}}\right)-\frac{N-r}{N-1}e^{-2}>0$$

Since $e^{-\frac{2}{N}}\leq 1-\frac{2}{N}+\frac{2}{N^2}$ (thanks to the second-order Taylor expansion with integral rest), we obtain that (3.3) is satisfied if:

$$(r-1)\left(1-\frac{N}{N-1}\left(1-\frac{2}{N}+\frac{2}{N^2}\right)\right)-\frac{N-r}{N-1}e^{-2}>0$$

And finally:

$$r>\frac{N^2e^{-2}+N-2}{Ne^{-2}+N-2}$$

□

Remark 23.

- If $N\geq 2$, the condition on r that we found in Lemma 3.4.6 is met if:

$$r>1+Ne^{-2}\left(\frac{N-1}{N-2}\right)$$

- If $N \geq 3$, this condition is satisfied if $r > 1 + 2Ne^{-2}$.

We could also express r as a function of ν (we refer to (3.2)):

Corollary 3.4.8. $\mathbb{E}[|S_{t+1}|] \geq (1+\nu)|S_t|$ *for some constant* $\nu > 0$ *and for*

$$N \geq r \geq \frac{\nu(N-1) + Ne^{-2} - Ne^{-\frac{2}{N}} + N - 1}{e^{-2} - Ne^{-\frac{2}{N}} + N - 1} := r_N(\nu)$$

Therefore $|S_t|$ *has an exponential growth with rate* ν *until it reaches the size of* 2^{N-r} *and for* $r \geq r_N(\nu)$*. Moreover, as expected, if we define* ν^* *as the bigger admissible* ν*,* i.e. $\nu^* = \sup\{\nu \mid r_N(\nu) \leq N\}$*, then* $\nu^* \leq 1$*.*

Proof. The proof consists in elementary computations, starting from (3.2). In particular as far as the second statement is concerned, we impose $r_N(\nu) \leq N$, and clearly this condition is satisfied iff:

$$\nu \leq N - 1 - Ne^{-\frac{2}{N}}$$

Then, as $e^{-\frac{2}{N}} \geq 1 - \frac{2}{N}$, we can conclude. □

We are now able to state the following result:

Theorem 3.4.9. *Given a simple 2-BRW-*$\mathcal{P}$*, there exists a time* T *such that* $T = \mathcal{O}(N)$ *and with high probability* $|S_T| \geq 2^{N-r}$*,* r *satisfying the hypothesis of Lemma 3.4.6.*

Proof. The proof is a direct application of a result obtained for generic expander graphs in [43], Section 4. This result applies to our specific case thanks to Lemma 3.4.6. The main idea to prove Theorem 3.4.9 is to describe the change in the number of active nodes as a Markov process which lower bounds the growth of the size of the active set $|S_t|$. The statement is proven for this Markov process and, consequently, it is true also for our BRW. □

3.4.3 Simple 2-BRW-$\mathcal{P}^{(k)}$

Let us start by examining an analog of Lemma 3.4.6 for the 2-BRW-$\mathcal{P}^{(k)}$, where we recall that $\mathcal{P}^{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathcal{P}^i$ (Definition 3.7). We show that in this case the BRW covers a significantly bigger proportion of vertices in a time $\mathcal{O}(N)$. We follow again the method used in [43].

First of all, we prove that the 2-BRW-$\mathcal{P}^{(k)}$ allows, for $k \geq 2$, an exponential growth until it covers at least a half of the vertex set of the hypercube:

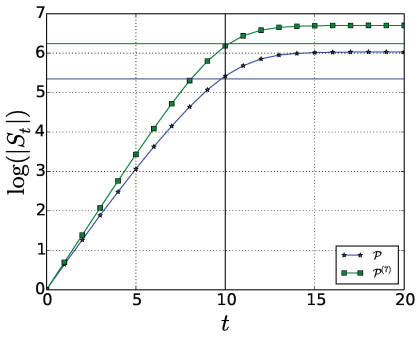


Figure 3.2: Evolution of the size of the active set in logarithmic scale comparing the 2-BRW for $\mathcal{P}$ (blue stars) and $\mathcal{P}^{(7)}$ (green squares) for $N = 10$ (average values obtained over 40 simulations). These simulations show that the BRW referring to $\mathcal{P}^{(k)}$ can explore the whole set of hypercube's vertices simultaneously (the graph underlying by $\mathcal{P}^{(k)}$ is not bipartite). Moreover, it is also the faster one in covering. The dark vertical line corresponds to $t = N$ and the blue and green horizontal lines represent the theoretical percentage of nodes we're supposed to cover in a time $\mathcal{O}(N)$, as proven in Theorem 3.4.9 for $\mathcal{P}$ and 3.4.13 for $\mathcal{P}^{(k)}$, for $\mathcal{P}$ and $\mathcal{P}^{(7)}$ respectively.

Theorem 3.4.10. *Given a 2-BRW-$\mathcal{P}^{(k)}$:*

$$\forall\, t \geq 0 \ \textit{s.t.}\ |S_t| \leq \delta 2^N \quad \Rightarrow \quad \mathbb{E}[|S_{t+1}|] \geq (1+\nu)|S_t|$$

for some constant $\nu > 0$, $\delta \leq 1/2$ and for N big enough.

In order to prove Theorem 3.4.10, we need two preliminary results (Propositions 3.4.11 and 3.4.12) about the characteristics of $\mathcal{P}^{(k)}$.

Proposition 3.4.11. *Let $M_{N,i} := \min_{j,l} \left\{ \left(\mathcal{P}^i\right)_{j,l} \;\middle|\; \left(\mathcal{P}^i\right)_{j,l} \neq 0 \right\}$. We have:*

$$\forall\, N \geq 1, \forall\, i \in \{0, \ldots, N\},\ M_{N,i} = i! \cdot N^{-i}$$

Proof. Due to the regularity of $\mathcal{H}_N$, we have that $\mathcal{P} = \frac{1}{N} A_N$. Then, $M_{N,i} = N^{-i} \min_{j\,l} \left\{ \left(A_N^i\right)_{j,l} \;\middle|\; \left(A_N^i\right)_{j,l} \neq 0 \right\}$ for all $1 \leq i \leq N$, while $\mathcal{P}^0 = A_N^0 = \mathcal{I}_{2^N}$ for all N (and consequently $M_{N,0} = 1$). We have now to prove that $\min_{j,l} \left\{ \left(A_N^i\right)_{j,l} \;\middle|\; \left(A_N^i\right)_{j,l} \neq 0 \right\} = i!$. The proof comes directly by applying the well known result [35]: if A is the adjacency matrix of a graph G, i a positive integer, then the $(j,l)^{\text{th}}$-entry of A^i corresponds to the number of i-length walks between vertex j and l in G.

First, in the case of the N-dimensional hypercube, the Hamming distance between j and l corresponds to the length of the minimal path (*i.e.* walk without

loops) connecting these vertices. Moreover, because of the bipartite structure of $\mathcal{H}_N$ with an i-length walk we can not pass from j to l so that $h(j,l) = i-(2t+1)$, $t \geq 0$ (*i.e.* with an i-length walk we can connect nodes having distance $k \leq i$, k with the same parity as i). It is also clear that if $h(j,l) > i$, then there does not exists any i-length walk from j to l. The minimal number of i-length walks to connect two nodes j, l s.t. $h(j,l) = i - 2t$, $t \geq 0$ corresponds to the case $t = 0$. First, if $h(j,l) = i$ we are counting the number of paths between j and l, and that corresponds to $i!$ (we have just to choose the order of switching of the i different bits). We briefly prove by combinatory arguments that given j_1, l_1, j_2, l_2 s.t. $h(j_1,l_1) = i$ and $h(j_2,l_2) = i-2$, $i \leq N$, then $\left(A^i\right)_{j_1,l_1} \leq \left(A^i\right)_{j_2,l_2}$ *i.e.* $\left(A^i\right)_{j_2,l_2} \geq i!$. In order to cover a distance $i-2$ with an i-length walk we need to change the $i-2$ different bits in i steps. Then the number of possible i-length walks to go from j_2 to l_2 is given by the sum for $k = 0$ to $i-2$ of those walks given by the compositions of:

- a k-length path from j_2 to $j_{2,1}$ s.t. $h(j_{2,1},l_2) = i-2-k$: $\binom{i-2}{k}k!$ possible choices;
- a step from $j_{2,1}$ to $j_{2,2}$ s.t. $h(j_{2,2},l_2) = i-1-k$: $(N-(i-2-k))$ possible choices;
- an $(i-k-1)$-length path from $j_{2,2}$ to l_2: $(i-k-1)!$ possible choices.

Finally:

$$\left(A^i\right)_{j_2,l_2} = \sum_{k=0}^{i-2} \binom{i-2}{k} k!(N-(i-2-k))(i-k-1)! \tag{3.5}$$

We have now to prove that (3.5)$\geq i!$:

$$\sum_{k=0}^{i-2} \binom{i-2}{k} k!(N-(i-2-k))(i-k-1)! = (i-2)! \sum_{k=0}^{i-2} (N-(i-2-k))(i-k-1)$$

And then (3.5)$\geq i! \Leftrightarrow \sum_{k=0}^{i-2}(N-(i-2-k))(i-k-1) \geq i(i-1)$. One can prove by an elementary computation that $\sum_{k=0}^{i-2}(N-(i-2-k))(i-k-1) = \frac{1}{6}i(i-1)(3N-2i+4)$. Consequently the result is proven if $3N-2i+4 \geq 6$:

$$3N-2i+4 \geq N+4 \text{ as } i \leq N, \text{ and } N+4 \geq 6 \text{ since } N \geq 2.$$

□

We give recursively the number of neighbors of each node within our graph:

Proposition 3.4.12. *Let $d_N^{(k)}$ be the number of neighbors of a generic node l (including possibly l) in the graph corresponding to $\mathcal{P}^{(k)}$:* $d_N^{(k)} = \left|\left\{l \,\middle|\, \left(\mathcal{P}^{(k)}\right)_{j,l} \neq 0\right\}\right|$

for all $l \in \{1, \dots, 2^N\}$ *fixed. Then,* $\forall\, N \geq 2$*:*

$$\begin{cases} d_N^{(1)} = N \\ d_N^{(2)} = N + d_{N-1}^{(2)} \\ d_N^{(k)} = d_{N-1}^{(k-1)} + d_{N-1}^{(k)} \text{ for } 3 \leq k \leq N-1 \\ d_N^{(N)} = 2^N \end{cases}$$

Proof. For $k = 1$ and $k = N$ the proof is straightforward. If $k = 1$ we are considering the standard N-dimensional hypercube, and $d_N^{(1)}$ corresponds to the regularity of the graph. If $k = N$, since we allow all possible switch-type mutations, each vertex is connected to itself and any other node within the graph. Therefore, $d_N^{(N)} = 2^N$, the size of the state-space. In order to prove both cases $k = 2$ and $3 \leq k \leq N-1$ we rewrite $d_N^{(k)}$ by using powers of A_N. Indeed, as $\mathcal{P}^{(k)} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{N}A_N\right)^i$, we have: $d_N^{(k)} = \left|\left\{l \,|\, \left(\sum_{i=1}^{k} A_N^i\right)_{j,l} \neq 0\right\}\right|$. Proposition 3.4.12 can now be proven by using the recursive construction of the adjacency matrix of $\mathcal{H}_N$ [51]. □

Proof. *(Theorem 3.4.10)* Let $t \geq 0$ so that $|S_t| \leq \delta 2^N$, for $\delta \leq 1/2$ still unknown. As we did while proving Lemma 3.4.6, our aim is to show:

$$\mathbb{E}[|\mathcal{N}(S_t) - S_{t+1}|] \leq |\mathcal{N}(S_t)| - (1+\nu)|S_t| \tag{3.6}$$

For all vertices $v \in \mathcal{N}(S_t)$, let X_v be the indicator variable:

$$X_v = \begin{cases} 1 & \text{if } v \notin S_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{P}[X_v = 1] = \prod_{j\sim v,\, j\in S_t}\left(1 - \mathcal{P}_{jv}^{(k)}\right)^2 =: p$. We can maximize p as follows:

$$p \leq \prod_{j\sim v,\, j\in S_t}\left(1 - \frac{1}{k}\sum_{i=1}^{k} M_{N,i}\right)^2 = \left(1 - \frac{1}{k}\sum_{i=1}^{k} M_{N,i}\right)^{2d_v},$$

where d_v represents the number of neighbors that v has in S_t $(1 \leq d_v \leq d_N^{(k)})$.

As $\mathbb{E}[X_v] = p$, we have:

$$\mathbb{E}[|\mathcal{N}(S_t) - S_{t+1}|] \leq \sum_{v \in \mathcal{N}(S_t)} \left(1 - \frac{1}{k}\sum_{i=1}^{k} M_{N,i}\right)^{2d_v} \tag{3.7}$$

Denoting by $\Delta := (1/k)\sum_{i=1}^{k} M_{N,i}$, we finally obtain:

$$(3.7) \leq \sum_{v \in \mathcal{N}(S_t)} e^{-2\Delta \cdot d_v} \tag{3.8}$$

Applying Lemma 3.4.7 this expression is maximized if for any v (except possibly for one) $d_v = 1$ or $d_v = d_N^{(k)}$. In particular let us suppose that all d_v are equal to 1 or to $d_N^{(k)}$ and let us denote $R_1 = |\{v \in \mathcal{N}(S_t) \,|\, d_v = 1\}|$ and $R_2 = |\{v \in \mathcal{N}(S_t) \,|\, d_v = d_N^{(k)}\}|$. We demonstrate the statement in this particular case. As $\sum_{v \in \mathcal{N}(S_t)} d_v = d_N^{(k)}|S_t|$:

$$\begin{cases} R_1 + R_2 = |\mathcal{N}(S_t)| \\ \\ R_1 + d_N^{(k)} R_2 = d_N^{(k)}|S_t| \end{cases} \Rightarrow \begin{cases} R_1 = \dfrac{d_N^{(k)}}{d_N^{(k)} - 1}(|\mathcal{N}(S_t)| - |S_t|) \\ \\ R_N = \dfrac{1}{d_N^{(k)} - 1}(d_N^{(k)}|S_t| - |\mathcal{N}(S_t)|) \end{cases}$$

Then we have:

$$\mathbb{E}[|\mathcal{N}(S_t) - S_{t+1}|] \leq \frac{d_N^{(k)}}{d_N^{(k)} - 1}(|\mathcal{N}(S_t)| - |S_t|)e^{-2\Delta} + \frac{1}{d_N^{(k)} - 1}(N|S_t| - |\mathcal{N}(S_t)|)e^{-2\Delta d_N^{(k)}}$$

Equation (3.6) is satisfied if:

$$\frac{d_N^{(k)}}{d_N^{(k)} - 1}(|\mathcal{N}(S_t)| - |S_t|)e^{-2\Delta} + \frac{1}{d_N^{(k)} - 1}(N|S_t| - |\mathcal{N}(S_t)|)e^{-2\Delta d_N^{(k)}} \leq |\mathcal{N}(S_t)| - (1+\nu)|S_t|$$

As the graph we are considering is a $\lambda_{N,2}^{(k)}$-expander graph (where $\lambda_{N,2}^{(k)} = \frac{N-2}{2k}\left(1 - \left(\frac{N-2}{N}\right)^k\right)$ is the second largest eigenvalue of $\mathcal{P}_N^{(k)}$, see Chapter 2), and applying Proposition 3.4.3, the last inequality is true if:

$$\alpha_N^{(k)}\left(1 - \frac{d_N^{(k)} \cdot e^{-2\Delta}}{d_N^{(k)} - 1} + \frac{e^{-2\Delta d_N^{(k)}}}{d_N^{(k)} - 1}\right) + \left(\frac{d_N^{(k)} \cdot e^{-2\Delta}}{d_N^{(k)} - 1} - \frac{d_N^{(k)} \cdot e^{-2\Delta d_N^{(k)}}}{d_N^{(k)} - 1} - 1\right) > 0, \tag{3.9}$$

where $\alpha_N^{(k)} = \dfrac{1}{\delta\left(1-{\lambda_{N,2}^{(k)}}^2\right)+{\lambda_{N,2}^{(k)}}^2}$. That means

$$\delta < \frac{e^{-2\Delta d_N^{(k)}} - d_N^{(k)} e^{-2\Delta} + d_N^{(k)} - 1}{\left(1-{\lambda_{N,2}^{(k)}}^2\right)\left(d_N^{(k)} e^{-2\Delta d_N^{(k)}} - d_N^{(k)} e^{-2\Delta} + d_N^{(k)} - 1\right)} - \frac{{\lambda_{N,2}^{(k)}}^2}{1-{\lambda_{N,2}^{(k)}}^2} := \delta_N^{(k)} \tag{3.10}$$

Finally, let us prove that for fixed $k \geq 2$, $\delta_N^{(k)}$ tends to 1 for N going to infinity. Indeed we have:

- Let $k \geq 2$: $\Delta = \frac{1}{k}\sum_{i=1}^{k} M_{N,i} = \frac{1}{k}\left(\frac{1}{N} + \frac{2}{N^2}\right) + \frac{1}{k}\sum_{i=1}^{k} \frac{i!}{N^i}$. Hence, for $N \to \infty$, $\Delta \sim \mathcal{O}\left(\frac{1}{N}\right)$
- For fixed k, $d_N(k)$ is monotonically increasing:
  - $k = 1 \Rightarrow d_N^{(1)} = N > N-1 = d_{N-1}^{(1)}$;
  - $k = 2 \Rightarrow d_N^{(2)} = N + d_{N-1}^{(2)} > d_{N-1}^{(2)}$;
  - $3 \leq k \leq N-1 \Rightarrow d_N^{(k)} = d_{N-1}^{(k-1)} + d_{N-1}^{(k)} > d_{N-1}^{(k)}$;
- Let $k \geq 2$: $d_N^{(k)} \geq d_N^{(2)}$. By definition: $d_N^{(2)} = N + d_{N-1}^{(2)} = \sum_{i=0}^{N-3}(N-i) + d_2^{(2)} = \frac{N^2+N+2}{2}$. Therefore, for fixed $k \geq 2$, $\Delta d_N^{(k)}$ tends to infinity for $N \to \infty$.

Finally we have, for $k \geq 2$ fixed:

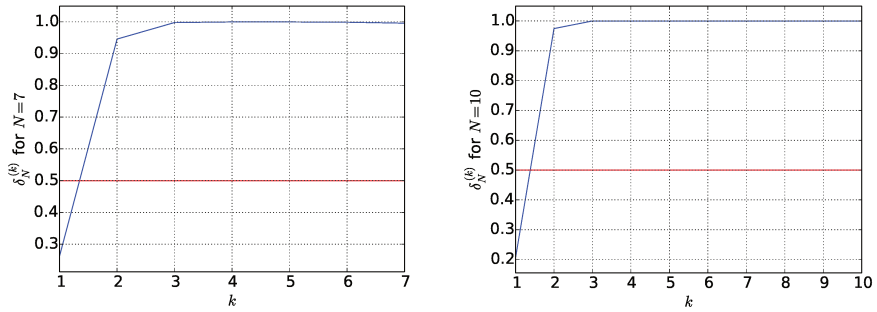
$$\delta_N^{(k)} = \frac{e^{-2\Delta d_N^{(k)}} - d_N^{(k)}\left(e^{-2\Delta}-1\right) - 1}{\left(1-{\lambda_{N,2}^{(k)}}^2\right)\left(d_N^{(k)} e^{-2\Delta d_N^{(k)}} - d_N^{(k)}\left(e^{-2\Delta}-1\right) - 1\right)} - \frac{{\lambda_{N,2}^{(k)}}^2}{1-{\lambda_{N,2}^{(k)}}^2} \to 1 \quad \text{for } N \to \infty$$

Consequently, the strongest condition on δ is the one given by Proposition 3.4.3 (that we need to obtain (3.9)): $\delta \leq 1/2$. Therefore, the 2-BRW-$\mathcal{P}^{(k)}$ grows exponentially until it covers half of the hypercube. The way the rest of the hypercube is covered is not known. □

As we saw in the previous section, we are now able to prove an equivalent of Theorem 3.4.9 for this BRW:

Theorem 3.4.13. *Given a simple 2-BRW-$\mathcal{P}^{(k)}$, there exists a time T such that $T = \mathcal{O}(N)$ and with high probability $|S_T| \geq \delta 2^N$, δ satisfying the hypothesis of Theorem 3.4.10.*

In Figure 3.3 we plot the value of the maximal proportion of vertices of the hypercube we can cover in $\mathcal{O}(N)$ considering a 2-BRW-$\mathcal{P}^{(k)}$. Of course, the case corresponding to $k = 1$ ($\mathcal{P}^{(k)} = \mathcal{P}$) is obtained by Lemma 3.4.6, and we denote $\delta_N^{(1)} := 2^{-r(N)}$ as obtained in (3.4). These simulations show that $\delta_N^{(k)} > 1/2$



(a) $N = 7$ **(b)** $N = 10$

Figure 3.3: The value of $\delta_N^{(k)}$ for $1 \leq k \leq N$ and $N = 7$, 10. The red line represents 1/2. While for the basic mutational model, the process covers a little portion of the hypercube in $\mathcal{O}(N)$, which is smaller for bigger N, allowing more than one mutation at each step, the process can cover at least a half of the graph in a time of the same order.

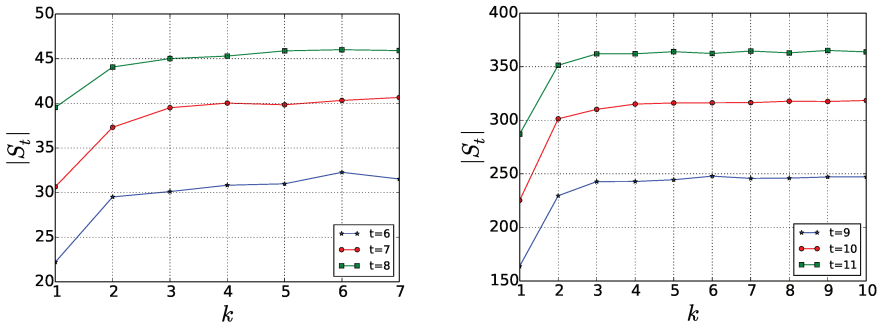
for all $k \geq 2$ even for small N. This result suggests that once we break the bipartiteness by allowing at least two switch-type mutations at each time step, the corresponding BRW invades at least half of the hypercube vertex set in $\mathcal{O}(N)$ (see Section 3.5.3 for a further overlook on this issue).

Remark 24. The definition of δ in Theorems 3.4.10 and 3.4.13 does not depend on $k \geq 2$: even for small values of k, we are able to cover at least a half of the hypercube vertex set in a time $\mathcal{O}(N)$. In Figure 3.4 we simulated the average size of S_t obtained by considering a 2-BRW-$\mathcal{P}^{(k)}$, with $k \in \{1, \dots, N\}$ for different time t. Simulations show that the size of S_t significantly increases passing from $\mathcal{P}$ to $\mathcal{P}^{(2)}$, and it is almost constant for k between 3 and N.

Remark 25. The method applied here does not allow to prove a better covering of $\mathcal{H}_N$ in a time $T = \mathcal{O}(N)$ than the one obtained in Theorem 3.4.13 for matrix $\mathcal{P}^{(k)}$.

Lemma 3.4.14. *Let $\mathcal{M}$ be a transition probability matrix over $\mathcal{H}_N$, representing a d-regular, connected and non bipartite graph. Let λ_2 be the second largest eigenvalue of $\mathcal{M}$. Given a simple 2-BRW-$\mathcal{M}$, there exists a $\delta(\mathcal{M}) := \frac{de^{-2}+d-2}{(1-\lambda_2^2)(d^2e^{-2}+d-2)} - \frac{\lambda_2^2}{1-\lambda_2^2}$ such that in a time $T = \mathcal{O}(N)$ with high probability $|S_T| \geq \delta(\mathcal{M})2^N$. For every such transition probability matrix $\mathcal{M}$, $\delta(\mathcal{M}) \leq 1/2$.*

In other words, applying the method used in Sections 3.4.2 and 3.4.3, the best result we can prove for a 2-BRW-$\mathcal{M}$ is $|S_T| \geq \delta 2^N$, $\delta \leq 1/2$ in a time $T = \mathcal{O}(N)$.



(a) $N = 7$ **(b)** $N = 10$

Figure 3.4: Average size of S_t after $t = N - 1$, $t = N$ and $t = N + 1$ time steps, comparing the 2-BRW-$\mathcal{P}^{(k)}$ with $k \in \{1, \dots, N\}$. Here we plot the average values obtained over 100 simulations.

Proof. The assumptions made over $\mathcal{M}$ and Corollary 3.4.4 imply that $\mathcal{M}$ expresses a (α, δ)-expander graph, with $\delta \leq 1/2$ and $\alpha = \left(\delta(1 - \lambda_2^2) + \lambda_2^2\right)^{-1}$. Let us consider a simple 2-BRW-$\mathcal{M}$. The method used in Sections 3.4.2 and 3.4.3 for $\mathcal{P}$ and $\mathcal{P}^{(k)}$ allows to find $\delta(\mathcal{M})$ (depending on d and λ_2, as given in Lemma 3.4.14) s.t. there exists a time $T = \mathcal{O}(N)$ s.t. with high probability $|S_T| \geq \delta(\mathcal{M})2^N$. However we have a restriction over $\delta(\mathcal{M})$, given by Proposition 3.4.3, which is $\delta(\mathcal{M}) \leq 1/2$. □

Furthermore, a similar threshold shall be explicit for a generic 2-BRW on a bipartite graph defined on the vertices of $\mathcal{H}_N$. At each time step we are observing the evolution of $|S_t|$ over a half part of $\mathcal{H}_N$, hence over a state-space of size 2^{N-1}. Proceeding as above, we obtain the same results with $N-1$ instead of N. Therefore, the best result we can expect in a time $T = \mathcal{O}(N)$ is $|S_T| \geq \delta 2^{N-1}$, $\delta \leq 1/2$.

In Figure 3.5 we test the evolution of the active set size for a 2-BRW corresponding to other transition probability matrices over $\mathcal{H}_N$ which assure good expansion properties. We show the ability of these simple 2-BRWs to cover $\mathcal{H}_N$ for $N = 10$, in logarithmic scale. In particular we consider 5 transition probability matrices:

- $\mathcal{P}$, in blue.
- $\mathcal{P}^7$, in red.
- $\mathcal{P}^{(7)}$, in green.

- $\mathcal{K}_{2^9,2^9}$, in magenta, defined as follows:

$$\mathcal{K}_{2^9,2^9} := \frac{1}{2^9}\begin{pmatrix} \mathbf{0}_{2^9} & \mathcal{J}_{2^9} \\ \mathcal{J}_{2^9} & \mathbf{0}_{2^9} \end{pmatrix},$$

  where $\mathbf{0}_{2^9}$ is a $2^9 \times 2^9$ matrix with all entries 0 and $\mathcal{J}_{2^9}$ is a $2^9 \times 2^9$ matrix with all entries 1. This is the transition probability matrix corresponding to the complete bipartite graph on 2×2^9 vertices.

- $\mathcal{K}_{2^{10}} := \frac{1}{2^{10}-1}(\mathcal{J}_{2^{10}} - \mathcal{I}_{2^{10}})$, in yellow, where $\mathcal{I}_{2^{10}}$ is the 2^{10}-identity matrix. This transition probability matrix corresponds to the complete graph on 2^{10} vertices.

We introduce the complete bipartite graph and the complete graph in order to test the ability in invading the state-space for two transition probability matrices with strong expansion properties. This choice is not biologically motivated and we do not expect that they describe actual mutation rules.

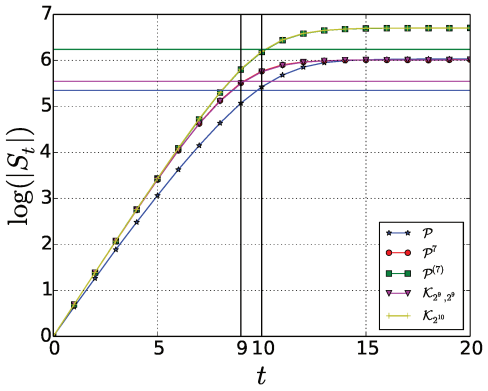


Figure 3.5: Evolution of $|S_t|$ on a log. scale. We compare the 2-BRW for different transition probability matrices and $N = 10$. Average values obtained over 40 simulations are plotted. The green horizontal line corresponds to $\log(2^9)$: we can observe that in a time $t = N$ we do not overtake this threshold, even while considering the complete graph over 2^N vertices. The magenta horizontal line corresponds to $\log(2^8)$: BRWs associated to bipartite graphs do not cover more than this value in a time $t = N - 1$. Finally, the blue horizontal line represents the theoretical size of S_t in a time $\mathcal{O}(N)$ for the simple 2-BRW-$\mathcal{P}$, as obtained in Theorem 3.4.9. The curves corresponding to $\mathcal{P}^{(7)}$ and $\mathcal{K}_{2^{10}}$ are almost overlapping: the expansion properties of both matrices ensure a covering of the same order. The same holds for curves corresponding to $\mathcal{P}^7$ and $\mathcal{K}_{2^9,2^9}$, which characterize bipartite graphs over $\{0,1\}^N$ with an appreciable vertex expansion.

We observe that although for the complete graph, which has the best expansion property, in a time $t = N$ we can cover about a half of the state-space, as with the simple 2-BRW-$\mathcal{P}^{(k)}$. Even for small $t > 0$, the process corresponding to $\mathcal{P}^{(7)}$ is faster when compared to the 2-BRW-$\mathcal{P}^7$. It is interesting to compare this fact with a phenomenon observed in Chapter 2 where we have investigated the typical time-scale of the exploration of $\mathcal{H}_N$ considering RWs without branching. We have demonstrated that for $k > 2$, $\mathcal{P}^k$ optimizes the hitting time to reach a certain configuration, if compared to $\mathcal{P}^{(k)}$. When we take into account the *branching* equivalent of these RWs, the exploration of $\mathcal{H}_N$ is more efficient using $\mathcal{P}^{(k)}$ as transition probability matrix instead of $\mathcal{P}^k$. That suggests that once added a branching process, the oscillations due to bipartiteness are of greater amplitude and forbid a quick covering even for small t.

3.5 Extensions of the model

In this Section we set some variants of the model considered so far, in which we take into account the multiplicity of each vertex. This adds a further building block to our model. Indeed, taking into account the number of particles lying on the same vertex allows to consider the size of the effective population and not only how many different BCR configurations are expressed at a certain time. Moreover, considering multiplicity also allows us to have a better chance of making $|S_t|$ grow faster, where $|S_t|$ represents here the number of vertices of $\{0,1\}^N$ on which at least one particle lies. In Section 3.5.1 we consider BRWs with multiplicity and fixed number of offsprings c at each time step. Then, in Section 3.5.2, we give to each individual a probability p to divide: we observe the impact of division on the limiting distribution. Finally, in Section 3.5.3, we observe and discuss, through computer simulations, a model for which the division rate depends on affinity.

3.5.1 c-BRW with multiplicity

At time $t \geq 0$ we have exactly c^t particles, as there is no death nor selection. We consider the distribution of these c^t particles within $\mathcal{H}_N$. In order to do so, we define the Markov process $(X_t^i)_{t\geq 0}$, where for all $i \in \{1, \dots, 2^N\}$, X_t^i corresponds to the number of particles lying on the i^{th} node at time t. Proposition 3.5.1 is given in the more general case of a c-BRW with multiplicity on a given d-regular graph: the case we are interested in is an application with $c = 2$ and $d = N$.

Proposition 3.5.1. *Given a c-BRW with multiplicity on a d-regular graph, for*

all $s \geq 0$*:*

$$\mathbb{P}\left[X_t^i = s \,\middle|\, \sum_{j\sim i} X_{t-1}^j = n\right] = \begin{cases} \binom{cn}{s} \dfrac{(d-1)^{cn-s}}{d^{cn}} & \text{if } s \leq cn, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Proof. We show that conditioning on $\sum_{j\sim i} X_{t-1}^j = n$, X_t^i follows a binomial distribution $\mathcal{B}\left(cn, \frac{1}{d}\right)$. For all $j \sim i$ let us define the random variables $\mathbf{Z}_{l,r}^j$, where $\mathbf{Z}_{l,r}^j$ corresponds to the vertex chosen by the l^{th}-particle lying on j in its r^{th}-trial, with $j \in S_{t-1} \cap \mathcal{N}(\{i\})$, $1 \leq l \leq X_{t-1}^j$ and $1 \leq r \leq c$. Then we have:

$$\mathbb{P}[\mathbf{Z}_{l,r}^j = i] = 1/d \quad \forall\, j, l, r$$

At each trial of each particle lying on a vertex j, the probability of success (*i.e.* going on vertex i) is exactly $1/d$ and the probability of failure is $1 - 1/d$. Moreover, there are exactly cn independent and identically distributed trials. The result follows. □

In particular:

Proposition 3.5.2. *Given a c-BRW with multiplicity on the complete graph on d vertices* $\mathcal{K}_d$*, the distribution of* X_t^i *given* $X_{t-1}^i = s'$ *is a binomial distribution with parameters* $c^t - cs'$ *and* $\frac{1}{d-1}$*,* i.e. *for all* $s \geq 0$*:*

$$\mathbb{P}[X_t^i = s \,|\, X_{t-1}^i = s'] = \begin{cases} \binom{c^t - cs'}{s} \left(\dfrac{1}{d-1}\right)^s \left(1 - \dfrac{1}{d-1}\right)^{c^t - cs' - s} & \text{if } s \leq c^t - cs', \\ \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 3.5.2 shows that, for a complete graph on N vertices, the probability of having s particles at time t on the i^{th}-node depends on the number of particles laying on i at time $t-1$.

Proof. In this particular case, i is connected to all nodes of the graph, except itself. Therefore each one of the c^t particles produced at time t has a probability $1/(d-1)$ to go to i: we have to remove the particles that will leave from i, and this is exactly cs'. □

We establish another property of the c-BRW with multiplicity: the asymptotic distribution of the c^t individuals for $t \to \infty$. This concludes this section.

Lemma 3.5.3. *Let $\mathcal{M}$ be the transition probability matrix corresponding to a finite connected graph $G = (V, E)$, $\mathbf{m}$ its stationary distribution. Let us suppose $\mathcal{M}$ aperiodic, and let us consider a c-BRW-$\mathcal{M}$ starting from a generic initial distribution $\mathbf{p}$. Therefore:*

$$\forall\, i \in V,\ \frac{X_t^i}{c^t} \to \mathbf{m}_i \text{ in probability, for } t \to \infty.$$

Proof. The position of each of the c^t individuals at time t corresponds to the position reached by a RW with $\mathcal{M}$ as transition probability matrix, starting from the initial distribution $\mathbf{p}$ and independently form others individuals. In other words, at time t we are considering the position of c^t parallel RWs-$\mathcal{M}$ starting from the same initial distribution. For all $j \in \{1, \dots, c^t\}$, let $(X_{j,t})_{t\geq 0}$ i.i.d RWs with transition probability matrix $\mathcal{M}$ and starting from the initial distribution $\mathbf{p}$. By hypothesis, for all $i \in V$, $\mathbb{P}(X_{j,t} = i) \to \mathbf{m}_i$ for $t \to \infty$. The result follows since convergence in law to a constant implies convergence in probability. □

Remark 26. Numerically, we compare the average size of S_t for $t = N = 10$ for the simple 2-BRW-$\mathcal{P}$, the simple 2-BRW-$\mathcal{K}_{2^9,2^9}$ and the 2-BRW-$\mathcal{P}$ with multiplicity. Table 3.2 below shows the average values obtained over 100 simulations. As expected, the 2-BRW-$\mathcal{P}$ with multiplicity is faster than the simple 2-BRW-$\mathcal{P}$ because of the number of particles within the population, which is not affected nor by selection or death, neither by coalescence. At each step, each particle can divide and colonize a new vertex of the hypercube, therefore we have a better chance to cover faster a half of the state-space (we recall that $\mathcal{P}$ is a bipartite graph). Moreover, we can observe that the simple 2-BRW-$\mathcal{K}_{2^9,2^9}$ is faster than the simple 2-BRW-$\mathcal{P}$. Indeed $\mathcal{K}_{2^9,2^9}$ has better expander properties, and thus the BRW invades more efficiently the state-space as noticed in Sec. 3.4.

Table 3.2: Average size of S_t after 10 time steps, comparing the simple 2-BRW-$\mathcal{P}$, the simple 2-BRW-$\mathcal{K}_{2^9,2^9}$ and the 2-BRW-$\mathcal{P}$ with multiplicity. We denote by $\widehat{|S_{10}|}_n$ the average value obtained over n simulations and by $\widehat{\sigma}_n$ its corresponding estimated standard deviation.

Model	$\boldsymbol{N}$	$\boldsymbol{n}$	$\widehat{\boldsymbol{\lvert S_{10}\rvert}}_{\boldsymbol{n}}$	$\frac{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{n}}}{\sqrt{\boldsymbol{n}}}$
Simple 2-BRW-$\mathcal{P}$	10	100	222.36	3.376
Simple 2-BRW-$\mathcal{K}_{2^9,2^9}$	10	100	318.04	1.231
2-BRW-$\mathcal{P}$ with multiplicity	10	100	398.42	0.972

3.5.2 Limiting distribution for the BRW-$\mathcal{P}$ with multiplicity and division rate p.

Lemma 3.5.3 can not be applied to the 2-BRW-$\mathcal{P}$ with multiplicity. Indeed, the bipartite structure of the corresponding graph prevents the convergence through the stationary distribution, *i.e.* the homogeneous probability distribution. We denote the homogeneous probability distribution by $\boldsymbol{\pi}$ (Chapter 2). We can overcome this problem by considering a BRW-$\mathcal{P}$ with multiplicity and with a non constant division rate p.

Definition 3.14. Let us fix $p \in]0,1[$. The process starts with a single individual located on an arbitrary node of $\mathcal{H}_N$. Each time step, a particle lying on a certain node $\mathbf{x}_i$ of $\mathcal{H}_N$ gives rise to 2 daughter cells and die with probability p. With probability $1-p$, it remains in the population for the next time step. When division occurs, each newborn particle choses a neighbor node according to matrix $\mathcal{P}$, independently and with replacement, and move on it.

The introduction of a division rate has two immediate consequences. First, it slows down the population's growth. In order to evaluate the expected number of individuals at time t, we consider a generic Galton-Watson process ([59], Chapter I).

Proposition 3.5.4. *Let Z_t be the random variable (rv) describing the number of individuals at generation t starting from $Z_0 = 1$ individual. We assume that each individual divides indepently from the others and from previous generations. Let $\mathbf{p} := (p_k, k = 0, 1, 2, \dots)$ be a probability distribution s.t. p_k gives the probability of having k offsprings in the next generation. At each time step, given $Z_t = k$, Z_{t+1} behaves as k independent copies of Z_1. Therefore: $\mathbb{E}(Z_t) = (\mathbb{E}(Z_1))^t$.*

In our specific case we have:

- $p_1 = 1 - p$
- $p_2 = p$
- $p_k = 0$ for all $k \neq 1, 2$

Which gives:

$$\mathbb{E}(Z_t) = (1+p)^t < 2^t \text{ as } p < 1. \tag{3.11}$$

Remark 27. One can observe that $Z_t = \sum_{i=1}^{2^N} X_t^i$, where X_t^i describes the number of individuals lying on vertex i at time t.

The addition of the parameter p overcomes issues related to the bipartite structure of the graph, discussed in Section 3.3.

Lemma 3.5.5. *Let us consider a BRW with multiplicity on a finite connected bipartite graph G_b. Let $\mathbf{p} := (p_k, k = 0, 1, 2, \dots)$ be the probability distribution of the number of offsprings of each individuals for the next generation, s.t. $p_1 > 0$ and $p_0 + p_1 < 1$. Then there exists a time $t \geq 0$ and two nodes $\mathbf{x}_1$, $\mathbf{x}_2$ s.t. $\mathbf{x}_1 \sim \mathbf{x}_2$ and $\mathbf{x}_1, \mathbf{x}_2 \in S_t$.*

Lemma 3.5.5 implies that for this type of BRWs, independently from the bipartite structure of $G_b = (V, E)$, there exists a time $t > 0$ s.t. $S_t = V$ (see Section 3.3.2).

Proof. Let $0 < T < \infty$ s.t. $\mathbf{x}_i \in S_T$ (T exists as G_b is finite and connected). As $p_0 + p_1 < 1$, $\exists\, k \geq 2$ s.t. $p_k > 0$. Then with probability p_k, $\exists\, \mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,k} \in \mathcal{N}(\{\mathbf{x}_i\})$ s.t. $\{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,k}\} \in S_{T+1}$. As $p_1 > 0$, with positive probability at least one among these k vertices does not divide: let $\overline{k} \in \{1, \dots, k\}$ s.t. $\mathbf{x}_{i,\overline{k}} \in S_{T+2}$. Moreover w.p.p. one among $\{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,k}\} \setminus \{\mathbf{x}_{i,\overline{k}}\}$ divides and w.p.p. one of its offsprings migrates to $\mathbf{x}_i$. Therefore, w.p.p. $\{\mathbf{x}_{i,\overline{k}}, \mathbf{x}_i\} \in S_{T+2}$, and $\mathbf{x}_{i,\overline{k}} \sim \mathbf{x}_i$. □

We give an equivalent of Lemma 3.5.3 for BRWs characterized by Definition 3.14.

Lemma 3.5.6. *Let $\mathcal{M}$ be the transition probability matrix corresponding to a finite connected graph $G = (V, E)$, $\mathbf{m}$ its stationary distribution. Let us consider a BRW-$\mathcal{M}$ with multiplicity starting from a generic initial distribution. Let $\mathbf{p} := (p_k, k = 0, 1, 2, \dots)$ be the probability distribution of the number of offsprings of each individual for the next generation, with $p_1 > 0$ and $p_0 + p_1 < 1$. We denote by Z_t the r.v. describing the population size at generation t (starting from $Z_0 = 1$). For all $i \in V$ let X_t^i be the r.v. describing the number of individuals lying on vertex i at time t. Therefore:*

$$\forall\, i \in V,\ \frac{X_t^i}{(\mathbb{E}(Z_1))^t} \to \mathbf{m}_i \text{ in probability for } t \to \infty.$$

Proof. The proof is the same as for Lemma 3.5.3. In this case, we do not need the hypothesis of aperiodicity of $\mathcal{M}$ as the problem of an eventual periodicity is overcome by the addition of the distribution of the number of offsprings $\mathbf{p}$, as shown in Lemma 3.5.5. □

Lemma 3.5.6 allows to prove:

Corollary 3.5.7. *Let us consider a BRW-$\mathcal{P}$ with multiplicity and division rate $p \in]0,1[$.*

$$\forall\, i \in \{0,1\}^N,\ \frac{X_t^i}{(1+p)^t} \to \frac{1}{2^N} \text{ in probability for } t \to \infty.$$

Proof. We have already determined $\mathbb{E}(Z_t)$ corresponding to the BRW-$\mathcal{P}$ with multiplicity and division rate p (cf. (3.11)). Therefore, in order to prove Corollary 3.5.7 we have just to observe that the stationary distribution for $\mathcal{P}$ is the homogeneous probability distribution on $\{0,1\}^N$. Then the result follows applying Lemma 3.5.6. □

Remark 28. In Chapter 2 we overcame the problem of the bipartiteness of the graph underlined by $\mathcal{P}$ by adding N loops at each node. That corresponds to take into account matrix $\mathcal{P}_L := \frac{1}{2}(\mathcal{P}+\mathcal{I}_{2^N})$ instead of $\mathcal{P}$. Considering a BRW-$\mathcal{P}$ with multiplicity and division rate $p = 1/2$ is equivalent to consider a 2-BRW-$\mathcal{P}_L$ with multiplicity, but with coalescence of offsprings which decide to remain in place. The only difference is the size of the population at time t, which is 2^t in the case of a 2-BRW-$\mathcal{P}$ with multiplicity and is expected to be $(3/2)^t$ in the other case. The choice of $\mathcal{P}_L$ as transition probability matrix has also biological motivations. Indeed division of B-cells in GCs is asymmetric [94, 15]: only one between the two daughter cells has a mutated trait.

Remark 29. More generally, let us consider a transition probability matrix $\mathcal{M}$ on a graph $G = (V, E)$, with $|V| = n$. We can see a BRW-$\mathcal{M}$ with multiplicity and division rate p as a 2-BRW-$\mathcal{M}_p$ with multiplicity, where $\mathcal{M}_p := p\mathcal{M} + (1-p)\mathcal{I}_n$. Of course, we need to take the same caution as in Remark 28 about the number of individuals at time t.

3.5.3 BRW-$\mathcal{P}$ with multiplicity and affinity dependent division

In previous sections, the limiting distribution of traits (with or without division rate) only depends on the stationary distribution of the considered transition probability matrix. In particular, if the stationary distribution is homogeneous, for t big enough all individuals are uniformly distributed over the state-space. From a biological point of view, it does not seem so efficient to explore all the state-space. It will be rather more interesting to drive mutations through the region of the state-space with greater affinity for the target trait. We can therefore propose a model in which we introduce a division rate dependent on the affinity of the cell.

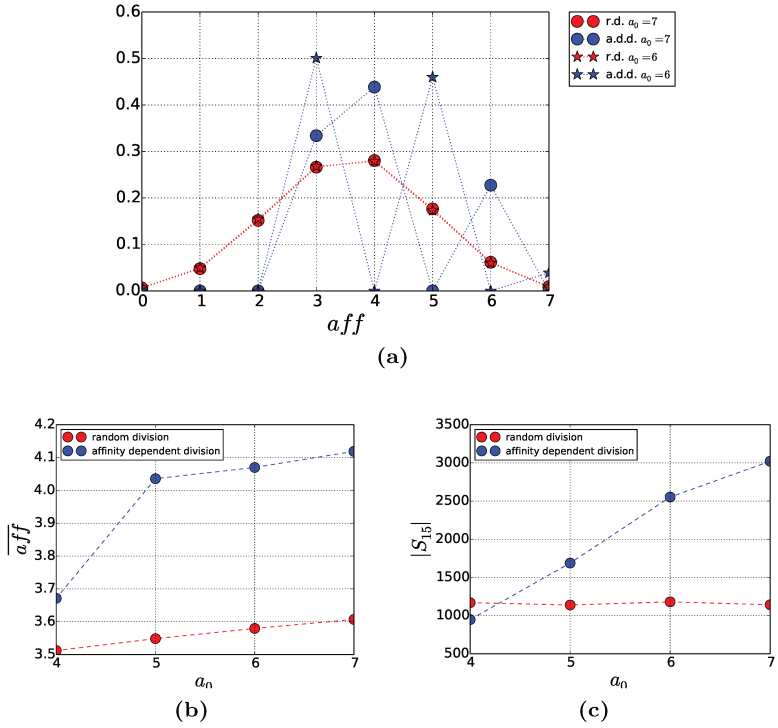


Figure 3.6: Simulations of the BRW-$\mathcal{P}$ with multiplicity, comparing a model with division rate $p = 0.6$ (in red) and a model with affinity dependent division (in blue). In this last case individuals having affinity at least 4 with the target vertex divide and mutate accordingly to matrix $\mathcal{P}$, they remain unchanged in the population otherwise. (a) Distribution of the affinity to the antigen after 15 time steps, starting from initial affinity 7 (circles) and 6 (stars) respectively. (b) Dependance of the average affinity (after 15 time steps) on the initial affinity a_0. (c) Dependance of the final population size (after 15 time steps) on the initial affinity a_0.

Formally, $\forall\, \mathbf{x}_i \in \mathcal{H}_N$, let $p_d(\mathbf{x}_i)$ be the probability of division of an individual lying on vertex $\mathbf{x}_i$. We can define an increasing function f s.t. $p_d(\mathbf{x}_i) = f\left(\text{aff}(\mathbf{x}_i, \overline{\mathbf{x}})\right)$, where $\text{aff}(\mathbf{x}_i, \overline{\mathbf{x}}) = N - h(\mathbf{x}_i, \overline{\mathbf{x}})$ is the affinity of $\mathbf{x}_i$ with respect to the target trait $\overline{\mathbf{x}}$ (Definition 3.9), and h return the Hamming distance. The aim is to be able to privilege those individuals having better fitness. This choice has biological motivations. Indeed, recent evidence shows that during GC reaction the acquisition of highest affinity for the presented antigen regulates proliferation and diversification of B-cells [55]. In Figure 3.6 we compare a model of BRW-$\mathcal{P}$ with multiplicity and division rate $p = 0.6$ with a model of BRW-$\mathcal{P}$ with multiplicity and affinity dependent division. In this case, we choose a very

simple function for the division rate, defined $\forall\, \mathbf{x}_i \in \mathcal{H}_N$, as follows:

$$p_d(\mathbf{x}_i) = \begin{cases} 0 & \text{if aff}(\mathbf{x}_i, \overline{\mathbf{x}}) < N - \overline{h}_s \\ 1 & \text{if aff}(\mathbf{x}_i, \overline{\mathbf{x}}) \geq N - \overline{h}_s \end{cases} \tag{3.12}$$

We plot results obtained for $N = 7$ and $\overline{h}_s = 3$: all individuals having affinity at least 4 with the target trait divide and mutate accordingly to matrix $\mathcal{P}$, they remain unchanged in the population otherwise.

In Figure 3.6 (a) we represent the final distribution of the affinity of traits within the population after 15 time steps. As expected, the distribution corresponding to the first model is binomial and does not depend on the initial Hamming distance. Indeed, from Corollary 3.5.7 we know that the distribution of traits is uniform on $\{0,1\}^N$. We have just to remark that in $\{0,1\}^N$ there are exactly $\binom{N}{h}$ nodes having Hamming distance h from a given vertex, $0 \leq h \leq N$: this determines the proportion of individuals having a given affinity after 15 time steps. The support of the distribution at time step 15 for the second model corresponds to vertices having affinity 3, 4 or 6 (resp. 3, 5, 7) with the target trait for an initial affinity $a_0 = 7$, (resp. $a_0 = 6$). Indeed, as $a_0 \geq 4$, the total population can be divided in two subpopulations. The sub-population whose affinity with the target trait is greater than 4 follows a standard 2-BRW-$\mathcal{P}$ with multiplicity. Therefore, we can observe the effects of the bipartiteness of the graph: only traits whose affinity has the same parity as a_0 are expressed at even time steps. On the contrary, at odd time steps only vertices with affinity having the opposite parity as a_0 are expressed. The other sub-population is composed by those individuals that after an unfavorable mutation obtain a trait having affinity exactly 3. They remain unchanged for all further time steps, as they can not divide nor die. Therefore, through further time steps, individuals with affinity 3 can only continue to accumulate. This is due to the definition of $p_d(\mathbf{x}_i)$ as a step function.

Figure 3.6 (b) shows the average affinity of the population after 15 time steps. We can see that for the BRW-$\mathcal{P}$ with division rate 0.6 this depends very lightly from the initial affinity, while, as expected, the initial affinity strongly influences the final one if we allow only individuals having affinity greater than 3 to divide. Finally in Figure 3.6 (c) we see the size of the population after 15 time steps. Again, in the case of random division with rate 0.6, the initial affinity does not affect the final population size, which is always approximately $1.6^{15} \simeq 1152.92$.

3.6 Conclusions and perspectives

In this Chapter, we introduce and study BRWs on binary strings, modeling the evolution of cells in a mutation-division process. The edge set (or graph) associated to $\mathcal{H}_N := \{0,1\}^N$, hence the corresponding transition probability matrix, reflects mutations allowed during the evolutionary process. Graph's characteristics determine the behavior of the BRW, *e.g.* its ability in covering $\mathcal{H}_N$ or the limiting distribution of the traits, as shown in Sections 3.4 and 3.5.

We particularly focus on the expander property of the graphs when giving quantitative results about the expected portion of $\mathcal{H}_N$ covered in $\mathcal{O}(N)$. We observe that strong expansion properties enable a faster invasion of the state-space. From a biological point of view, this property is significant since it ensures that starting from one or a few B-cells, the GC can produce, hence test a huge variety of BCRs against the target antigen. Indeed, GCs seem to be oligoclonal [81, 88], which means that they develop from very few initial naive B-cells (three, on average). Therefore, starting from a single clonal population, it is of interest to understand how a B-cell population invades the BCR state-space.

For this reason, in Section 3.4, we consider the state-space $\mathcal{H}_N$ of every possible N-length strings (modeling B-cell traits), and compare the ability of different mutation rules in colonizing $\mathcal{H}_N$ in a time $\mathcal{O}(N)$. We develop upon a method used in [43] to evaluate partial cover times on expander graphs. Nevertheless, our approach differs from [43]. Indeed, we fix the state-space and the main question becomes: how many nodes we are able to activate in a time $\mathcal{O}(N)$ for a given graph? In particular, we observe that while matrix $\mathcal{P}$, which denotes the structure of the standard N-dimensional hypercube, can cover a quite small portion of $\mathcal{H}_N$ in a time $\mathcal{O}(N)$, the mutation rule $\mathcal{P}^{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathcal{P}^i$ leads to a significantly bigger expansion which does not strongly depends on k, for values of k greater than 2.

In Section 3.4, we show that if we simply consider the expansion properties of the structure built over $\mathcal{H}_N$, the covering in $\mathcal{O}(N)$ is limited at a half the state-space (Lemma 3.4.14). This favors the hypothesis that the expansion property is not enough to insure a quick covering of a large portion of the state-space: considering self-avoiding BRWs on connected graphs could be more efficient, although these are not necessarily good expanders. On the other hand, from a biological point of view, it may not be so efficient to explore the whole state-space, but rather to steer mutations toward a specific region of the state-space with the best affinity. Indeed, the production of new clones has a cost in terms

of time and energy, therefore it does not make sense to produce a huge variety of cells with any possible fitness with the presented antigen. Models considered in this Chapter share this drawback: even if a bigger portion of possible traits is expressed in a time $\mathcal{O}(N)$, we can not say much about their average fitness.

We can propose many possible solutions to this problem. We can for example privilege individuals with good fitness by considering a model with affinity dependent division, as discussed in Section 3.5.3. Another possibility is to consider transition probability matrices whose stationary distribution is concentrated on a specific region of the state-space containing the fittest traits. Indeed, as we observe in Section 3.5.1, given this hypothesis than the distribution of traits for a 2-BRW with multiplicity only depends on the stationary distribution of the transition probability matrix under consideration. In this case the problem is: does this matrix accounts for realistic mutations? Another way to drive mutations towards a specific region of the state-space is, of course, the introduction of a selection mechanism, which we study in Chapter 4.

Chapter 4

Multi-type Galton-Watson processes with affinity-dependent selection applied to antibody affinity maturation

Summary We analyze the interactions between division, mutation and selection in a simplified evolutionary model, assuming that the population observed can be classified into fitness levels. The construction of our mathematical framework is motivated by the modeling of antibody affinity maturation of B-cells in Germinal Centers during an immune response. This is a key process in adaptive immunity leading to the production of high affinity antibodies against a presented antigen. Our aim is to understand how the different biological parameters affect the system's functionality. We identify the existence of an optimal value of the selection rate, able to maximize the number of selected B-cells for a given generation.

4.1 Introduction

Antibody Affinity Maturation (AAM) takes place in Germinal Centers (GCs), specialized micro-environnements which form in the peripheral lymphoid organs upon infection or immunization [137, 36]. GCs are seeded by ten to hundreds distinct B-cells [132], activated after the encounter with an antigen, which ini-

tially undergo a phase of intense proliferation [36]. Then, AAM is achieved thanks to multiple rounds of division, Somatic Hypermutation (SHM) of the B-cell receptor proteins, and subsequent selection of B-cells with improved ability of antigen-binding [89]. B-cells which successfully complete the GC reaction output as memory B-cells or plasma cells [138, 36]. Indirect evidences suggest that only B-cells exceeding a certain threshold of antigen-affinity differentiate into plasma cells [109]. The efficiency of GCs is assured by the contribution of other immune molecules, for instance Follicular Dendritic Cells (FDCs) and follicular helper T-cells (Tfh). Nowadays the key dynamics of GCs are well characterized [89, 36, 55, 132]. Despite this there are still mechanisms which remain unclear, such as the dynamics of clonal competition of B-cells, hence how the selection acts. In recent years a number of mathematical models of the GC reaction appear to investigate these questions, such as [96, 142], where the authors have developed agent-based models, mostly analyzed through extensive numerical simulations, or [148] where the authors have established a coarse-grained model, looking for optimal values of *e.g.* the selection strength and the initial B-cell fitness maximizing the affinity improvement.

Our aim in this Chapter is to contribute to the mathematical foundations of adaptive immunity by introducing and study a simplified evolutionary model inspired by AAM, including division, mutation, affinity-dependent selection and death. We focus on interactions between these mechanisms, identify and analyze the parameters which mostly influence the system functionality, through a rigorous mathematical analysis. This research is motivated by important biotechnological applications. The fundamental understanding of the evolutionary mechanisms involved in AAM have been inspiring many methods for the synthetic production of specific antibodies for drugs, vaccines or cancer immunotherapy [6, 79, 122]. Indeed, this production process involves the selection of high affinity peptides and requires smart methods to generate an appropriate diversity [34]. Beyond biomedical motivations, the study of this learning process has also given rise in recent years to a new class of bio-inspired algorithms [30, 107, 134], mainly addressed to solve optimization and learning problems.

We consider a model in which B-cells are classified into $N+1$ affinity classes with respect to a presented antigen, N being an integer big enough to opportunely describe the possible fitness levels of a B-cell with respect to a specific antigen [143, 146]. A B-cell is able to increase its fitness thanks to SHMs of its receptors: only about 20% of all mutations are estimated to be affinity-affecting mutations [121, 123]. By conveniently define a transition probability matrix, we can characterize the probability that a B-cell belonging to a given affinity class

passes to another one by mutating its receptors thanks to SHMs. Therefore we define a selection mechanism which acts on B-cells differently depending on their fitness. We mainly focus on a model of positive and negative selection in which B-cells submitted to selection either die or exit the GC as output cells, according to the strength of their affinity with the antigen. Hence, in this case, no recycling mechanism is taken into account. Nevertheless the framework we set is very easy to manipulate: we can define and study other kinds of affinity-dependent selection mechanisms, and eventually include recycling mechanisms, which have been demonstrated to play an important role in AAM [139]. We demonstrate that independently from the transition probability matrix defining the mutational mechanism and the affinity threshold chosen for positive selection, the optimal selection rate maximizing the number of output cells for the t^{th} generation is $1/t$ (Corollary 4.3.11).

From a mathematical point of view, we study a class of multi-types Galton-Watson (GW) processes in which, by considering dead and selected B-cells as two distinct types, we are able to formalize the evolution of a population submitted to an affinity-dependent selection mechanism. To our knowledge, the problem of affinity-dependent selection in GW processes has not been deeply investigated so far.

In Section 4.2 we define the main model analyzed in this Chapter. We give as well some definitions that we will use in next sections, such as affinity classes and mutational model. Section 4.3 contains the main mathematical results. A conveniently use of a multi-type GW process allows to study the evolution of both GC and output cells during time. Proposition 4.3.9 collects the formulas which describe the expected size and average affinity of both populations. Moreover, in Section 4.3.3 we determine the optimal value of the selection rate which maximizes the expected number of selected B-cells at time t. This value is $1/t$ independently from all other parameters. We conclude Section 4.3 with some numerical simulations. In Section 4.4 we define two possible variants of the model described in previous sections, and provide some mathematical results and numerical simulations as well. Section 4.4 evidences how the mathematical tools used in Section 4.3 easily apply to define other affinity-dependent selection models. Finally, in Section 4.5 we discuss our modeling assumptions and give possible extensions and limitations of our mathematical model.

4.2 Definitions and modeling assumptions

This section provides the mathematical framework of this Chapter. Let us suppose that given an antigen target cell $\overline{\mathbf{x}}$, all B-cell traits can be divided in exactly $N+1$ distinct affinity classes, named 0 to N.

Definition 4.1. Let $\overline{\mathbf{x}}$ be the antigen target trait. Given a B-cell trait $\mathbf{x}$, we denote by $a_{\overline{\mathbf{x}}}(\mathbf{x})$ the affinity class it belongs to with respect to $\overline{\mathbf{x}}$, $a_{\overline{\mathbf{x}}}(\mathbf{x}) \in \{0, \dots, N\}$. The maximal affinity corresponds to the first class, 0, and the minimal one to N.

Definition 4.2. Let $\mathbf{x}$ be a B-cell trait belonging to the affinity class $a_{\overline{\mathbf{x}}}(\mathbf{x})$ with respect to $\overline{\mathbf{x}}$. We say that its affinity with $\overline{\mathbf{x}}$ is given by:

$$\text{aff}(\mathbf{x}, \overline{\mathbf{x}}) = N - a_{\overline{\mathbf{x}}}(\mathbf{x})$$

Of course, this is not the only possible choice of affinity. Typically affinity is represented as a Gaussian function [142, 96], having as argument the distance between the B-cell trait and the antigen in the shape space of possible traits. In our model this distance corresponds to the index of the affinity class the B-cell belongs to. Nevertheless the choice of the affinity function does not affect our model.

During the GC reaction B-cells are submitted to random mutations. This implies switches from one affinity class to another with a given probability.

Definition 4.3. Let $(\mathbf{X}_t)_{t\geq 0}$ be a RW on the state-space of B-cell traits describing a pure mutational process of a B-cell during the GC reaction. We denote by $\mathcal{Q}_N = (q_{ij})_{0\leq i,j\leq N}$ the transition probability matrix over $\{0, \dots, N\}$ which gives the probability of passing from an affinity class to another during the given mutational model. For all $0 \leq i, j \leq N$:

$$q_{ij} = \mathbb{P}(a_{\overline{\mathbf{x}}}(\mathbf{X}_{t+1}) = j \,|\, a_{\overline{\mathbf{x}}}(\mathbf{X}_t) = i)$$

The main model we study in this Chapter is defined as follows:

Definition 4.4. The process starts with $z_0 \geq 1$ B-cells entering the GC, belonging to some affinity classes in $\{0, \dots, N\}$. In case they are all identical, we denote by a_0 the affinity class they belong to, with respect to the antigen target cell $\overline{\mathbf{x}}$. At each time step, each GC B-cell can die with a given rate r_d. If not, each B-cell can divide with rate r_{div}: each daughter cell may have a mutated trait, according to the mutational rule allowed. Hence it eventually belongs to a different affinity class than its mother cell. Clearly, it also happens that a B-cell

stays in the GC without neither die nor divide. Finally, with rate r_s each B-cell can be submitted to selection, which is made according to its affinity with $\overline{\mathbf{x}}$. A threshold $\overline{a}_s$ is fixed: if the B-cell belongs to an affinity class with index greater than $\overline{a}_s$, the B-cell dies. Otherwise, the B-cell exits the GC pool and reaches the selected pool. Therefore, for any GC B-cell and at any generation, we have:

- $\mathbb{P}(\text{death}) = r_d$
- $\mathbb{P}(\text{division}) = r_{div}$
- $\mathbb{P}(\text{selection}) = r_s$

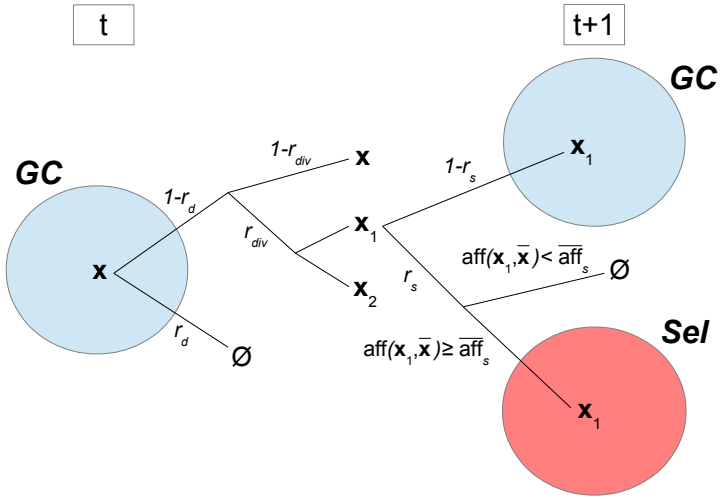


Figure 4.1: Schematic representation of model described by Definition 4.4. Here we denote by $\overline{\text{aff}}_s$ the fitness corresponding to the affinity class $\overline{a}_s$.

The mutation rule reflects the edge set associated to the state-space $\{0, \dots, N\}$: this is given by a transition probability matrix.

Once the GC reaction is fully established ($\sim$ day 7 after immunization), it is polarized into two compartments, named Dark Zone (DZ) and Light Zone (LZ) respectively. The DZ is characterized by densly packed dividing B-cells, while the LZ is less densely populated and contains FDCs and Tfh cells. This is the preferential zone for selection [36]. The transition of B-cells from the DZ to the LZ seems to be determined by a timed cellular program: over a 6 hours period about 50% of DZ B-cells transit to the LZ, where they compete for positive

selection signaling [14, 136]. In our simplified mathematical model we do not take into account any spatial factor and in a single time step a GC B-cell can eventually undergo both division (with mutation) and selection. Hence the time unity has to be chosen big enough to take into account both mechanisms.

In Chapters 2 and 3 we have modeled B-cells and antigens as N-length binary strings, hence their traits correspond to elements of $\{0,1\}^N$. In this context we have characterized affinity thanks to the Hamming distance between B-cell and antigen representing strings.

Definition 4.5. For all $\mathbf{x} = (x_1, \dots, x_N)$, $\mathbf{y} = (y_1, \dots, y_N) \in \{0,1\}^N$, their Hamming distance is given by:

$$h(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \delta_i \qquad \text{where} \qquad \delta_i = \begin{cases} 1 & \text{if} \quad x_i \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

Consequently, in this specific case, the i^{th}-affinity class contains B-cells having Hamming distance i from $\overline{\mathbf{x}}$ and the fitness is defined as follows:

Definition 4.6. For all $\mathbf{x}_i \in \{0,1\}^N$, its affinity with a given vertex $\overline{\mathbf{x}}$ is given by $\text{aff}(\mathbf{x}_i, \overline{\mathbf{x}}) := N - h(\mathbf{x}_i, \overline{\mathbf{x}})$.

While performing numerical simulations (Sections 4.3.4 and 4.4.2) we refer to the following transition probability matrix on $\{0, \dots, N\}$:

Definition 4.7. For all $i, j \in \{0, \dots, N\}$:

$$q_{ij} = \mathbb{P}(h(\mathbf{X}_t, \overline{\mathbf{x}}) = j \,|\, h(\mathbf{X}_t, \overline{\mathbf{x}}) = i) = \begin{cases} i/N & \text{if} \quad j = i-1 \\ (N-i)/N & \text{if} \quad j = i+1 \\ 0 & \text{if} \quad |j-i| \neq 1 \end{cases}$$

$\mathcal{Q}_N := (q_{ij})_{0 \leq i,j \leq N}$ is a tridiagonal matrix where the main diagonal consists of zeros.

If we model B-cell traits as vertices of the state-space $\{0,1\}^N$, this corresponds to a model of simple point mutations (see Chapter 2 for more details and variants of this basic mutational model on binary strings).

Except for numerical simulations, in this Chapter we do not restrict to Definitions 4.5 to 4.7. All mathematical results obtained in following sections are independent from the hypotheses corresponding to Definitions 4.5-4.7. Indeed,

in order to define our model we only need to determine $N+1$ distinct affinity classes of B-cell traits with respect to a presented antigen and the probabilities that a GC B-cell passes from a given affinity class to another one thanks to SHMs during the GC reaction.

4.3 Results

In this Section we formalize mathematically the model introduced above. This enables the estimation of various qualitative and quantitative measures of the GC evolution and of the selected pool as well. In Section 4.3.1 we show that a simple GW process describes the evolution of the size of the GC and determine a condition for its extinction. In order to do this we do not need to know the mutational model. Nevertheless, if we want to understand deeply the whole reaction we need to consider a $(N+3)$-type GW process, which we introduce in Section 4.3.2. Therefore we determine explicitly other quantities, such as the average affinity in the GC and the selected pool, or the evolution of the size of the latter. We conclude this section by numerical simulations (Section 4.3.4).

4.3.1 Evolution of the GC size

The aim of this section is to estimate the evolution of the GC size and its extinction probability. In order to do so we define a simple GW process, with respect to the parameters r_d, r_{div} and r_s. Indeed, each B-cell submitted to selection exits the GC pool, independently from its affinity with $\overline{\mathbf{x}}$. Hence we apply some classical results about generating functions and GW processes (see [59], Chapter I). We collect these results for our specific case in Theorem 4.3.2. Corollary 4.3.3 gives explicitly the expected size of the GC at time t and conditions for the extinction of the GC.

Definition 4.8. Let $Z_t^{(z_0)}$, $t \geq 0$ be the random variable (rv) describing the GC-population size at time t, starting from $z_0 \geq 1$ initial B-cells. $(Z_t^{(z_0)})_{t\in\mathbb{N}}$ is a MC (as each cell behaves independently from the others and from the previous generations) on $\{0, 1, 2, \dots\}$.

If $z_0 = 1$ and there is no confusion, we denote $Z_t := Z_t^{(1)}$. By Definition 4.8, Z_1 corresponds to the number of cells in the GC at the first generation, starting from a single seed cell. Thanks to Definition 4.4 one can claim that $Z_1 \in \{0, 1, 2\}$, with the following probabilities:

$$\begin{cases} p_0 := \mathbb{P}(Z_1 = 0) = r_d + (1 - r_d)r_s(1 - r_{div} + r_{div}r_s) \\ p_1 := \mathbb{P}(Z_1 = 1) = (1 - r_d)(1 - r_s)(1 - r_{div} + 2r_{div}r_s) \\ p_2 := \mathbb{P}(Z_1 = 2) = r_{div}(1 - r_d)(1 - r_s)^2 \end{cases} \tag{4.1}$$

As far as next generations are concerned, conditioning to $Z_t = k$, Z_{t+1} is distributed as the sum of k independent copies of Z_1, i.e. $\mathbb{P}(Z_{t+1} = k' \mid Z_t = k) = \mathbb{P}\left(\sum_{i=1}^{k} Z_1 = k'\right)$.

Definition 4.9. Let X be an integer valued rv, $p_k := \mathbb{P}(X = k)$ for all $k \geq 0$. Its probability generating function (pgf) is given by:

$$F_X(s) = \sum_{k=0}^{+\infty} p_k s^k$$

F_X is a convex monotonically increasing function over $[0, 1]$, and $F_X(1) = 1$. If $p_0 \neq 0$ and $p_0 + p_1 < 1$ then F is a strictly increasing function.

Definition 4.10. Given F, the pgf of a rv X, the iterates of F are given by:

$$F_0(s) = s$$

$$F_1(s) = F(s)$$

$$F_t(s) = F(F_{t-1}(s)) \text{ for } t \geq 2$$

Proposition 4.3.1.

(i) *If $\mathbb{E}(X)$ exists (respectively $\mathbb{V}(X)$), then $\mathbb{E}(X) = F'_X(1)$ (respectively $\mathbb{V}(X) = F''_X(1) - (\mathbb{E}(X))^2 + \mathbb{E}(X)$).*

(ii) *If X and Y are two integer valued independent rvs, then $X + Y$ is still an integer valued rv and its pgf is given by $F_{X+Y} = F_X F_Y$.*

The pgf for Z_1 is given by:

$$\begin{aligned} F(s) &= p_0 + p_1 s + p_2 s^2 \\ &= r_d + (1 - r_d)r_s(1 - r_{div} + r_{div}r_s) \\ &\quad +(1 - r_d)(1 - r_s)(1 - r_{div} + 2r_{div}r_s)s \\ &\quad +r_{div}(1 - r_d)(1 - r_s)^2 s^2 \end{aligned} \tag{4.2}$$

Definition 4.11. We denote by η the extinction probability of the process $(Z_t)_{t\in\mathbb{N}}$:

$$\eta := \lim_{t\to\infty} F_t(0)$$

Theorem 4.3.2.

(i) *The pgf of $Z_t^{(z_0)}$, $t \in \mathbb{N}$, which represents the population size of the t^{th}-generation starting from $z_0 \geq 1$ seed cells, is $F_t^{(z_0)} = (F_t)^{z_0}$, F_t being the t^{th}-iterate of F (Equation (4.2)).*

(ii) *The expected size of the GC at time t and starting from z_0 B-cells is given by:*

$$\mathbb{E}(Z_t^{(z_0)}) = (\mathbb{E}(Z_t))^{z_0} = \left((\mathbb{E}(Z_1))^t\right)^{z_0} , \tag{4.3}$$

(iii) *η is the smallest fixed point of the generating function F, i.e. η is the smallest s s.t. $F(s) = s$.*

(iv) *If $\mathbb{E}(Z_1) =: m$ is finite, then:*

- *if $m \leq 1$ then F has only 1 as fixed point and consequently $\eta = 1$;*
- *if $m > 1$ then F as exactly a fixed point on $[0, 1[$ and then $\eta < 1$.*

(v) *Denoted by η_{z_0} the probability of extinction of $(Z_t^{(z_0)})$, one has:*

$$\eta_{z_0} = \eta^{z_0}$$

where η is given by (iii).

By applying Theorem 4.3.2 and Equation (4.1) above, one can prove:

Corollary 4.3.3.

(i) *The expected size of the GC at time t and starting from z_0 initial B-cells is given by:*

$$\mathbb{E}(Z_t^{(z_0)}) = ((1 - r_d)(1 + r_{div})(1 - r_s))^{tz_0} \tag{4.4}$$

(ii) *Denoted by η_{z_0} the extinction probability of the GC population starting from z_0 initial B-cells, one has:*

- *if $r_s \geq 1 - \dfrac{1}{(1 - r_d)(1 + r_{div})}$, then $\eta_{z_0} = 1$*
- *otherwise $\eta_{z_0} = \eta^{z_0} < 1$, η being the smallest fixed point of (4.2)*

In particular the process is subcritical or supercritical independently from z_0. In the supercritical case, increasing the number of B-cells at the beginning of the process makes the probability of extinction decrease. More precisely, in

the case $\eta < 1$, then $\eta_{z_0} \to 0$ if $z_0 \to \infty$.

This section shows that a classical use of a simple GW process enables to understand quantitatively the GC growth. Moreover, Corollary 4.3.3 (ii) gives a condition over the main parameters for the extinction of the GC: if the selection pressure is too high, with probability 1 the GC size goes to 0, independently from the initial number of seed cells. Intuitively, a too high selection pressure prevents those B-cells with bad affinity to improve their fitness undergoing further rounds of mutation and division. Most B-cells will be rapidly submitted to selection, hence either exit the GC as output cells or die by apoptosis if they fail to receive positive selection signals [89].

4.3.2 Evolution of the size and fitness of GC and selected pools

The GW process defined in the previous Section only describes the size of the GC. Indeed, we are not able to say anything about the average fitness of GC clones, or the expected number of selected B-cells, or their average affinity. Hence, we need to consider a more complex model and take into account the parameter $\overline{a}_s$ and the transition probability matrix characterizing the mutational rule. We introduce a multi-type GW Process (see for instance [7], chapter V).

Definition 4.12. Let $\mathbf{Z}_t^{(\mathbf{i})} = (Z_{t,0}^{(\mathbf{i})}, \dots, Z_{t,N+2}^{(\mathbf{i})})$, $t \geq 0$ be a MC where for all $0 \leq j \leq N$, $Z_{t,j}^{(\mathbf{i})}$ describes the number of GC B-cells belonging to the j^{th}-affinity class with respect to $\overline{\mathbf{x}}$, $Z_{t,N+1}^{(\mathbf{i})}$ the number of selected B-cells and $Z_{t,N+2}^{(\mathbf{i})}$ the number of dead B-cells at generation t, when the process is initiated in state $\mathbf{i} = (i_0, \dots, i_N, 0, 0)$.

For all $j \in \{0, \dots, N+2\}$ the generating function gives the number of offsprings of each type that a type j particle can produce. It is defined as follows:

$$f^{(j)}(s_0, \dots, s_{N+2}) = \sum_{k_0, \dots, k_{N+2} \geq 0} p^{(j)}(k_0, \dots, k_{N+2}) s_0^{k_0} \cdots s_{N+2}^{k_{N+2}}, \tag{4.5}$$

$$0 \leq s_\alpha \leq 1 \text{ for all } \alpha \in \{0, \dots, N+2\}$$

where $p^{(j)}(k_0, \dots, k_{N+2})$ is the probability that a type j cell produces k_0 cells of type 0, k_1 of type 1, …, k_{N+2} of type $N+2$ for the next generation.
We denote:

- $\mathbf{p}(\mathbf{k}) = (p^{(0)}(\mathbf{k}), \dots, p^{(N+2)}(\mathbf{k}))$, for $\mathbf{k} = (k_0, \dots, k_{N+2}) \in \mathbb{Z}_+^{N+3}$

- $\mathbf{f}(\mathbf{s}) = (f^{(1)}(\mathbf{s}), \ldots, f^{(N+1)}(\mathbf{s}))$, for $\mathbf{s} = (s_0, \ldots, s_{N+2}) \in \mathcal{C}^{N+3}$ where $\mathcal{C}^{N+3} := \{\mathbf{x} \in \mathbb{R}^{N+3} \,|\, 0 \leq x_\alpha \leq 1, \, \alpha \in \{0, \ldots, N+2\}\}$

The probability generating function of $\mathbf{Z}_1$ is given by:

$$\mathbf{f}(\mathbf{s}) = \sum_{\mathbf{k} \in \mathbb{Z}_+^{N+3}} \mathbf{p}(\mathbf{k}) \mathbf{s}^{\mathbf{k}}, \, \mathbf{s} \in \mathcal{C}^{N+3} \tag{4.6}$$

Again, the generating function of $\mathbf{Z}_t$, $\mathbf{f}_t(\mathbf{s})$, is obtained as the t^{th}-iterate of $\mathbf{f}$, and it holds true that:

$$\mathbf{f}_{t+r}(\mathbf{s}) = \mathbf{f}_t[\mathbf{f}_r(\mathbf{s})], \, \mathbf{s} \in \mathcal{C}^{N+3}.$$

Let $m_{ij} := \mathbb{E}[Z_{1,j}^{(i)}]$ the expected number of offspring of type j of a cell of type i in one generation. We collect all m_{ij} in a matrix, $\mathcal{M} = (m_{ij})_{0 \leq i,j \leq N+2}$. We have:

$$m_{ij} = \frac{\partial f^{(i)}}{\partial s_j}(\mathbf{1})$$

and:

$$\mathbb{E}[Z_{t,j}^{(i)}] = \frac{\partial f_t^{(i)}}{\partial s_j}(\mathbf{1}) \tag{4.7}$$

Finally:

$$\mathbb{E}[\mathbf{Z}_t^{(\mathbf{i})}] = \mathbf{i}\mathcal{M}^t \tag{4.8}$$

We can now explicitly give the elements of $\mathcal{M}$ by using matrix $\mathcal{Q}_N$ given by Definition 4.3.

Proposition 4.3.4. *$\mathcal{M}$ is a $(N+3) \times (N+3)$ matrix, which we can define as a block matrix in the following way:*

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}_1 & \mathcal{M}_2 \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix}$$

Where:

- $\mathbf{0}_{2\times(N+1)}$ *is a $2 \times (N+1)$ matrix with all entries 0;*
- $\mathcal{I}_n$ *is the identity matrix of size n;*
- $\mathcal{M}_1 = 2(1-r_d)r_{div}(1-r_s)\mathcal{Q}_N + (1-r_d)(1-r_{div})(1-r_s)\mathcal{I}_{N+1}$
- $\mathcal{M}_2 = (m_{2,ij})$ *is a $(N+1) \times 2$ matrix where for all $i \in \{0, \ldots, N\}$:*
  - *if $i \leq \overline{a}_s$:*

$$m_{2,i1} = (1-r_d)(1-r_{div})r_s + 2(1-r_d)r_{div}r_s \sum_{j=0}^{\overline{a}_s} q_{ij},$$

$$m_{2,i2} = r_d + 2(1-r_d)r_{div}r_s \sum_{j=\overline{a}_s+1}^{N} q_{ij}$$

– *if* $i > \overline{a}_s$*:*

$$m_{2,i1} = 2(1-r_d)r_{div}r_s \sum_{j=0}^{\overline{a}_s} q_{ij},$$

$$m_{2,i2} = r_d + (1-r_d)(1-r_{div})r_s + 2(1-r_d)r_{div}r_s \sum_{j=\overline{a}_s+1}^{N} q_{ij}$$

Proof. It suffices to compute $f^{(i)}(\mathbf{s})$ for $i = 0, \ldots, N+2$, which depend on r_d, r_{div}, r_s and the elements of $\mathcal{Q}_N$. First, the elements of the $(N+2)^{\text{th}}$ and $(N+3)^{\text{th}}$-lines are obviously determined: all selected (resp. dead) cells remain selected (resp. dead) for next generations, as they can not give rise to any other cell type offspring (we do not take into account here any type of recycling mechanism). Let $i \in \{0, \ldots, N\}$ be a fixed index: we evaluate m_{ij} for all $j \in \{0, \ldots, N+2\}$. The first step is to determine the value of $p^{(i)}(\mathbf{k})$ for $\mathbf{k} = (k_0, \ldots, k_{N+2}) \in \mathbb{Z}_+^{N+3}$. There exists only a few cases in which $p^{(i)}(\mathbf{k}) \neq 0$, which can be explicitly evaluated:

- $p^{(i)}(0, \ldots, 0, 1) = \begin{cases} r_d & \text{if } i \leq \overline{a}_s \\ r_d + (1-r_d)(1-r_{div})r_s & \text{otherwise} \end{cases}$
- $p^{(i)}(0, \ldots, 0, 1, 0) = \begin{cases} (1-r_d)(1-r_{div})r_s & \text{if } i \leq \overline{a}_s \\ 0 & \text{otherwise} \end{cases}$
- $p^{(i)}(0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0, 0) = (1-r_d)(1-r_{div})(1-r_s)$
- $p^{(i)}(0, \ldots, 0, 2) = (1-r_d)r_{div}r_s^2 \sum_{j_1=\overline{a}_s+1}^{N} q_{ij_1} \sum_{j_2=\overline{a}_s+1}^{N} q_{ij_2}$
- $p^{(i)}(0, \ldots, 0, 2, 0) = (1-r_d)r_{div}r_s^2 \sum_{j_1=0}^{\overline{a}_s} q_{ij_1} \sum_{j_2=0}^{\overline{a}_s} q_{ij_2}$
- $p^{(i)}(0, \ldots, 0, 1, 1) = 2(1-r_d)r_{div}r_s^2 \sum_{j_1=0}^{\overline{a}_s} q_{ij_1} \sum_{j_2=\overline{a}_s+1}^{N} q_{ij_2}$
- For all $j_1 < j_2 \in \{0, \ldots, N\}$:
  - $p^{(i)}(0, \ldots, 0, \underset{j_1}{2}, 0, \ldots, 0, 0) = (1-r_d)r_{div}(1-r_s)^2 q_{ij_1}^2$
  - $p^{(i)}(0, \ldots, 0, \underset{j_1}{1}, 0, \ldots, 0, \underset{j_2}{1}, 0, \ldots, 0, 0) = 2(1-r_d)r_{div}(1-r_s)^2 q_{ij_1}q_{ij_2}$

– $p^{(i)}(0,\ldots,0,\underset{j_1}{1},0,\ldots,0,1) = 2(1-r_d)r_{div}r_s(1-r_s)q_{ij_1}\sum_{j_2=\overline{a}_s+1}^{N} q_{ij_2}$

– $p^{(i)}(0,\ldots,0,\underset{j_1}{1},0,\ldots,0,1,0) = 2(1-r_d)r_{div}r_s(1-r_s)q_{ij_1}\sum_{j_2=0}^{\overline{a}_s} q_{ij_2}$

- $p^{(i)}(\mathbf{k}) = 0$ otherwise

We can therefore evaluate $f^{(i)}(\mathbf{s})$, with $\mathbf{s} = (s_0,\ldots,s_{N+2}) \in \mathcal{C}^{N+3}$.

For all $i \leq \overline{a}_s$:

$$
\begin{aligned}
f^{(i)}(\mathbf{s}) &= r_d s_{N+2} + (1-r_d)(1-r_{div})r_s s_{N+1} + (1-r_d)(1-r_{div})(1-r_s)s_i \\
&+ (1-r_d)r_{div}r_s^2\left(\sum_{j_1=\overline{a}_s+1}^{N} q_{ij_1}\sum_{j_2=\overline{a}_s+1}^{N} q_{ij_2}s_{N+2}^2\right. \\
&\left.+\sum_{j_1=0}^{\overline{a}_s} q_{ij_1}\sum_{j_2=0}^{\overline{a}_s} q_{ij_2}s_{N+1}^2 + 2\sum_{j_1=0}^{\overline{a}_s} q_{ij_1}\sum_{j_2=\overline{a}_s+1}^{N} q_{ij_2}s_{N+1}s_{N+2}\right) \\
&+ (1-r_d)r_{div}(1-r_s)^2\left(\sum_{j_1=0}^{N} q_{ij_1}^2 s_{j_1}^2 + 2\sum_{j_1=0}^{N} q_{ij_1}\sum_{j_2<j_1=0}^{N} q_{ij_2}s_{j_1}s_{j_2}\right) \\
&+ 2(1-r_d)r_{div}r_s(1-r_s)\sum_{j_1=0}^{N} q_{ij_1}\left(\sum_{j_2=\overline{a}_s+1}^{N} q_{ij_2}s_{N+2} + \sum_{j_2=0}^{\overline{a}_s} q_{ij_2}s_{N+1}\right)s_{j_1}
\end{aligned}
\tag{4.9}
$$

If $i > \overline{a}_s$ then $f^{(i)}(\mathbf{s})$ is the same except for the first line, which becomes:

$$(r_d + (1-r_d)(1-r_{div})r_s)s_{N+2} + (1-r_d)(1-r_{div})(1-r_s)s_i$$

The values of each m_{ij} are now obtained by evaluating all partial derivatives of $f^{(i)}(\mathbf{s})$ in $\mathbf{1}$, keeping in mind that for all $i \in \{0,\ldots,N\}$, $\sum_{j=0}^{N} q_{ij} = 1$.

□

Example 2. One can give explicitly the form of matrix $\mathcal{M}_2$ corresponding to the mutational model defined in Definition 4.7:

$$\mathcal{M}_2 = \begin{array}{c} 0 \\ \vdots \\ \overline{a}_s - 1 \\ \overline{a}_s \\ \overline{a}_s + 1 \\ \overline{a}_s + 2 \\ \vdots \\ N \end{array} \begin{pmatrix} \alpha & r_d \\ \vdots & \vdots \\ \alpha & r_d \\ \alpha - \beta + \beta\frac{\overline{a}_s}{N} & r_d + \beta\frac{N-\overline{a}_s}{N} \\ \beta\frac{\overline{a}_s+1}{N} & r_d + \alpha - \beta + \beta\frac{N-(\overline{a}_s+1)}{N} \\ 0 & r_d + \alpha \\ \vdots & \vdots \\ 0 & r_d + \alpha \end{pmatrix},$$

where:

- $\alpha := (1 - r_d)(1 + r_{div})r_s$
- $\beta := 2(1 - r_d)r_{div}r_s$

Remark 30. Independently from the given mutational model, $\alpha + r_d$ corresponds to the expected number of selected or dead B-cells that each GC B-cell can produce in a single time step.

Of course in the multi-type context we recover again results from Section 4.3.1. For this sake, let us recall some results about the extinction probability for multi-type GW processes [7].

Definition 4.13. Let $q^{(i)}$ be the probability of eventual extinction of the process, when it starts from a single type i cell. As above bold symbols denote vectors *i.e.* $\mathbf{q} := (q^{(0)}, \ldots, q^{(N+2)}) \geq 0$.

Definition 4.14. We say that $(\mathbf{Z}_t)$ is singular if each particle has exactly one offspring, which implies that the branching process becomes a simple MC.

Definition 4.15. Matrix $\mathcal{M}$ is said to be strictly positive if it has non-negative entries and there exists a t s.t. $(\mathcal{M}^t)_{ij} > 0$ for all i, j. $(\mathbf{Z}_t)$ is positive regular iff $\mathcal{M}$ is strictly positive.

Notation 4. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. We say that $\mathbf{u} \leq \mathbf{v}$ if $u_i \leq v_i$ for all $i \in \{1, \ldots, n\}$. Moreover, we say that $\mathbf{u} < \mathbf{v}$ if $\mathbf{u} \leq \mathbf{v}$ and there exists at least an index j s.t. $u_j < v_j$.

Theorem 4.3.5. *Let $(\mathbf{Z}_t)$ be non singular and strictly positive. Let ρ be the maximum eigenvalue of $\mathcal{M}$. The following three results hold:*

1. *If $\rho < 1$ (subcritical case) or $\rho = 1$ (critical case) then $\mathbf{q} = \mathbf{1}$. Otherwise, if $\rho > 1$ (supercritical case), then $\mathbf{q} < \mathbf{1}$.*
2. $\lim_{t\to\infty} \mathbf{f}_t(\mathbf{s}) = \mathbf{q}$, *for all* $\mathbf{s} \in \mathcal{C}^{N+3}$.
3. $\mathbf{q}$ *is the only solution of* $\mathbf{f}(\mathbf{s}) = \mathbf{s}$ *in* $\mathcal{C}^{N+3}$.

The spectra of matrix $\mathcal{M}$ defined in Definition 4.3.4 is obtained as follows:

Proposition 4.3.6. *Let $\mathcal{M}$ be defined as a block matrix as in 4.3.4. Let $\lambda_{\mathcal{M},i}$ be its i^{th}-eigenvalue. The spectra of $\mathcal{M}$ is given by:*

- *For all $i \in \{0, \dots, N\}$, $\lambda_{\mathcal{M},i} = (1 - r_d)(1 - r_s)(1 + r_{div}(2\lambda_i - 1))$, where λ_i is the i^{th}-eigenvalue of matrix $\mathcal{Q}_N$.*
- *whereas $\lambda_{\mathcal{M},N+1} = 1$ with multiplicity 2.*

Proof. As $\mathcal{M}$ is a block matrix with the lower left block composed of zeros, then $Spec(\mathcal{M}) = Spec(\mathcal{M}_1) \cup Spec(\mathcal{I}_2)$. The result follows.

□

Therefore we obtain the same condition as in Corollary 4.3.3 for the extinction probability in the GC:

Proposition 4.3.7. *Let $\mathbf{q}$ be the extinction probability for the process $(\mathbf{Z}_t)$ defined in Definition 4.20 and restricted to the first $N + 1$ components (*i.e. *we refer only to matrix $\mathcal{M}_1$, which defines the expectations of GC B-cells). Therefore:*

- *if $r_s \geq 1 - \dfrac{1}{(1 - r_d)(1 + r_{div})}$, then $\mathbf{q} = \mathbf{1}$*
- *otherwise $\mathbf{q} < \mathbf{1}$ is the smallest fixed point of $\mathbf{f}(\mathbf{s})$ in $\mathcal{C}^{N+3}$.*

Proof. $\mathcal{Q}_N$ is a stochastic matrix, therefore its largest eigenvalue is 1. The corresponding eigenvalue of matrix $\mathcal{M}_1$ is: $\lambda_{\mathcal{M}_1,1} = (1-r_d)(1-r_s)(1+r_{div})$. The proposition is proved by observing that $\lambda_{\mathcal{M}_1,1} \leq 1 \Leftrightarrow r_s \geq 1 - \dfrac{1}{(1 - r_d)(1 + r_{div})}$ and applying Theorem 4.3.5 (note that $\mathcal{M}_1$ is positive regular: this is not the case for matrix $\mathcal{M}$).

□

In order to determine the expected number of selected cells at a given time t, we need to introduce another multi-type GW process.

Definition 4.16. Let $\widetilde{\mathbf{Z}}_t^{(\mathbf{i})} = (\widetilde{Z}_{t,0}^{(\mathbf{i})}, \dots, \widetilde{Z}_{t,N+2}^{(\mathbf{i})})$, $t \geq 0$ be a MC where for all $0 \leq j \leq N$, $\widetilde{Z}_{t,j}^{(\mathbf{i})}$ describes the number of GC B-cells belonging to the j^{th}-affinity class with respect to $\overline{\mathbf{x}}$, $\widetilde{Z}_{t,N+1}^{(\mathbf{i})}$ the number of selected B-cells and $\widetilde{Z}_{t,N+2}^{(\mathbf{i})}$ the number of dead B-cells at generation t, when the process is initiated in state $\mathbf{i} = (i_0, \dots, i_N, 0, 0)$ and before the selection mechanism is performed for the t^{th}-generation.

Proceeding as we did for $\mathbf{Z}_t^{(\mathbf{i})}$, we can determine matrix $\widetilde{\mathcal{M}}$ whose elements are $\widetilde{m}_{ij} := \mathbb{E}[\widetilde{Z}_{1,j}^{(i)}]$ for all $i, j \in \{0, \dots, N+2\}$.

Proposition 4.3.8. *$\widetilde{\mathcal{M}}$ is a $(N+3) \times (N+3)$ matrix, which only depends on matrix $\mathcal{Q}_N$, r_d and r_{div} and can be defined as a block matrix as follows:*

$$\widetilde{\mathcal{M}} = \begin{pmatrix} \widetilde{\mathcal{M}}_1 & \widetilde{\mathcal{M}}_2 \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix}$$

Where:

- $\widetilde{\mathcal{M}}_1 = 2(1-r_d)r_{div}\mathcal{Q}_N + (1-r_d)(1-r_{div})\mathcal{I}_{N+1}$
- $\widetilde{\mathcal{M}}_2 = (\mathbf{0}_{N+1}, r_d \cdot \mathbf{1}_{N+1})$, *where $\mathbf{0}_{N+1}$ (resp. $\mathbf{1}_{N+1}$) is a $(N+1)$-column vector whose elements are all 0 (resp. 1).*

Notation 5. Let S_t, $t \geq 0$ be the random variable describing the number of selected B-cells at time t. By hypothesis $S_0 = 0$. $(S_t)_{t\in\mathbb{N}}$ is a MC on $\{0, 1, 2, \dots\}$.

We can therefore prove the following results:

Proposition 4.3.9. *Let $\mathbf{i}$ be the initial state.*

- *The expected size of the GC at time t is given by:*

$$\sum_{k=0}^{N} (\mathbf{i}\mathcal{M}^t)_k \tag{4.10}$$

- *The average affinity in the GC at time t is given by:*

$$\frac{\sum_{k=0}^{N} (N-k)(\mathbf{i}\mathcal{M}^t)_k}{\sum_{k=0}^{N} (\mathbf{i}\mathcal{M}^t)_k} \tag{4.11}$$

- *The expected number of selected B-cells at time t is given by:*

$$\mathbb{E}(S_t) = r_s \sum_{k=0}^{\overline{a}_s} \left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)_k \tag{4.12}$$

- *The expected number of selected B-cells produced until time t is given by:*

$$\mathbb{E}\left[\sum_{n=0}^{t} S_n\right] = \mathbb{E}\left[\left(\mathbf{Z}_t^{(\mathbf{i})}\right)_{N+2}\right] = \left(\mathbf{i}\mathcal{M}^t\right)_{N+2} \tag{4.13}$$

- *The average affinity of selected B-cells at time t is given by:*

$$\frac{\sum_{k=0}^{\overline{a}_s}(N-k)\left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)_k}{\sum_{k=0}^{\overline{a}_s}\left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)_k} \tag{4.14}$$

- *The average affinity of selected B-cells until time t is given by:*

$$\frac{r_s \sum_{s=1}^{t}\sum_{k=0}^{\overline{a}_s}(N-k)\left(\mathbf{i}\mathcal{M}^{s-1}\widetilde{\mathcal{M}}\right)_k}{\left(\mathbf{i}\mathcal{M}^t\right)_{N+2}} \tag{4.15}$$

Proof. Equations (4.10), (4.11) and (4.13) are a direct application of what stated in Equation (4.8). In order to prove Equation (4.13) we have to observe that:

$$\mathbb{E}\left[\widetilde{\mathbf{Z}}_t^{(\mathbf{i})}\right] = \mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\,, \tag{4.16}$$

since due to the Markov property of the process, the behavior of $\widetilde{\mathbf{Z}}_t^{(\mathbf{i})}$ only depends on the distribution of $\mathbf{Z}_{t-1}^{(\mathbf{i})}$. Moreover, we have to remark that the expected number of selected B-cells at time t is obtained from the expected number of B-cells in GC at time t (before the selection mechanism is performed) having fitness good enough to be positive selected. This is clearly given by $\sum_{k=0}^{\overline{a}_s}\left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)_k$, thanks to (4.16). We have just to multiply this expectation for the probability that each of these B-cells is submitted to mutation, *i.e.* r_s. Finally, results about the average affinity in both the GC and the selected pool (Equations (4.11), (4.14) and (4.15)) are obtained from the previous ones by multiplying the number of individuals belonging to the same class by their fitness (Definition 4.2), and dividing by the total number of individuals in the considered pool. The definition of affinity as function of the affinity classes, determines Equations (4.11), (4.14) and (4.15).

□

The expected size of the GC at time t can be obtained applying a simple GW process (Section 4.3.1) and is given by (4.4). It is possible to prove the same result starting from the $(N+3)$-type GW process (4.10). For the sake of simplicity, let us suppose that the process starts from a single B-cell belonging to the affinity class $a_0 = i$ with respect to the target trait. We do not need to specify the transition probability matrix used to define the mutational model allowed.

We can easily prove by iteration that:

$$\mathcal{M}^t = \begin{pmatrix} \mathcal{M}_1^t & \sum_{k=0}^{t-1} \mathcal{M}_1^k \mathcal{M}_2 \\ \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix} \tag{4.17}$$

Therefore we can claim that $(\mathbf{i}\mathcal{M}^t)_k$ corresponds to the k^{th}-component of the i^{th}-row of matrix $\mathcal{M}_1^t = (2(1-r_d)r_{div}(1-r_s)\mathcal{Q}_N + (1-r_d)(1-r_{div})(1-r_s)\mathcal{I}_{N+1})^t$, where $\mathcal{Q}_N$ is a stochastic matrix. Matrices $\mathcal{A} := 2(1-r_d)r_{div}(1-r_s)\mathcal{Q}_N$ and $\mathcal{B} := (1-r_d)(1-r_{div})(1-r_s)\mathcal{I}_{N+1}$ clearly commute, therefore we write:

$$(\mathcal{A}+\mathcal{B})^t = \sum_{j=0}^{t} C_t^j \mathcal{A}^{t-j}\mathcal{B}^j \tag{4.18}$$

For all j, $0 \le j \le t$:

$$\begin{aligned} \mathcal{A}^{t-j}\mathcal{B}^j &= 2^{t-j}(1-r_d)^{t-j} r_{div}^{t-j}(1-r_s)^{t-j}(1-r_d)^j(1-r_{div})^j(1-r_s)^j \mathcal{Q}_N^{t-j} \\ &= (1-r_d)^t(1-r_s)^t(2r_{div})^{t-j}(1-r_{div})^j \mathcal{Q}_N^{t-j} \end{aligned}$$

Hence:

$$(\mathcal{A}+\mathcal{B})^t = (1-r_d)^t(1-r_s)^t \sum_{j=0}^{t} C_t^j (2r_{div})^{t-j}(1-r_{div})^j \mathcal{Q}_N^{t-j}$$

And consequently:

$$\begin{aligned} \sum_{k=0}^{N} (\mathbf{i}\mathcal{M}^t)_k &= \sum_{k=0}^{N} \left(\mathbf{i}\,(\mathcal{A}+\mathcal{B})^t\right)_k \\ &= (1-r_d)^t(1-r_s)^t \sum_{j=0}^{t} C_t^j (2r_{div})^{t-j}(1-r_{div})^j \sum_{k=0}^{N} \left(\mathbf{i}\mathcal{Q}_N^{t-j}\right)_k \end{aligned}$$

Since $\mathcal{Q}_N$ is a stochastic matrix, for all n, $\mathcal{Q}_N^n$ is still a stochastic matrix, *i.e.* the entries of each row of $\mathcal{Q}_N^n$ sum to 1. Therefore:

$$\begin{aligned}\sum_{k=0}^{N}(\mathbf{i}\mathcal{M}^t)_k &= (1-r_d)^t(1-r_s)^t\sum_{j=0}^{t}C_t^j(2r_{div})^{t-j}(1-r_{div})^j \\ &= (1-r_d)^t(1-r_s)^t(2r_{div}+1-r_{div})^t = (1-r_d)^t(1-r_s)^t(1+r_{div})^t \ ,\end{aligned}$$

as stated by Equation (4.4) for $z_0 = 1$. This result can be easily generalized to the case of $z_0 \geq 1$ initial B-cells.

4.3.3 r_s maximizing the expectation of selected B-cells at time t

What is the behavior of the expected number of selected B-cells as a function of the model parameters ? In particular, is there an optimal value of the selection rate which maximizes this number ? In this section we show that, indeed, the answer is positive.

To do so we detail hereafter the computation of $\mathbb{E}(S_t)$ (Equation (4.12)), given by Proposition 4.3.9.

Let us suppose, for the sake of simplicity, that $\mathcal{Q}_N$ is diagonalizable:

$$\mathcal{Q}_N = R\Lambda_N L \ , \tag{4.19}$$

where $\Lambda_N = \mathrm{diag}(\lambda_0, \ldots, \lambda_N)$, and $R = (r_{ij})$ (resp. $L = (l_{ij})$) is the transition matrix whose rows (resp. lines) contain the right (resp. left) eigenvectors of $\mathcal{Q}_N$, corresponding to $\lambda_0, \ldots, \lambda_N$. This is the case, for example, if we consider the mutational model given by Definition 4.7. Moreover, in this specific case, the $N+1$ distinct eigenvalues of $\mathcal{Q}_N$ are known explicitly (Chapter 2):

$$\lambda_0 = 1 \leq 1 - \frac{1}{N} \leq 1 - \frac{2}{N} \leq \cdots \leq -1 + \frac{2}{N} \leq -1 + \frac{1}{N} \leq -1 = \lambda_N$$

It follows from (4.17) and (4.19) that for all $t \geq 1$, $\mathcal{M}^t$ can be written as:

$$\mathcal{M}^t = \begin{pmatrix} RD^tL & \left(R\sum_{k=0}^{t-1}D^kL\right)\mathcal{M}_2 \\ \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix} \ , \tag{4.20}$$

where $D = 2(1-r_d)r_{div}(1-r_s)\Lambda_N + (1-r_d)(1-r_{div})(1-r_s)\mathcal{I}_{N+1}$ is a diagonal matrix. We obtain its expression thanks to Proposition 4.3.4.

Moreover, by Proposition 4.3.8 and Equation (4.19) we have:

$$\widetilde{\mathcal{M}} = \begin{pmatrix} R\widetilde{D}L & \widetilde{\mathcal{M}}_2 \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix} , \tag{4.21}$$

where $\widetilde{D} = 2(1-r_d)r_{div}\Lambda_N + (1-r_d)(1-r_{div})\mathcal{I}_{N+1}$ is a diagonal matrix.

Proposition 4.3.10. *Let us suppose that at time $t = 0$ there is a single B-cell entering the GC belonging to the i^{th}-affinity class with respect to the target cell. Moreover, let us suppose that $\mathcal{Q}_N = R\Lambda_N L$. For all $t \geq 1$, the expected number of selected B-cells at time t, is:*

$$\mathbb{E}(S_t) = r_s(1-r_s)^{t-1}(1-r_d)^t \sum_{\ell=0}^{N}(2\lambda_\ell r_{div} + 1 - r_{div})^t \sum_{k=0}^{\overline{a}_s} r_{i\ell} l_{\ell k} \ ,$$

Proof. Proposition 4.3.9 claims:

$$\mathbb{E}(S_t) = r_s \sum_{k=0}^{\overline{a}_s} \left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)_k$$

We have to explicitly write $\left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)_k$. From Equations (4.20) and (4.21):

$$\mathcal{M}^{t-1}\widetilde{\mathcal{M}} = \begin{pmatrix} RD^{t-1}\widetilde{D}L & RD^{t-1}L\widetilde{\mathcal{M}}_2 + \left(R\sum_{k=0}^{t-2} D^k L\right)\mathcal{M}_2 \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix}$$

Since, by hypothesis, $\mathbf{i} = (0,\dots,0,1,0,\dots,0,0)$, with the only 1 being at position i, $0 \leq i \leq N$, then $\left(\mathbf{i}\mathcal{M}^{t-1}\widetilde{\mathcal{M}}\right)$ denotes the i^{th}-row of matrix $\mathcal{M}^{t-1}\widetilde{\mathcal{M}}$. Therefore, we are interested in the sum between 0 and $\overline{a}_s$ of the elements of the i^{th}-row of matrix $\mathcal{M}^{t-1}\widetilde{\mathcal{M}}$, *i.e.* of the i^{th}-row of matrix $RD^{t-1}\widetilde{D}L$, since clearly $\overline{a}_s \leq N$. $D^{t-1}\widetilde{D}$ is a diagonal matrix whose ℓ^{th}-diagonal element is given by:

$$\begin{aligned} \left(D^{t-1}\widetilde{D}\right)_\ell &= (2(1-r_d)r_{div}(1-r_s)\lambda_\ell + (1-r_d)(1-r_{div})(1-r_s))^{t-1} \\ &\quad \cdot(2(1-r_d)r_{div}\lambda_\ell + (1-r_d)(1-r_{div})) \\ &= (1-r_s)^{t-1}(1-r_d)^t\,(2\lambda_\ell r_{div} + 1 - r_{div})^t \end{aligned}$$

The result follows observing that: $\left(RD^{t-1}\widetilde{D}L\right)_{ik} = \sum_{\ell=0}^{N}\left(D^{t-1}\widetilde{D}\right)_{\ell} r_{i\ell}l_{\ell k}$.

□

As an immediate consequence of Proposition 4.3.10, we can claim:

Corollary 4.3.11. *For all time $t \geq 1$ the value $r_s(t)$ which maximizes the expected number of selected B-cells at time t is:*

$$r_s(t) = \frac{1}{t}$$

Proof. Since $(1-r_d)^t \sum_{\ell=0}^{N}(2\lambda_\ell r_{div} + 1 - r_{div})^t \sum_{k=0}^{\bar{a}_s} r_{i\ell}l_{\ell k}$ is a non negative quantity independent from r_s, the value of r_s which maximizes $\mathbb{E}(S_t)$ is the one that maximizes $r_s(1-r_s)^{t-1}$. The result trivially follows.

□

Under certain hypotheses about the mutational model and the GC evolution, one could justify the claim of Corollary 4.3.11 by heuristic arguments, without considering the $(N+3)$-type GW process. This leads to approximately estimate the expected number of selected B-cells at time t.

Hypothesis 1. $\mathcal{Q}_N$ converges through its stationary distribution, denoted by $\mathbf{m} = (m_i)$, $i \in \{0, \dots, N\}$.

Hypothesis 2. Z_t explodes, where $(Z_t)_{t\in\mathbb{N}}$ is given by Definition 4.8.

Let $\widetilde{Z}_t$, $t \geq 0$ be the random variable describing the GC-population size at time t before the selection mechanism is performed for this generation. For the sake of simplicity, let us suppose $\widetilde{Z}_0 = 1$. $(\widetilde{Z}_t)_{t\in\mathbb{N}}$ is a MC on $\{0, 1, 2, \dots\}$. Denoted by $\tilde{p}_k := \mathbb{P}(\widetilde{Z}_1 = k)$, $k \in \{0, 1, 2\}$:

$$\begin{cases} \tilde{p}_0 = r_d \\ \tilde{p}_1 = (1-r_d)(1-r_{div}) \\ \tilde{p}_2 = (1-r_d)r_{div} \end{cases} \tag{4.22}$$

It follows: $\tilde{m} := \mathbb{E}(\widetilde{Z}_1) = (1-r_d)(1-r_{div}) + 2(1-r_d)r_{div} = (1-r_d)(1+r_{div})$.

Conditioning to $Z_t = k$, $\widetilde{Z}_{t+1}$ is distributed as the sum of k independent copies of $\widetilde{Z}_1$, which gives:

$$\mathbb{E}(\widetilde{Z}_t) = \mathbb{E}(Z_{t-1})\mathbb{E}(\widetilde{Z}_1) = \mathbb{E}(Z_1)^{t-1}\mathbb{E}(\widetilde{Z}_1) = (1-r_d)^t(1+r_{div})^t(1-r_s)^{t-1} \tag{4.23}$$

Thanks to Hypotheses 1 and 2, if t is big enough, there is approximately a proportion of m_i elements in the i^{th}-affinity class with respect to $\overline{\mathbf{x}}$. Therefore, on average at time t there are approximately $\sum_{i=0}^{\overline{a}_s} m_i \mathbb{E}(\widetilde{Z}_t)$ B-cells in the GC belonging to an affinity class with index at most equal to $\overline{a}_s$ with respect to $\overline{\mathbf{x}}$, before the selection mechanism is performed for this generation. Each one of these cells can be submitted to selection with probability r_s, and in this case it will be positively selected. Hence:

$$\mathbb{E}(S_t) \simeq r_s \sum_{i=0}^{\overline{a}_s} m_i \mathbb{E}(\widetilde{Z}_t) = (1-r_d)^t (1+r_{div})^t (1-r_s)^{t-1} r_s \sum_{i=0}^{\overline{a}_s} m_i \ , \qquad (4.24)$$

which is maximized at time $t \geq 1$ for $r_s(t) = 1/t$.

Remark 31. One observes that the approximation in (4.24) gives the same value for the optimal $r_s(t)$ as in Corollary 4.3.11. Nevertheless, it does not allow to describe exactly the behavior of $\mathbb{E}(S_t)$, since it is obtained by approximating the distribution of B-cells in the GC with their stationary distribution.

4.3.4 Numerical simulations

We evaluate numerically results of Proposition 4.3.9. The $(N+3)$-type GW process allows a deeper understanding of the dynamics of both populations: inside the GC and in the selected pool. Through numerical simulations we emphasize the dependence of the quantities defined in Proposition 4.3.9 on parameters involved in the model.

We suppose that at the beginning of the process there is a single B-cell entering the GC belonging to the affinity class a_0. Of course, the model we set allows to simulate any possible initial conditions. Indeed, by fixing the initial vector $\mathbf{i}$, we can decide to start the reaction with more B-cells, in different affinity classes. We consider $\mathcal{Q}_N$ given by Definition 4.7 as transition probability matrix characterizing the mutational mechanism. When it is not stated otherwise, we set $N = 10$, $r_s = 0.1$, $r_d = 0.1$, $r_{div} = 0.9$, $a_0 = 3$ and $\overline{a}_s = 3$. This parameter choice implies a small extinction probability, hence a great probability of explosion of the GC population (Corollary 4.3.3).

Evolution of the GC population

The evolution of the size of the GC can be studied by using the simple GW process defined in Section 4.3.1. Equation (4.4), in the case of a single initial B-cell, evidences that the expected number of B-cells within the GC for this

model only depends on r_d, r_{div} and r_s and it is not driven by the initial affinity, nor by the threshold chosen for positive selection $\overline{a}_s$, nor by the mutational rule.

Equation (4.4) evidences that, independently from the transition probability matrix defining the mutational mechanism, the GC size at time t increases with r_{div} and decreases for increasing r_s and r_d. Moreover, the impact of these last two parameters is the same for the growth of the GC. One could expect this behavior since the effect of both the death and the selection on a B-cell is the exit from the GC.

In order to study the evolution of the average affinity within the GC, we need to refer to the $(N+3)$-type GW process defined in Section 4.3.2.

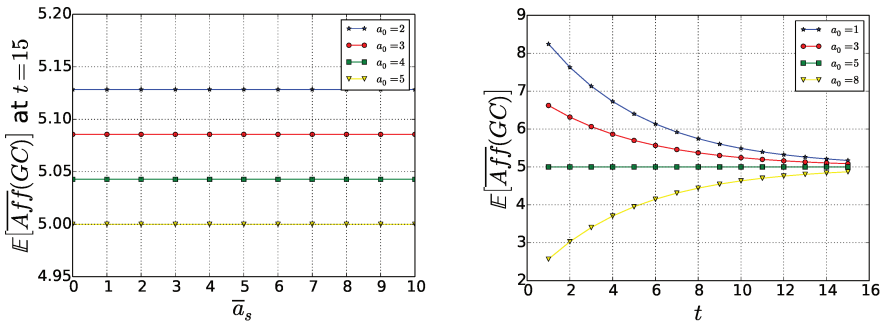


Figure 4.2: (a) Dependence of the expected average affinity in the GC on $\overline{a}_s$ at time $t = 15$, for different values of a_0. The average affinity in the GC is constant with respect to $\overline{a}_s$. (b) The evolution during time of the expected average affinity in the GC for different values of a_0. The average affinity converges through $N/2$, due to the stationary distribution of Q_N, the binomial probability distribution.

Proposition 4.3.12. *Let us suppose that $\mathcal{Q}_N = R\Lambda_N L$. The average affinity within the GC at time t, starting from a single B-cell belonging to the i^{th}-affinity class with respect to $\overline{\mathbf{x}}$ is given by:*

$$N - \frac{\sum_{\ell=0}^{N}(2\lambda_\ell r_{div} + 1 - r_{div})^t \sum_{k=0}^{N} k \cdot r_{i\ell} l_{\ell k}}{(1+r_{div})^t},$$

Proof. It follows directly from Equations (4.11) and (4.20): since we are considering the sum of the first $N+1$ components of the i^{th}-row of matrix $\mathcal{M}^t$, this is the sum of the elements of the i^{th}-row of matrix RD^tL.

□

It is obvious from Proposition 4.3.12 that this quantity only depends on the initial affinity with the target trait, the transition probability matrix $\mathcal{Q}_N$ and the division rate r_{div}. The average affinity within the GC does not depend on $\overline{a}_s$ (as one can clearly see in Figure 4.2 (a)), nor by r_s or r_d. One can intuitively understand this behavior: independently from their fitness, all B-cells submitted to mutation exit the GC. Moreover, r_s and r_d impact the GC size, but not its average affinity, as selection and death affect all individuals of the GC independently from their fitness.

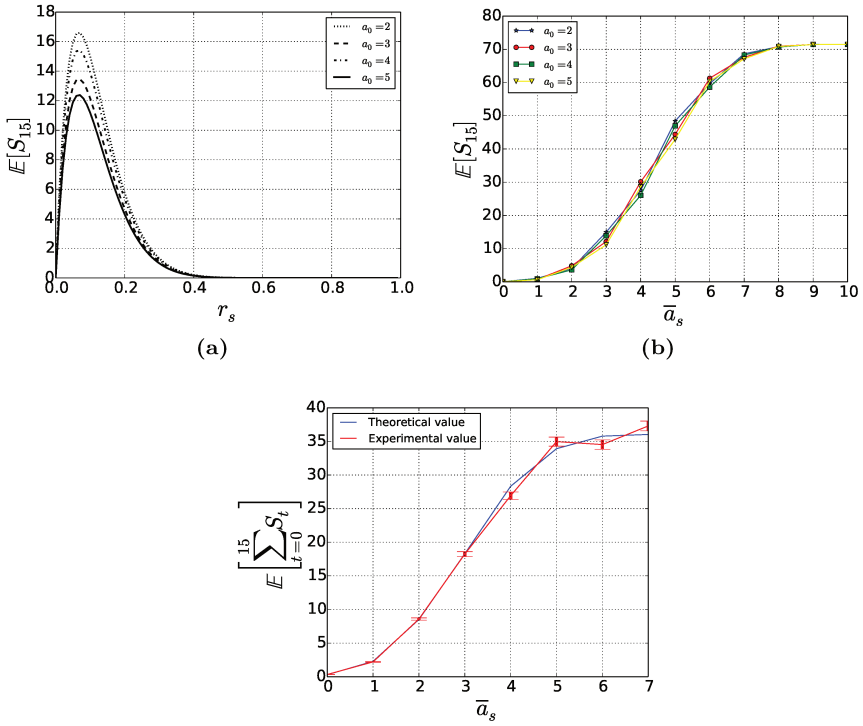
It can be interesting to observe the evolution of the expected average affinity within the GC during time. Simulations shows that the expected average affinity in the GC converges through $N/2$, independently from the affinity of the first naive B-cell (Figure 4.2 (b)). This depends on the mutational model we choose for these simulations. Indeed, providing that the GC is in a situation of explosion, for t big enough the distribution of GC clones within the affinity classes is governed by the stationary distribution of matrix $\mathcal{Q}_N$. Since for $\mathcal{Q}_N$ given by Definition 4.7 one can prove that the stationary distribution over $\{0, \dots, N\}$ is the binomial probability distribution (Chapter 2), the average affinity within the GC will quickly stabilizes at a value of $N/2$.

Evolution of the selected pool

The evolution of the number of selected B-cells during time necessarily depends on the evolution of the GC. In particular, let us suppose we are in the supercritical case, *i.e.* the extinction probability of the GC is strictly smaller than 1. Than, with positive probability, the GC explodes and so does the selected pool. On the other hand, if the GC extinguishes, the number of selected B-cells will stabilize at a constant value, as once a B-cell is selected it can only stay unchanged in the selected pool.

As already mentioned in Section 4.3.3, there exists an optimal value of the parameter r_s which maximizes the expected number of selected B-cells at time t. Figure 4.3 (a) evidences this fact. Moreover, as expected, simulations show that the expected size of selected B-cells at a given time t increases with the threshold $\overline{a}_s$ chosen for positive selection (Figure 4.3 (b)). This is a consequence of Proposition 4.3.10: $\overline{a}_s$ determines the number of elements of the sum $\sum_{k=0}^{\overline{a}_s} r_{i\ell} l_{\ell k}$.

Figure 4.3 (c) underlines the correspondence between theoretical results given by Proposition 4.3.9 and numerical values obtained by simulating the



(a) (b)

(c)

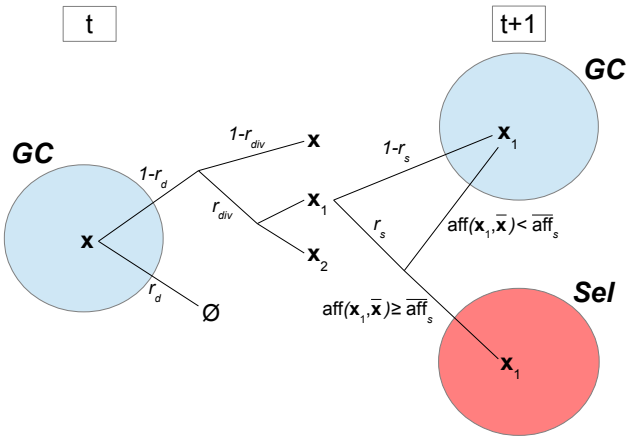
Figure 4.3: (a-b) Expected number of selected B-cells for the time step $t = 15$ for different values of a_0, depending on r_s and $\overline{a}_s$ respectively. There exists an optimal value of r_s maximizing the expected number of selected B-cells for a given time step. This value is independent from a_0. (c) Comparison between the expected number of selected B-cells until time t given by the theoretical formula (Equation (4.13)), and the experimental value obtained as the mean over 4000 simulations. Vertical bars denotes the corresponding estimated standard deviations. Here $N = 7$ and $r_s = 0.3$.

evolutionary process described by Definition 4.4. In particular Figure 4.3 (c) shows the expected (resp. average) number of selected B-cells produced until time $t = 15$ depending on the threshold chosen for positive selection, $\overline{a}_s$.

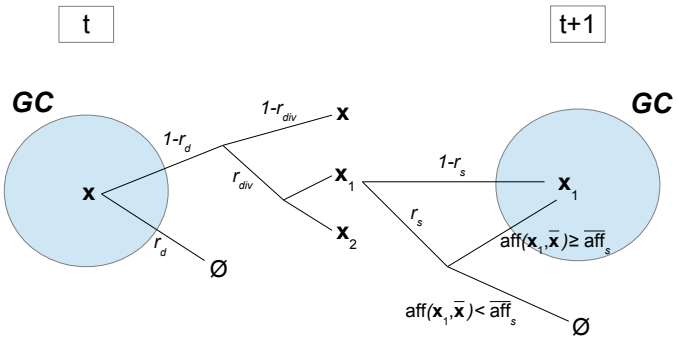
4.4 Extensions of the model

Proceeding as in Section 4.3.2, we can define and study many different models of affinity-dependent selection. Here we propose a model in which we perform only positive selection and a model reflecting a Darwinian evolutionary system, in which the selection is only negative. For the latter, we will take into account only $N + 2$ types instead of $N + 3$: we do not have to consider a selected

pool. Indeed the selected population remains in the GC. Here below we give the definitions of both models. In Section 4.4.1 we formalize these problems mathematically, then in Section 4.4.2 we show some numerical results.



(a) Positive selection



(b) Negative selection

Figure 4.4: Schematic representations of models described (a) by Definitions 4.17 and (b) by Definitions 4.18 of exclusively positive (resp. exclusively negative) selection.

4.4.1 Definitions and results

Let us consider the process described in Definition 4.4. We change only the selection mechanism.

Definition 4.17 (Positive selection)**.** If a B-cell submitted to selection belongs to an affinity class with index greater than $\overline{a}_s$, nothing happens. Otherwise, the B-cell exits the GC pool and reaches the selected pool.

Definition 4.18 (Negative selection)**.** If a B-cell submitted to selection belongs to an affinity class with index greater than $\overline{a}_s$, it dies. Otherwise, nothing happens.

In Figure 4.4 we represent schematically both processes of positive selection and of negative selection. It is clear from Figure 4.4 (b) that in the case of Definition 4.18 we do not need to consider the selected pool anymore.

Positive selection

Definition 4.19. Let $\mathbf{Z}_t^{+(\mathbf{i})} = (Z_{t,0}^{+\ (\mathbf{i})}, \ldots, Z_{t,N+2}^{+\ (\mathbf{i})})$, $t \geq 0$ be a MC where for all $0 \leq j \leq N$, $Z_{t,j}^{+\ (\mathbf{i})}$ describes the number of GC B-cells belonging to the j^{th}-affinity class with respect to $\overline{\mathbf{x}}$, $Z_{t,N+1}^{+\ (\mathbf{i})}$ the number of selected B-cells and $Z_{t,N+2}^{+\ (\mathbf{i})}$ the number of dead B-cells at generation t, when the process is initiated in state $\mathbf{i} = (i_0, \ldots, i_N, 0, 0)$, and following the evolutionary model described by Definition 4.17.

Let us denote by $\mathcal{M}^+ = (m_{ij}^+)_{0 \leq i,j \leq N+2}$ the matrix containing the expected number of type-j offsprings of a type-i cell corresponding to the model defined by Definition 4.17. We can explicitly write the value of all m_{ij}^+ depending on r_d, r_{div}, r_s, and the elements of matrix $\mathcal{Q}_N$.

Proposition 4.4.1. *$\mathcal{M}^+$ is a $(N+3)^2$ matrix, which we can define as a block matrix in the following way:*

$$\mathcal{M}^+ = \begin{pmatrix} \mathcal{M}_1^+ & \mathcal{M}_2^+ \\ \mathbf{0}_{2\times(N+1)} & \mathcal{I}_2 \end{pmatrix}$$

Where:

- $\mathcal{M}_1^+ = (m_{1,ij}^+)$ *is a* $(N+1)^2$ *matrix. For all* $i \in \{0, \ldots, N\}$*:*
  - $\forall\, j \leq \overline{a}_s$*:* $m_{1,ij}^+ = 2(1-r_d)r_{div}(1-r_s)q_{ij} + (1-r_d)(1-r_{div})(1-r_s)\delta_{ij}$
  - $\forall\, j > \overline{a}_s$*:* $m_{1,ij}^+ = 2(1-r_d)r_{div}q_{ij} + (1-r_d)(1-r_{div})\delta_{ij}$

  where δ_{ij} *is the Kronecker delta.*

- $\mathcal{M}_2^+ = (m_{2,ij}^+)$ *is a* $(N+1) \times 2$ *matrix where for all* $i \in \{0, \dots, N\}$ $m_{2,i1}^+ = m_{2,i1}$, *and* $m_{2,i2}^+ = r_d$. *We recall that* $m_{2,i1}$ *is the* i^{th}*-component of the first column of matrix* $\mathcal{M}_2$, *given in Proposition 4.3.4.*

Negative selection

Definition 4.20. Let $\mathbf{Z}_t^{-\,(\mathbf{i})} = (Z_{t,0}^{-\,(\mathbf{i})}, \dots, Z_{t,N+1}^{-\,(\mathbf{i})})$, $t \geq 0$ be a MC where for all $0 \leq j \leq N$, $Z_{t,j}^{-\,(\mathbf{i})}$ describes the number of GC B-cells belonging to the j^{th}-affinity class with respect to $\overline{\mathbf{x}}$ and $Z_{t,N+1}^{-\,(\mathbf{i})}$ the number of dead B-cells at generation t, when the process is initiated in state $\mathbf{i} = (i_0, \dots, i_N, 0, 0)$, and following the evolutionary model described by 4.18.

Let us denote by $\mathcal{M}^- = (m_{ij}^-)_{0 \leq i,j \leq N+1}$ the matrix containing the expected number of type-j offsprings of a type-i cell corresponding to the model defined by Definition 4.20.

Proposition 4.4.2. $\mathcal{M}^-$ *is a* $(N+2)^2$ *matrix, which we can define as a block matrix in the following way:*

$$\mathcal{M}^- = \begin{pmatrix} \mathcal{M}_1^- & \mathbf{m}_2^- \\ \mathbf{0}'_{N+1} & 1 \end{pmatrix}$$

Where:

- $\mathcal{M}_1^- = (m_{1,ij}^-)$ *is a* $(N+1)^2$ *matrix. For all* $i \in \{0, \dots, N\}$*:*
  - $\forall\, j \leq \overline{a}_s$*:* $m_{1,ij}^- = 2(1-r_d)r_{div}q_{ij} + (1-r_d)(1-r_{div})\delta_{ij}$
  - $\forall\, j > \overline{a}_s$*:* $m_{1,ij}^- = 2(1-r_d)r_{div}(1-r_s)q_{ij} + (1-r_d)(1-r_{div})(1-r_s)\delta_{ij}$
- $\mathbf{m}_2^-$ *is a* $(N+1)$ *column vector s.t. for all* $i \in \{0, \dots, N\}$ $m_i^+ = m_{2,i2}$, $m_{2,i2}$ *being the* i^{th}*-component of the second column of matrix* $\mathcal{M}_2$, *given in Proposition 4.3.4.*
- $\mathbf{0}'_{N+1}$ *is a* $(N+1)$ *row vector composing of zeros.*

We do not prove Propositions 4.4.1 and 4.4.2, since the proofs are the same as for Proposition 4.3.4.

Results stated in Proposition 4.3.9 hold true for these new models, by simply replacing matrix $\mathcal{M}$ with $\mathcal{M}^+$ (resp. $\mathcal{M}^-$). Of course, in the case of negative selection, as we do not consider the selected pool, we only refer to (4.10) and (4.11) quatifying the growth and average affinity of the GC. Matrix $\widetilde{\mathcal{M}}$ is the same for both models as only selection principles change.

Because of peculiar structures of matrices $\mathcal{M}^+$ and $\mathcal{M}^-$, we are not able to compute explicitly their spectra. Henceforth we can not give an explicit formula for the extinction probability or evaluate the optimal values of the selection rate r_s as we did in Sections 4.3.2 and 4.3.3.

Nevertheless, by using standard arguments for positive matrices, the greatest eigenvalue of both matrices $\mathcal{M}_1^+$ and $\mathcal{M}_1^-$ can be bounded, and hence give sufficient conditions for extinction.

Proposition 4.4.3. *Let $\mathbf{q}^+$ (resp. $\mathbf{q}^-$) be the extinction probability of the GC for the model corresponding to matrix $\mathcal{M}_1^+$ (resp. $\mathcal{M}_1^-$).*

- *If $r_{div} \leq \dfrac{r_d}{1-r_d}$, then $\mathbf{q}^+ = \mathbf{q}^- = \mathbf{1}$.*
- *If $r_s < 1 - \dfrac{1}{(1-r_d)(1+r_{div})}$, then $\mathbf{q}^+ < \mathbf{1}$ and $\mathbf{q}^- < \mathbf{1}$.*

Proof. Since both matrices $\mathcal{M}_1^+$ and $\mathcal{M}_1^-$ are strictly positive matrices (Definition 4.15), the Perron Frobenius Theorem insures that the spectral radius is also the greatest eigenvalue. Then the following classical result holds [99]:

Theorem 4.4.4. *Let $A = (a_{ij})$ be a square nonnegative matrix with spectral radius $\rho(A)$ and let $r_i(A)$ denote the sum of the elements along the i^{th}-row of A. Then:*

$$\min_i r_i(A) \leq \rho(A) \leq \max_i r_i(A)$$

Simple calculations provide:

$$\min_i r_i(\mathcal{M}_1^+) = (1-r_d)(1+r_{div}) - r_s(1-r_d)\left(2r_{div}\min_i \sum_{j=0}^{\overline{a}_s} q_{ij} + 1 - r_{div}\right)$$

$$\max_i r_i(\mathcal{M}_1^+) = (1-r_d)(1+r_{div}) - 2r_s r_{div}(1-r_d)\max_i \sum_{j=0}^{\overline{a}_s} q_{ij}$$

$$\min_i r_i(\mathcal{M}_1^-) = (1-r_d)(1+r_{div}) - r_s(1-r_d)\left(2r_{div}\min_i \sum_{j=\overline{a}_s+1}^{N} q_{ij} + 1 - r_{div}\right)$$

$$\max_i r_i(\mathcal{M}_1^-) = (1-r_d)(1+r_{div}) - 2r_s r_{div}(1-r_d)\max_i \sum_{j=\overline{a}_s+1}^{N} q_{ij}$$

The result follows by observing that for all $i \in \{0,\dots,N\}$, $0 \leq \sum_{j=0}^{\overline{a}_s} q_{ij}$, $\sum_{j=\overline{a}_s+1}^{N} q_{ij} \leq 1$, and applying Theorem 4.3.5.

□

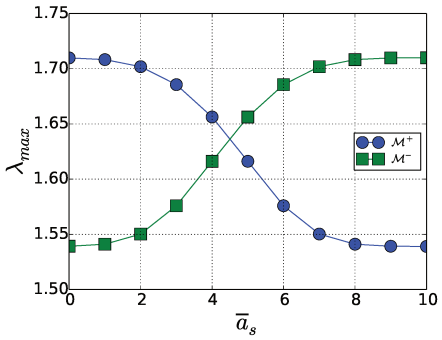


Figure 4.5: Dependence of greater eigenvalues of matrices $\mathcal{M}^+$ (blue circles) and $\mathcal{M}^-$ (green squares) respectively on $\overline{a}_s$ for $N = 10$, $r_{div} = 0.9$, $r_d = r_s = 0.1$. Hence $(1 - r_d)(1 + r_{div})(1 - r_s) = 1.539$ and $(1 - r_d)(1 + r_{div}) = 1.71$.

Remark 32. One can intuitively obtain the second claim of Proposition 4.4.3, as this condition over the parameters implies that the probability of extinction of the GC for the model underlined by $\mathcal{M}_1$ (Proposition 4.3.7) of positive and negative selection is strictly smaller than 1. Indeed keeping the same parameters for all models, the size of the GC for the model of positive and negative selection is smaller than the size of GCs corresponding to both models of only positive and only negative selection. Consequently if the GC corresponding to $\mathcal{M}$ has a positive probability of explosion, it will be necessarily the same for $\mathcal{M}^+$ and $\mathcal{M}^-$.

Remark 33. The values of both $\rho(\mathcal{M}_1^+)$ and $\rho(\mathcal{M}_1^-)$ depend on $\overline{a}_s$, varying from a minimum of $(1 - r_d)(1 + r_{div})(1 - r_s)$ and a maximum of $(1 - r_d)(1 + r_{div})$. Figure 4.5 evidences the dependence on $\overline{a}_s$ of the spectral radius of $\mathcal{M}_1^+$ and $\mathcal{M}_1^-$, using matrix $\mathcal{Q}_N$ given by Definition 4.7 as transition probability matrix.

Remark 33 and Figure 4.5 evidences that, conversely to the previous case of positive and negative selection, in both cases of exclusively positive (resp. exclusively negative) selection the parameter $\overline{a}_s$ plays an important role in the GC dynamics, affecting its extinction probability. In particular, keeping unchanged all other parameters, if $\overline{a}_s \to N$ (resp. $\overline{a}_s \to 0$), then $\rho(\mathcal{M}_1^+)$ (resp. $\rho(\mathcal{M}_1^-)$) $\to (1 - r_d)(1 + r_{div})(1 - r_s)$, which implies $\mathbf{q}^+$ (resp. $\mathbf{q}^-$) $\to \mathbf{1}$.

4.4.2 Numerical simulations

The evolution of GCs corresponding to matrices $\mathcal{M}^+$ and $\mathcal{M}^-$ respectively are complementary. Moreover, in both cases, keeping all parameters fixed one expects a faster expansion if compared to the model of positive and negative selection, since the selection acts only positively (resp. negatively) on good (resp. bad) clones. In particular, the model of negative selection corresponds to the

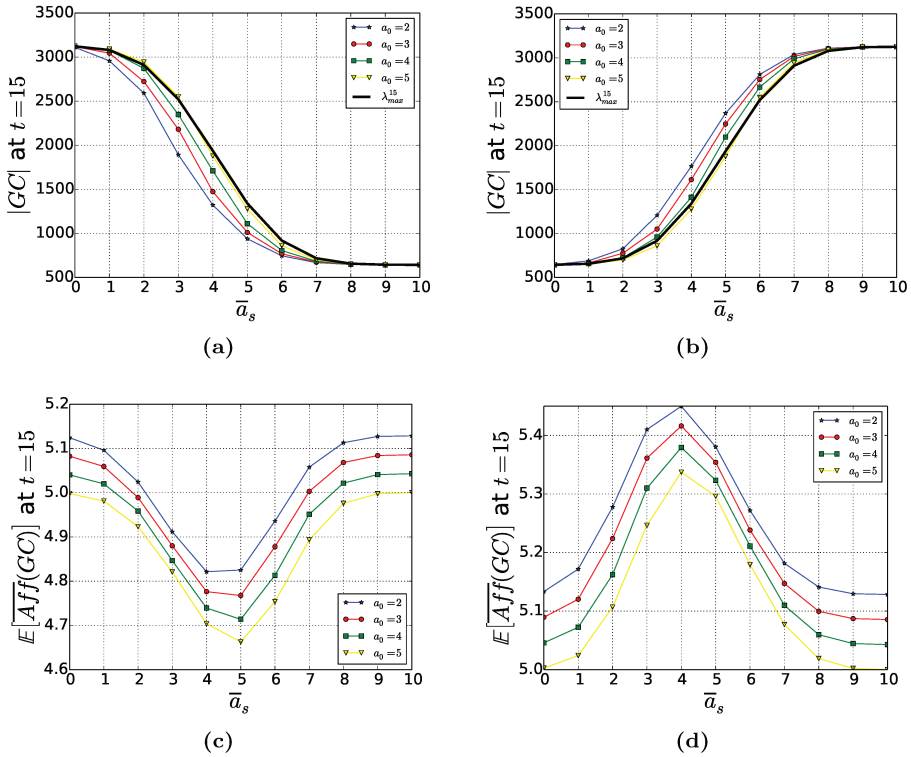
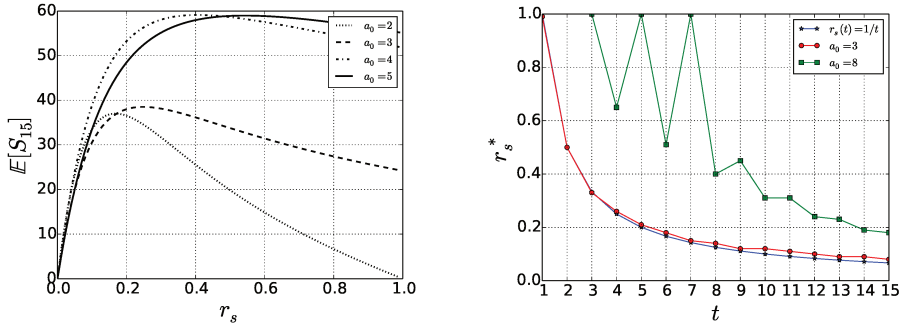


Figure 4.6: (a,b) Dependence of the expected size of the GC after 15 time steps on $\overline{a}_s$ for different values of a_0. The thick black line corresponds in both figures to the value of the greater eigenvalue of matrices $\mathcal{M}_1^+$ and $\mathcal{M}_1^-$ respectively, raised to the power of $t = 15$ (see Figure 4.5). Note that thanks to Corollary 4.3.3 we know that for this parameter choice the expected size of the GC for the model of positive and negative selection corresponds to $((1 - r_d)(1 + r_{div})(1 - r_s))^{15}$, which is equivalently λ_{max}^{15} for $\overline{a}_s = 10$ in Figure 4.6 (a) or λ_{max}^{15} for $\overline{a}_s = 0$ in Figure 4.6 (b). (c,d) Dependence of the expected average affinity in the GC after $t = 15$ time steps on $\overline{a}_s$ for different values of a_0. The left column of Figure 4.6 refer to the model of positive selection, while the right column to the model of negative selection.

case of 100% of recycling, meaning that all positively selected B-cells stay in the GC for further rounds of mutation, division and selection.

Figure 4.6 shows the dependence on $\overline{a}_s$ of the GC size and fitness, comparing $\mathcal{M}^+$ (left column) and $\mathcal{M}^-$ (right column). Indeed, for these models the GC depends on the selection threshold, conversely to the previous case of positive and negative selection, and not only on the selection rate. The effects of $\overline{a}_s$ on the GC are perfectly symmetric: it is interesting to observe that when both selection mechanisms are coupled, then $\overline{a}_s$ does not affect the GC dynamics

anymore, as shown for instance in Figure 4.2 (a). Moreover, Figures 4.6 (c,d) evidence the existence of a value of $\overline{a}_s$ that minimizes (resp. maximizes) the expected average affinity in the GC for $\mathcal{M}^+$ (resp. $\mathcal{M}^-$). In both cases this value is approximately $N/2$. This probably depends on the transition probability matrix chosen for the mutational model, which converges to a binomial probability distribution over $\{0, \ldots, N\}$.

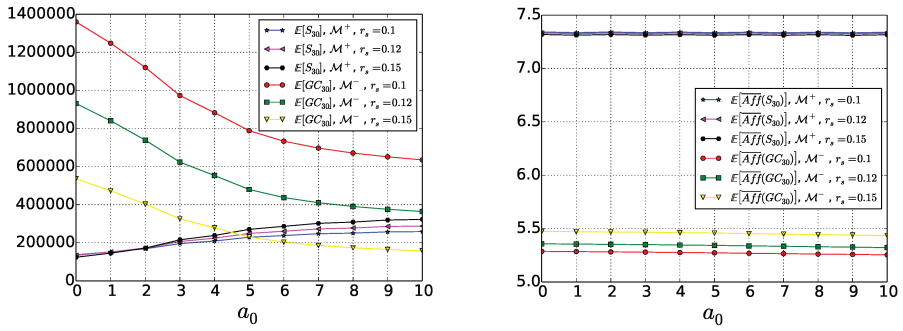


(a) **(b)**

Figure 4.7: Model of positive selection. (a) Expected number of selected B-cells for the time step $t = 15$ for different values of a_0, depending on r_s. (b) Estimation of the optimal r_s^* maximizing the expected number of selected B-cells for a given generation, comparing the model of positive selection for different values of a_0 and the model described in Section 4.3 (we plot the exact value, $r_s(t) = 1/t$, as obtained by Corollary 4.3.11). In (b), for simulations corresponding to the model of positive selection we set $\overline{a}_s = 5$.

The evolution of the selected pool for the model of positive selection have some important differences if compared to the model described in Section 4.3. For instance, it is not easy to identify an optimal value of r_s which maximizes the expected number of selected B-cells at time t. Indeed it depends both on a_0 and $\overline{a}_s$: if $a_0 \leq \overline{a}_s$ we find curves similar to those plotted in Figure 4.3 (a), otherwise Figure 4.7 (a) shows a substantial different behavior. Indeed, if $a_0 > \overline{a}_s$, choosing a big value for r_s does not negatively affect the number of selected B-cells at time t. In this case, for the first time steps no (or a very few) B-cells will be positively selected, since they still need to improve their affinity to the target. Therefore, they stay in the GC and continue to proliferate for next generations. This fact is further underlined in Figure 4.7 (b), where we estimate numerically the optimal r_s^* which maximizes the expected number of selected B-cells at time t. Simulations show that for $a_0 \leq \overline{a}_s$ the value of r_s^* for the model of positive selection is really close to the one obtained by Corollary 4.3.11. On the other hand if we start from an initial affinity class $a_0 > \overline{a}_s$ the

result we obtain is substantially different from the previous one, especially for small t. Moreover we observe important oscillations, which are probably due to the mutational model, and to the fact that the total GC size is still small for small t, since the process starts from a single B-cell. Nevertheless, it seems that for t big enough also in this case the value of r_s^* tends to approach $1/t$.



(a) **(b)**

Figure 4.8: (a) Expected number of B-cells which have been selected until time $t = 30$ for $\mathcal{M}^+$ compared to the expected size of the GC for $\mathcal{M}^-$ for different values of r_s. (b) Expected corresponding average affinity for the selected pool (case of positive selection) and the GC (case of negative selection). For some choice of the parameter r_s, the size of the selected pool for $\mathcal{M}^+$, and the GC for $\mathcal{M}^-$, are comparable. Nevertheless, the corresponding average affinities are significantly different.

Since in the case of negative selection there is no selected pool, one can suppose that at a given time t the process stops and all clones in the GC pool exit the GC as selected clones. Hence it can be interesting to compare the selected pool of the model of positive selection and the GC pool of the model of negative selection at time t. Clearly to make these two compartments comparable, the main parameters of both systems have to be opportunely chosen. In Figure 4.8 we compare the size and average fitness of the selected pool for $\mathcal{M}^+$ and the GC for $\mathcal{M}^-$ at time $t = 30$. We test different values of the parameter r_s. In particular, we observe that increasing r_s the GC size for the model of negative selection decreases and its average fitness increases. For the parameter choices we made for these simulations, Figure 4.8 (a) shows that the size of the GC for $\mathcal{M}^-$ is comparable to the size of the selected pool for $\mathcal{M}^+$ at time $t = 30$ if, keeping all other parameters fixed, $r_s = 0.15$ for $\mathcal{M}^-$. Nevertheless, this does not implies a comparable value for the average affinity: the clones of the selected pool for $\mathcal{M}^+$ have a significantly greater average affinity than those of the GC for $\mathcal{M}^-$. In order to increase the average fitness in the GC for the model of

negative selection one has to consider greater values for the parameter r_s, but this affects the probability of extinction of the process.

We can expect this discrepancy between the average affinity for the selected pool for $\mathcal{M}^+$ and the one of the GC for $\mathcal{M}^-$. Indeed, in the first case we are looking to all those B-cells which have been positive selected, hence belong at most to the $\overline{a}_s^{\text{th}}$-affinity class. On the contrary in the case of $\mathcal{M}^-$, we consider the average affinity of all B-cells which are still alive in the GC at a given time step. Among these clones, if $r_s < 1$, with positive probability there are also individuals with affinity smaller than the one required for escaping negative selection, which remain in the GC because they have not been submitted to selection. These B-cells make the average affinity decrease. Of course r_s is not the only parameter affecting the quantities plotted in Figure 4.8. In particular, one can observe that choosing a greater value for $\overline{a}_s$ also have a significant effect over the growth of both pools, as discussed in Remark 33.

4.5 Conclusions and perspectives

In this Chapter we formalize and analyze a mathematical model describing an evolutionary process with affinity-dependent selection. We use a multi-type GW process, obtaining a discrete-time probabilistic model, which includes division, mutation, death and selection. In the main model developed here, we chose a selection mechanism which acts both positively and negatively on individuals submitted to selection. This leads to build matrix $\mathcal{M}$, which contains the expectations of each type (Proposition 4.3.4) and enables to describe the average behavior of all components of the process. Moreover, thanks to the spectral decomposition of $\mathcal{M}$ we were able to obtain explicitly some formulas giving the expected dynamics of all types. In addition, we exhibited an optimal value of the selection rate maximizing the expected number of selected clones for the t^{th}-generation (Corollary 4.3.11).

This is one possible choice of the selection mechanism. From a mathematical point of view, the matrix $\mathcal{M}$ is particularly easy to manipulate, as we can obtain explicitly its spectra. On the other hand, the positive and negative selection model leads, for example, to a selection threshold that does not have any impact on the evolution of the GC size. From a biological point of view this seems counterintuitive, since we could expect that the GC dynamics is sensible to the minimal fitness required for positive selection. Moreover, this process does not take into account any recycling mechanism, which has been confirmed by experiments [139] and which improves GCs' efficiency. In addition, we considered

that only the selection mechanism is affinity dependent, while in the GC reaction other mechanisms, such as the death and proliferation rate, may depend on fitness [55, 5]. Of course it is possible to define models with affinity-dependent division and death mechanisms with our formalism. This would clearly lead to a more complicated model, which can be at least studied numerically.

The mathematical tools used in Section 4.3 can be applied to define and study other selection mechanisms. For instance in Section 4.4 we propose two variants of the model analyzed in Section 4.3, in which selection acts only positively, resp. only negatively. This Section shows how our mathematical environment can be modified to describe different selection mechanisms, which can be studied at least numerically. Moreover, it gives a deeper insight of the previous model of positive and negative selection, by highlighting the effects of each selection mechanism individually, when they are not coupled.

From a biological viewpoint there exist many possibilities to improve the models proposed in this Chapter. First of all it is extremely important to fix the system parameters, which have to be consistent with the real biological process. The choice of N defines the number of affinity level with respect to a given antigen. This value can be interpreted in different ways. On the one hand it can correspond to the number of key mutations observed during the process of Antigen Affinity Maturation, hence be even smaller than 10. On the other hand, each mutational event implies a change in the B-cell affinity, slight or not if it is a key mutation. In this case the affinity can be modeled as a continuous function, hence N corresponds to a possible discretization [143, 146]. To this choice corresponds an appropriate choice of the transition probability matrix defining the mutational model over the affinity classes, $\mathcal{Q}_N$. In most numerical simulations we set $N = 10$, which is a sensible value since experimentalists observe that high-affinity B-cells differ in their BCR coding gene by about 9 mutations from germline genes [64, 148]. Nevertheless all mathematical results are independent from this choice and hold true for all $N \geq 1$. The selection, division and death rates have also an important impact in the GC and selected pool dynamics: in the simulations we set them in order to be in a case of explosion of the GC hence appreciate the effects of all parameters over the main quantities, but they are not biologically justified. For instance, if we suppose that a single time step corresponds to one day, then the typical proliferation rate of a B-cell has been estimated between 2 and 4 per day and in the literature we found B-cell death rates of the order of 0.5-0.8 per day [96, 148, 77].

In Section 4.3.3 we have explicitly determined the optimal value of the selec-

tion rate maximizing the production of output cells at time t for the main model of positive and negative selection. It is equal to $1/t$ independently from all other parameters. Moreover, numerical estimations we made in Section 4.4.2 for the model of positive selection suggest that also in this case there exists an optimal value of $r_s(t)$, which tends to $1/t$ at least for t big enough. One has to interpret this result as the ideal optimal strength of the selection pressure to obtain a peak of the GC production of output cells at a given time step. For example, let us suppose that a time step corresponds to 1 day. The peak of the GC reaction has been measured to be close to day 12 [144]: for the kind of models we built and analyzed in this Chapter, a constant selection pressure of $1/12$ assures that the production of plasma and memory B-cells at the GC peak is maximized.

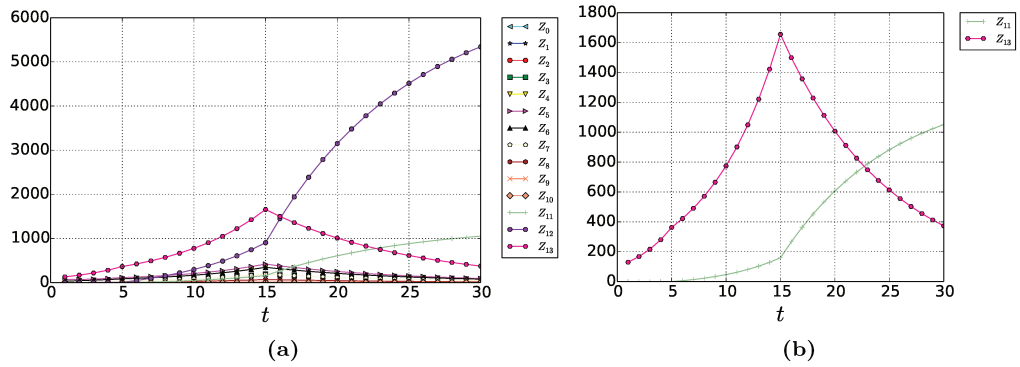


Figure 4.9: (a) Evolution during time of the expected values of all types for the model of positive and negative selection, with r_s varying during time and $N = 10$. In particular we set $r_s = 0$ until $t = 5$, $r_s = 0.1$ from $t = 6$ to $t = 15$ and $r_s = 0.3$ from $t = 16$ to $t = 30$. Z_{13} denotes the total size of the GC (*i.e.* $\sum_{k=0}^{N} Z_k$), and we recall that Z_{11} corresponds to selected B-cells and Z_{12} to dead B-cells. We set $r_{div} = 0.3$, $r_d = 0.005$ and $z_0 = 100$ initial naive B-cells. All initial B-cells belong to $a_0 = 5$, and the selection threshold is $\overline{a}_s = 3$. (b) Evolution during time of the expected total size of the GC and the selected pool respectively, for the same set of parameters as in Figure 4.9 (a).

In our models the selection pressure is constant. Since the optimal selection rate above depends on time, this suggests to go further in this direction. Moreover, this would allow to take into account, for instance, the early GC phase in which simple clonal expansion of B-cells with no selection occurs [36]. The hypothesis of a selection pressure changing over time can be easily integrated in our model. Indeed let us suppose that we fix a selection rate $r_{s,1}$ until time t_1 and $r_{s,2}$ for all $t > t_1$. Then starting from the initial condition $\mathbf{i}$ the expectations of each type at time t are given by $(\mathbf{i}\mathcal{M}^t_{r_{s,1}})$ if $t \leq t_1$ and $(\mathbf{i}\mathcal{M}^{t_1}_{r_{s,1}}\mathcal{M}^{t-t_1}_{r_{s,2}})$

if $t > t_1$, where $\mathcal{M}_{r_{s,i}}$ is the matrix containing the expectations of each type for an evolutionary process with constant selection rate $r_{s,i}$, $i = 1, 2$. In Figure 4.9 we plot the expected evolution during time of all types considering an increasing selection rate. We evaluate the expectations of all types following a process with positive and negative selection. We set $r_s = 0$ until $t = 5$, $r_s = 0.1$ from $t = 6$ to $t = 15$ and $r_s = 0.3$ for $t > 15$. Numerical simulations show that a time dependent selection rate allows initial explosion of the GC, and then progressive extinction, while when parameters are fixed, a GW process gives only rise either to explosion or to extinction, as shown above. The regulation and termination of the GC reaction has not yet been fully understood. In the literature, an increasing differentiation rate of the GC B-cells is thought to be a good explanation [100], here we show that other reasons could be of importance as well. Similarly, we can let other parameters vary for fixed time intervals, as well as decide to alternatively switch on and off the mutation mechanism, as already proposed in [108]. This can be obtained by alternatively use the identity matrix in place of $\mathcal{Q}_N$.

Chapter 5

Discussion

The aim of the work developed in this report is to introduce a very flexible mathematical environment which could be variously modified in order to pattern and study different mutation-division-selection processes. We want to contribute to the mathematical foundations of AAM, a key process in adaptive immunity. AAM produces high-affinity antibodies against immunizing antigens through iterative rounds of SHM, clonal expansion and selection for improved affinity. We enrich the model adding further fundamental bricks, which we analyse using probabilistic tools and numerical simulations. Although the evolutionary model we consider is highly simplified, it already leads to interesting mathematical problems, which we rigorously analyze in Chapters 2-4. Of course it is possible to argue many modeling assumptions and envisage improvements in order to make these models more coherent with the biological process under consideration.

In Chapter 2, we introduce and analyze several mutational processes, seen as RWs on graphs. Each mutation rule defines a specific graph. For each graph we compute the characteristic time-scales of the state-space exploration. This characterizes the efficiency of these mutational processes modeling SHM in AAM.

We define the state-space of B-cell traits as the set of N-length binary strings. From one side this assumption is justified as these two amino acid classes could represent amino acids positively charged and negatively charged. These are effectively the most responsible amino-acids in creating the non-covalent bonds which determine the antigen-antibody interaction. Nevertheless it implies a great simplification and in other papers (*e.g.* [108, 101]) models with an alphabet of 3 or more amino acids have already been proposed.

In Chapter 2 and 3 we model the BCR-antigen interaction as a linear contact between BCR and antigen representing strings. This allows us to solve the problem of defining the affinity between BCR and antigen. We are aware that the effects of genetic mutations on the new generated protein could be even more complex. It could be interesting to consider the creation of bonds among amino-acids of the BCR (resp. the antigen) itself, which determines the geometrical structure of the corresponding proteins and consequently the portion of the BCR and the antigen that can actually be in contact. To consider the tridimensional contact between two proteins is a really hard challenge and would lead us to another class of very interesting and complicated mathematical problems [23].

We define the affinity between strings in the most natural way through the Hamming distance. Other definitions of affinity are often constructed as functions of the state-space distance, given for instance by the Gaussian probability density function (*e.g.* [92]). Nevertheless in our models the choice of the affinity function does not have any influence on results. Indeed in Chapters 2 and 3, the graph structures reflecting the mutational rules are not predefined and the RWs (resp. BRWs) we perform on them are not biased by the affinity gradient. Moreover, in Chapter 4, we simply refer to affinity classes without specifying how the affinity between the antigen and B-cells belonging to the same affinity class are evaluated.

In both Chapter 2 and 3 we essentially consider mutational processes given by combinations of single point mutation mechanisms. SHM introduces mostly single nucleotide exchanges, together with small deletions and duplications, *i.e.* the insertion of extra copies of a portion of genetic material already present within the DNA code [63, 26, 27]. Allowing for indels mutations has two main consequences. Firstly it means that the length of the BCR representing string could actually change during the process, while we consider it as constant and equal to the length of the antigen representing string. We overcome this problem considering that the chain in our model corresponds to a portion of BCR in contact with the antigen, and this is approximately composed by 15 amino-acids [80]. Moreover these mutations can imply substantial changes into the amino-acid chain, enabling for long range connections in the BCR state-space. Therefore, even if these are rare mutational events, they may have an important effect in AAM and consequently it could be interesting to take also insertions and deletions into account. Another possibility is to consider that mutations at one site are influenced by other amino acids composing the string. This assumption has been firstly proposed in a highly theoretical context by S. A. Kauffman and E. D. Weinberger in [70], where they have introduced the NK

models. More recently Y. Elhanati *et al* in [45] have found biological evidence for an evolutionary model where substitution rates strictly depend on the context. Nevertheless, they only consider SHM events at the DNA level, without taking into account the effects of nucleotide substitutions on the expressed BCR and its affinity for the target antigen.

In Chapter 3 we enrich the previously analyzed mutational models by considering the division of B-cell clones. This allows to evaluate the efficiency of different mutational rules in determining the variety of the repertoire of an exponentially growing B-cell population. We observe that strong expansion properties of the graph characterizing the mutational mechanism, enable a faster invasion of the state-space. From a biological viewpoint, this property is significant since it ensures that starting from a few seeder B-cells, the GC can produce, hence test a huge variety of BCRs against the target antigen. Indeed, GCs seem to be oligoclonal [81, 88], which means that they develop from very few initial naive B-cells. Therefore, starting from a single clonal population, it is of interest to understand how a B-cells population invades the BCR state-space.

We show that if we simply consider the expansion properties of the structure built over the BCR state-space, the covering in $\mathcal{O}(N)$ is limited at a half the state-space. This suggests that the expansion property is not enough to insure a quick covering of a large portion of the state-space: considering self-avoiding BRWs on connected graphs could be more efficient, although these are not necessarily good expanders. On the other hand, from a biological point of view, it may not be so efficient to explore the whole state-space, but rather to steer mutations toward a specific region of the state-space with the best affinity. Indeed, the production of new clones has a cost in terms of time and energy, therefore it does not make sense to produce a huge variety of cells with any possible fitness with the presented antigen. It is for this reason and since SHMs are random events, that during the GCR B-cells are submitted to powerful selection mechanisms.

We discuss the consequences of defining an affinity-dependent division rate. We show that this allows to privilege individuals with good fitness. Another possibility is to consider transition probability matrices whose stationary distributions are concentrated on a specific region of the state-space containing the fittest traits. Indeed, we prove that, without any biasing mechanism, the distribution of traits for a 2-BRW only depends on the stationary distribution of the transition probability matrix under consideration.

Another way to drive mutations towards a specific region of the state-space is, of course, the introduction of a selection mechanism, which we investigate in Chapter 4. There we introduce and analyze some variants of an evolutionary model including mutation, division and affinity-dependent selection, based on the assumption that all B-cell traits can be classified into some affinity classes with respect to their binding abilities for the immunizing antigen. We use multi-type Galton Watson processes modeling the evolution of each affinity class of the B-cell population, together with dead and differentiated B-cells. In addition, we exhibited an optimal value of the selection rate maximizing the expected number of selected clones for the t^{th}-generation.

The mathematical tools used in Chapter 4 can be applied to define and study other selection mechanisms. From a biological viewpoint there exist many possibilities to improve these models. First of all it is extremely important to fix the system parameters, which have to be consistent with the real biological process. The choice of N defines the number of affinity levels with respect to a given antigen. This value can be interpreted in different ways. On the one hand it may correspond to the number of key mutations observed during the process of AAM, hence be even smaller than 10. On the other hand, each mutational event implies a change in the B-cell affinity, slight or not if it is a key mutation. In this case the affinity can be modeled as a continuous function [98], hence N corresponds to a possible discretization [143, 146]. To this choice corresponds an appropriate choice of the transition probability matrix defining the mutational model over the affinity classes. In most numerical simulations we set $N = 10$, which is a sensible value since experimentalists observe that high-affinity B-cells differ in their BCR coding gene by about 9 mutations from germline genes [64, 148].

The selection, division and death rates have also an important impact on the GC and selected pool dynamics. In the simulations we fix the parameters such that the GC's population grows exponentially, this is not biologically sound for the whole GC duration. The typical proliferation rate of a B-cell has been estimated between 2 and 4 per day and in the literature we found B-cell death rates of the order of 0.5-0.8 per day [96, 148, 77]. Depending on the selection strength we can obtain either explosion or extinction of the GC. It can be interesting to determine a reasonable parameter choice for our model to observe *e.g.* a realistic evolution of the GC size.

In the models set and studied here, all rates are kept constant during time: this implies that we shall observe either explosion or extinction of the GC only.

It is of course mandatory to allow one ore more parameters be time dependent. For instance, letting the selection pressure increase during time would account for the early GC phase in which simple clonal expansion of B-cells with no selection occurs [36]. Moreover, in the literature, an increasing differentiation rate of the GC B-cells is thought to be a good explanation for GC termination [100]. The hypothesis of a selection pressure changing over time can be easily integrated in our model. Similarly, we can let other parameters vary for fixed time intervals, as well as decide to alternatively switch on and off the mutation mechanism, as already proposed in [108].

We do not include in our models all details and biological facts discussed above, since the aim of this project was not to build a comprehensive model of AAM. Our objective is to simplify this learning evolutionary process focusing on its fundamental features and be able to provide a rigorous mathematical analysis. Hence our results remain theoretical by means of a high simplification of the biological process under examination. Nevertheless all biologically motivated improvements proposed here can be included within our models and analyzed numerically, even if they sometimes depend on experimental data which is still hard to gather. Another essential specificity of this work is that it is based on probabilistic models including B-cell traits and the evolution their affinity due to mutations. Most of the models of GCR that have been proposed in the literature are based on ODE systems (*e.g.* [77, 100]). The deterministic continuum approach has certainly many advantages, but it is not able to capture the stochastic fluctuations of reactions nor take into account the discrete nature of cells.

It is possible to add further bricks to our models and enrich them in many directions. For example, since both the selection and death rates have an impact on the regulation of the GC reaction, we can define models in which in a single time step a B-cell can undergo only one among these two mechanisms. This could be studied in a similar way as in Chapter 4. Another possibility is to increase the size of the matrix containing the average behavior of each type and define types which can only proliferate and mutate and types which can only be submitted to selection. This would allows us to take into account the compartmentalization of the GC in DZ and LZ, in which B-cells undergo distinct genetic programs. These improvements are matters of forthcoming works.

Bibliography

[1] Abul K Abbas, Andrew HH Lichtman, and Shiv Pillai. *Basic immunology: functions and disorders of the immune system*. Elsevier Health Sciences, 2012.

[2] Uwe Aickelin, Dipankar Dasgupta, and Feng Gu. Artificial immune systems. In *Search Methodologies*, pages 187–211. Springer, 2014.

[3] Sergio A Albeverio, LV Bogachev, SA Molchanov, and EB Yarovaya. *Annealed moment Lyapunov exponents for a branching random walk in a homogeneous random branching environment*. Universität Bonn. SFB 256. Nichtlineare Partielle Differentialgleichungen, 2000.

[4] David Aldous and Jim Fill. Reversible markov chains and random walks on graphs, 2002.

[5] Shannon M Anderson, Ashraf Khalil, Mohamed Uduman, Uri Hershberg, Yoram Louzoun, Ann M Haberman, Steven H Kleinstein, and Mark J Shlomchik. Taking advantage: high-affinity b cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *The Journal of Immunology*, 183(11):7314–7325, 2009.

[6] Hifzur Rahman Ansari and Gajendra PS Raghava. Identification of conformational b-cell epitopes in an antigen from its primary sequence. *Immunome research*, 6(1):1, 2010.

[7] Krishna B Athreya and Peter E Ney. *Branching processes*, volume 196. Springer Science & Business Media, 2012.

[8] Chen Avin and Carlos Brito. Efficient and robust query processing in dynamic environments using random walk techniques. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 277–286. ACM, 2004.

[9] Thomas Bäck. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press, 1996.

[10] Irene Balelli, Vuk Milisic, and Gilles Wainrib. Random walks on binary strings applied to the somatic hypermutation of b-cells. *arXiv preprint arXiv:1501.07806*, 2015.

[11] Irene Balelli, Vuk Milisic, and Gilles Wainrib. Branching random walks on binary strings for evolutionary processes. *arXiv preprint arXiv:1607.00927*, 2016.

[12] Irene Balelli, Vuk Milišić, and Gilles Wainrib. Multi-type galton-watson processes with affinity-dependent selection applied to antibody affinity maturation. *arXiv preprint arXiv:1609.00823*, 2016.

[13] Frank Ball and Peter Donnelly. Strong approximations for epidemic models. *Stochastic processes and their applications*, 55(1):1–21, 1995.

[14] Oliver Bannard, Robert M Horton, Christopher DC Allen, Jinping An, Takashi Nagasawa, and Jason G Cyster. Germinal center centroblasts transition to a centrocyte phenotype according to a timed program and depend on the dark zone for effective selection. *Immunity*, 39(5):912–924, 2013.

[15] Burton E Barnett, Maria L Ciocca, Radhika Goenka, Lisa G Barnett, Junmin Wu, Terri M Laufer, Janis K Burkhardt, Michael P Cancro, and Steven L Reiner. Asymmetric b cell division in the germinal center reaction. *Science*, 335(6066):342–344, 2012.

[16] Richard Bellman and Theodore Harris. On age-dependent binary branching processes. *Annals of Mathematics*, pages 280–295, 1952.

[17] Micah J Benson, Loren D Erickson, Michael W Gleeson, and Randolph J Noelle. Affinity of antigen encounter and other early b-cell signals determine b-cell fate. *Current opinion in immunology*, 19(3):275–280, 2007.

[18] Nathanaël Edouard Berestycki. *Phase transitions for the distance of random walks with applications to genome rearrangements*. PhD thesis, Cornell University, 2005.

[19] Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.

[20] Daniela Bertacchi and Fabio Zucca. Critical behaviors and critical values of branching random walks on multigraphs. *J.Appl.Prob.*, 45(2):481–497, 2008.

[21] Daniela Bertacchi and Fabio Zucca. Characterization of critical values of branching random walks on weighted graphs through infinite-type branching processes. *Journal of statistical physics*, 134(1):53–65, 2009.

[22] Eva Besmer, Polyxeni Gourzi, and F Nina Papavasiliou. The regulation of somatic hypermutation. *Current opinion in immunology*, 16(2):241–245, 2004.

[23] Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, et al. Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research*, page gku340, 2014.

[24] Laure Bidou, Valérie Allamand, Jean-Pierre Rousset, and Olivier Namy. Sense from nonsense: therapies for premature stop codon diseases. *Trends in molecular medicine*, 18(11):679–688, 2012.

[25] S Binitha and S Siva Sathya. A survey of bio inspired optimization algorithms. *International Journal of Soft Computing and Engineering*, 2(2):137–151, 2012.

[26] Peter M Bowers, Petra Verdino, Zhengyuan Wang, Jean da Silva Correia, Mark Chhoa, Griffin Macondray, Minjee Do, Tamlyn Y Neben, Robert A Horlick, Robyn L Stanfield, et al. Nucleotide insertions and deletions complement point mutations to massively expand the diversity created by somatic hypermutation of antibodies. *Journal of Biological Chemistry*, 289(48):33557–33567, 2014.

[27] Bryan S Briney, Jordan R Willis, and JE Crowe. Location and length distribution of somatic hypermutation-associated dna insertions and deletions reveals regions of antibody structural plasticity. *Genes and immunity*, 13(7):523–529, 2012.

[28] John Cardy and Uwe C Täuber. Theory of branching and annihilating random walks. *Physical review letters*, 77(23):4780, 1996.

[29] John L Cardy and Uwe C Täuber. Field theory of branching and annihilating random walks. *Journal of statistical physics*, 90(1-2):1–56, 1998.

[30] Leandro N. De Castro and Fernando J. Von Zuben. Learning and optimization using the clonal selection principle. *Evolutionary Computation, IEEE Transactions on*, 6(3):239–251, 2002.

[31] Nicholas Chiorazzi, Kanti R Rai, and Manlio Ferrarini. Chronic lymphocytic leukemia. *New England Journal of Medicine*, 352(8):804–815, 2005.

[32] Sarah Cobey, Patrick Wilson, and Frederick A Matsen. The evolution within us. *Phil. Trans. R. Soc. B*, 370(1676):20140235, 2015.

[33] Colin Cooper, Tomasz Radzik, and Nicolas Rivera. The coalescing-branching random walk on expanders and the dual epidemic process. *arXiv preprint arXiv:1602.05768*, 2016.

[34] Andrew Currin, Neil Swainston, Philip J Day, and Douglas B Kell. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chemical Society Reviews*, 44(5):1172–1239, 2015.

[35] Dragos Cvetkovic, Michael Doob, and Horst Sachs. Spectra of graphs-theory and applications, iii revised and enlarged edition. *Johan Ambrosius Bart Verlag, Heidelberg-Leipzig*, 1995.

[36] Nilushi S De Silva and Ulf Klein. Dynamics of b cells in germinal centres. *Nature Reviews Immunology*, 15(3):137–148, 2015.

[37] J. Di Noia and M. S. Neuberger. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature*, 419(6902):43–48, Sep 2002.

[38] Javier M Di Noia and Michael S Neuberger. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.*, 76:1–22, 2007.

[39] Persi Diaconis, Ronald L. Graham, and John A. Morrison. Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Structures & Algorithms*, 1(1):51–72, 1990.

[40] G Dighiero and TJ Hamblin. Chronic lymphocytic leukaemia. *The Lancet*, 371(9617):1017–1029, 2008.

[41] Peter G Doyle and J Laurie Snell. Random walks and electric networks. *AMC*, 10:12, 1984.

[42] Deborah K Dunn-Walters, Alex Belelovsky, Hanna Edelman, Monica Banerjee, and Ramit Mehr. The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees. *Clinical and Developmental Immunology*, 9(4):233–243, 2002.

[43] Chinmoy Dutta, Gopal Pandurangan, Rajmohan Rajaraman, and Scott Roche. Coalescing-branching random walks on graphs, 2013.

[44] B Eichhorst, T Robak, E Montserrat, Paolo Ghia, P Hillmen, M Hallek, and C Buske. Chronic lymphocytic leukaemia: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 26(suppl 5):v78–v84, 2015.

[45] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. Inferring processes underlying b-cell repertoire diversity. *Phil. Trans. R. Soc. B*, 370(1676):20140243, 2015.

[46] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[47] Warren J Ewens. Mathematical population genetics. i. theoretical introduction. interdisciplinary applied mathematics, 27, 2004.

[48] Jose Faro and Michal Or-Guil. How oligoclonal are germinal centers? a new method for estimating clonal diversity from immunohistological sections. *BMC bioinformatics*, 14(Suppl 6):S8, 2013.

[49] Marc Thilo Figge. Stochastic discrete event simulation of germinal center reactions. *Physical Review E*, 71(5):051907, 2005.

[50] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[51] Stanley F. Florkowski. Spectral graph theory of the hypercube. Master's thesis, Naval Postgraduate School, Monterey, California, 2008.

[52] Robert E Smith Stephanie Forrest and Alan S Perelson. Population diversity in an immune system model: Implications for genetic search. *Foundations of Genetic Algorithms 1993 (FOGA 2)*, 2:153, 2014.

[53] A Franklin, PJ Milburn, RV Blanden, and EJ Steele. Human dna polymerase-eta an at mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase (vol 82, pg 219, 2004). *IMMUNOLOGY AND CELL BIOLOGY*, 82(3):342–342, 2004.

[54] Simon DW Frost, Ben Murrell, AS Md Mukarram Hossain, Gregg J Silverman, and Sergei L Kosakovsky Pond. Assigning and visualizing germline genes in antibody repertoires. *Phil. Trans. R. Soc. B*, 370(1676):20140240, 2015.

[55] Alexander D Gitlin, Ziv Shulman, and Michel C Nussenzweig. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature*, 2014.

[56] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[57] John Burdon Sanderson Haldane. The cost of natural selection. *Journal of Genetics*, 55(3):511–524, 1957.

[58] Frank Harary, John P Hayes, and Horng-Jyh Wu. A survey of the theory of hypercube graphs. *Computers & Mathematics with Applications*, 15(4):277–289, 1988.

[59] Theodore E Harris. *The theory of branching processes*. Springer-Verlag, 1963.

[60] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[61] Kenneth B Hoehn, Anna Fowler, Gerton Lunter, and Oliver G Pybus. The diversity and molecular evolution of b cell receptors during infection. *Molecular biology and evolution*, page msw015, 2016.

[62] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[63] Joyce K Hwang, Frederick W Alt, and Leng-Siew Yeap. Related mechanisms of antibody somatic hypermutation and class switch recombination. *Microbiology spectrum*, 3(1), 2015.

[64] Dagmar Iber and Philip K. Maini. A mathematical model for germinal centre kinetics and affinity maturation. *Journal of theoretical biology*, 219(2):153–175, 2002.

[65] Nobuyuki Ikeda, Masao Nagasawa, Shinzo Watanabe, et al. Branching markov processes i. *Journal of Mathematics of Kyoto University*, 8(2):233–278, 1968.

[66] Nobuyuki Ikeda, Masao Nagasawa, Shinzo Watanabe, et al. Branching markov processes ii. *Journal of Mathematics of Kyoto University*, 8(3):365–410, 1968.

[67] Nobuyuki Ikeda, Masao Nagasawa, Shinzo Watanabe, et al. Branching markov processes iii. *Journal of Mathematics of Kyoto University*, 9(1):95–160, 1969.

[68] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[69] Christel Kamp and Stefan Bornholdt. Coevolution of quasispecies: B-cell mutation rates maximize viral error catastrophes. *Physical Review Letters*, 88(6):068104, 2002.

[70] Stuart A Kauffman and Edward D Weinberger. The nk model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, 141(2):211–245, 1989.

[71] Celia Keim, David Kazadi, Gerson Rothschild, and Uttiya Basu. Regulation of AID, the B-cell genome mutator. *Genes Dev.*, 27(1):1–17, Jan 2013.

[72] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 482–491. IEEE, 2003.

[73] David G Kendall. On the generalized "birth-and-death" process. *The annals of mathematical statistics*, pages 1–15, 1948.

[74] Thomas B Kepler, Supriya Munshaw, Kevin Wiehe, Ruijun Zhang, Jae-Sung Yu, Christopher W Woods, Thomas N Denny, Georgia D Tomaras, S Munir Alam, M Anthony Moody, et al. Reconstructing a b-cell clonal lineage. ii. mutation, selection, and affinity maturation. *Immune system modeling and analysis*, page 83, 2015.

[75] Thomas B Kepler and Alan S Perelson. Cyclic re-entry of germinal center b cells and the efficiency of affinity maturation. *Immunology today*, 14(8):412–415, 1993.

[76] Thomas B Kepler and Alan S Perelson. Somatic hypermutation in b cells: an optimal control treatment. *Journal of theoretical biology*, 164(1):37–64, 1993.

[77] Can Keşmir and Rob J De Boer. A mathematical model on germinal center kinetics and termination. *The Journal of Immunology*, 163(5):2463–2469, 1999.

[78] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202. ACM, 2009.

[79] Jens Vindahl Kringelum, Claus Lundegaard, Ole Lund, and Morten Nielsen. Reliable b cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol*, 8(12):e1002829, 2012.

[80] Jens Vindahl Kringelum, Morten Nielsen, Søren Berg Padkjær, and Ole Lund. Structural analysis of b-cell epitopes in antibody: protein complexes. *Molecular immunology*, 53(1):24–34, 2013.

[81] Frans GM Kroese, Auk S Wubbena, Hendrik G Seijen, and Paul Nieuwenhuis. Germinal centers develop oligoclonally. *European journal of immunology*, 17(7):1069–1072, 1987.

[82] Hari Krovi and Todd A. Brun. Hitting time for quantum walks on the hypercube. *Physical Review A*, 73(3):032341, 2006.

[83] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. Amer Mathematical Society, 2009.

[84] Dan R Littman. Releasing the brakes on cancer immunotherapy. *Cell*, 162(6):1186–1190, 2015.

[85] László Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

[86] FP Machado and S Yu Popov. Branching random walk in random environment on trees. *Stochastic processes and their applications*, 106(1):95–106, 2003.

[87] Karen J Mackenzie, Paul M Fitch, Melanie D Leech, Anne Ilchmann, Claire Wilson, Amanda J McFarlane, Sarah EM Howie, Stephen M Anderton, and Jürgen Schwarze. Combination peptide immunotherapy based on t-cell epitope mapping reduces allergen-specific ige and eosinophilia in allergic airway inflammation. *Immunology*, 138(3):258–268, 2013.

[88] Ian CM MacLennan. Germinal centers. *Annual review of immunology*, 12(1):117–139, 1994.

[89] Ian CM MacLennan, Carola García de Vinuesa, and Montserrat Casamayor-Palleja. B-cell memory and the persistence of antibody responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 355(1395):345–350, 2000.

[90] Joaquín Marro and Ronald Dickman. *Nonequilibrium phase transitions in lattice models*. Cambridge University Press, 2005.

[91] Connor O McCoy, Trevor Bedford, Vladimir N Minin, Philip Bradley, Harlan Robins, and Frederick A Matsen. Quantifying evolutionary constraints on b-cell affinity maturation. *Phil. Trans. R. Soc. B*, 370(1676):20140244, 2015.

[92] Michael Meyer-Hermann. A mathematical model for the germinal center morphology and affinity maturation. *Journal of theoretical Biology*, 216(3):273–300, 2002.

[93] Michael Meyer-Hermann. A concerted action of b cell selection mechanisms. *Advances in Complex Systems*, 10(04):557–580, 2007.

[94] Michael Meyer-Hermann, Elodie Mohr, Nadége Pelletier, Yang Zhang, Gabriel D Victora, and Kai-Michael Toellner. A theory of germinal center b cell selection, division, and exit. *Cell reports*, 2(1):162–174, 2012.

[95] Michael E Meyer-Hermann and Philip K Maini. Cutting edge: back to ?one-way? germinal centers. *The Journal of Immunology*, 174(5):2489–2493, 2005.

[96] Michael E Meyer-Hermann, Philip K Maini, and Dagmar Iber. An analysis of b cell selection mechanisms in germinal centers. *Mathematical Medicine and Biology*, 23(3):255–277, 2006.

[97] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, 2009.

[98] Vuk Milisic and Gilles Wainrib. Mathematical modeling of lymphocytes selection in the germinal center. *arXiv preprint arXiv:1501.06725*, 2015.

[99] H Minc. Nonnegative matrices, 1988, 1988.

[100] Joana S Moreira and Jose Faro. Modelling two possible mechanisms for the regulation of the germinal center dynamics. *The Journal of Immunology*, 177(6):3705–3710, 2006.

[101] Enrique Muñoz and Michael W Deem. Amino acid alphabet size in protein evolution experiments: better to search a small library thoroughly or a large library sparsely? *Protein Engineering Design and Selection*, 21(5):311–317, 2008.

[102] Kenneth M. Murphy, Paul Travers, Mark Walport, et al. *Janeway's immunobiology*, volume 7. Garland Science New York, NY, USA, 2012.

[103] MS Neuberger. Antibodies: a paradigm for the evolution of molecular recognition. *Biochemical Society Transactions*, 30(4):341–350, 2002.

[104] James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[105] Mihaela Oprea and Alan S Perelson. Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts. *The Journal of Immunology*, 158(11):5155–5162, 1997.

[106] Michal Or-Guil and Jose Faro. A major hindrance in antibody affinity maturation investigation: We never succeeded in falsifying the hypothesis of single-step selection. *Frontiers in Immunology*, 5, 2014.

[107] Wei Pang, Kangping Wang, Yan Wang, Ge Ou, Hanbing Li, and Lan Huang. Clonal selection algorithm for solving permutation optimisation problems: A case study of travelling salesman problem. In *International Conference on Logistics Engineering, Management and Computer Science (LEMCS 2015)*. Atlantis Press, 2015.

[108] Alan S. Perelson and Gérard Weisbuch. Immunology for physicists. *Reviews of modern physics*, 69(4):1219–1267, 1997.

[109] Tri Giang Phan, Didrik Paus, Tyani D Chan, Marian L Turner, Stephen L Nutt, Antony Basten, and Robert Brink. High affinity germinal center b cells are actively selected into the plasma cell compartment. *The Journal of experimental medicine*, 203(11):2419–2424, 2006.

[110] Roybel R Ramiscal and Carola G Vinuesa. T-cell subsets in the germinal center. *Immunological reviews*, 252(1):146–155, 2013.

[111] L Chris G Rogers and David Williams. *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*. Cambridge university press, 2000.

[112] Igor B Rogozin and Marilyn Diaz. Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step

activation-induced cytidine deaminase-triggered process. *J. Immunol.*, 172(6):3382–3384, Mar 2004.

[113] Steven A Rosenberg and Nicholas P Restifo. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science*, 348(6230):62–68, 2015.

[114] Jimmy Salvatore. Bipartite graphs and problem solving. *University of Chicago*, 2007.

[115] Huseyin Saribasak and Patricia J Gearhart. Does dna repair occur during somatic hypermutation? In *Seminars in immunology*, volume 24, pages 287–292. Elsevier, 2012.

[116] Stanley Sawyer. Branching diffusion processes in population genetics. *Advances in Applied Probability*, pages 659–689, 1976.

[117] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[118] Ton N Schumacher and Robert D Schreiber. Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74, 2015.

[119] Tanja A Schwickert, Gabriel D Victora, David R Fooksman, Alice O Kamphorst, Monica R Mugnier, Alexander D Gitlin, Michael L Dustin, and Michel C Nussenzweig. A dynamic t cell–limited checkpoint regulates affinity-dependent b cell entry into the germinal center. *The Journal of experimental medicine*, 208(6):1243–1252, 2011.

[120] Roger Sciammas, Ying Li, Aryeh Warmflash, Yiqiang Song, Aaron R Dinner, and Harinder Singh. An incoherent regulatory network architecture that orchestrates b cell diversification in response to antigen signaling. *Molecular systems biology*, 7(1):495, 2011.

[121] Michele Shannon and Ramit Mehr. Reconciling repertoire shift with affinity maturation: the role of deleterious mutations. *The Journal of Immunology*, 162(7):3950–3956, 1999.

[122] Wen-Jun Shen, Hau-San Wong, Quan-Wu Xiao, Xin Guo, and Stephen Smale. Towards a mathematical foundation of immunology and amino acid chains. *arXiv preprint arXiv:1205.6031*, 2012.

[123] MJ Shlomchik, P Watts, MG Weigert, and S Litwin. Clone: a monte-carlo computer simulation of b cell clonal expansion, somatic mutation, and antigen-driven selection. In *Somatic Diversification of Immune Responses*, pages 173–197. Springer, 1998.

[124] Ziv Shulman, Alexander D Gitlin, Sasha Targ, Mila Jankovic, Giulia Pasqual, Michel C Nussenzweig, and Gabriel D Victora. T follicular helper cell dynamics in germinal centers. *Science*, 341(6146):673–677, 2013.

[125] Ann Smith. Nucleic acids to amino acids: Dna specifies protein. *Nature Education*, 1(1):126, 2008.

[126] Lauren Sompayrac. *How the immune system works*. Wiley-Blackwell, 2012.

[127] Joel NH Stern, Kevin C O'Connor, David A Hafler, Uri Laserson, Francois Vigneault, and Steven H Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Immune system modeling and analysis*, page 55, 2015.

[128] György Szabó. Branching annihilating random walk on random regular graphs. *Physical Review E*, 62(5):7474, 2000.

[129] R Michael Tanner. Explicit concentrators from generalized n-gons. *SIAM Journal on Algebraic Discrete Methods*, 5(3):287–293, 1984.

[130] David M Tarlinton and Kenneth GC Smith. Dissecting affinity maturation: a model explaining selection of antibody-forming cells and memory b cells in the germinal centre. *Immunology today*, 21(9):436–441, 2000.

[131] J. M. Tas, L. Mesin, G. Pasqual, S. Targ, J. T. Jacobsen, Y. M. Mano, C. S. Chen, J. C. Weill, C. A. Reynaud, E. P. Browne, M. Meyer-Hermann, and G. D. Victora. Visualizing antibody affinity maturation in germinal centers. *Science*, 351(6277):1048–1054, Mar 2016.

[132] Jeroen MJ Tas, Luka Mesin, Giulia Pasqual, Sasha Targ, Johanne T Jacobsen, Yasuko M Mano, Casie S Chen, Jean-Claude Weill, Claude-Agnès Reynaud, Edward P Browne, et al. Visualizing antibody affinity maturation in germinal centers. *Science*, 351(6277):1048–1054, 2016.

[133] Grace Teng and F Nina Papavasiliou. Immunoglobulin somatic hypermutation. *Annu. Rev. Genet.*, 41:107–120, 2007.

[134] Jonathan Timmis, Andrew Hone, Thomas Stibor, and Edward Clark. Theoretical advances in artificial immune systems. *Theoretical Computer Science*, 403(1):11–32, 2008.

[135] Susumu Tonegawa. Somatic generation of immune diversity. *Bioscience reports*, 8(1):3–26, 1988.

[136] Gabriel D Victora. Snapshot: the germinal center reaction. *Cell*, 159(3):700–700, 2014.

[137] Gabriel D Victora and Luka Mesin. Clonal and cellular dynamics in germinal centers. *Current opinion in immunology*, 28:90–96, 2014.

[138] Gabriel D Victora and Michel C Nussenzweig. Germinal centers. *Annual review of immunology*, 30:429–457, 2012.

[139] Gabriel D Victora, Tanja A Schwickert, David R Fooksman, Alice O Kamphorst, Michael Meyer-Hermann, Michael L Dustin, and Michel C Nussenzweig. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell*, 143(4):592–605, 2010.

[140] Michael Voit. Asymptotic distributions for the ehrenfest urn and related random walks. *Journal of Applied Probability*, pages 340–356, 1996.

[141] Feng Wang, Shiladitya Sen, Yong Zhang, Insha Ahmad, Xueyong Zhu, Ian A Wilson, Vaughn V Smider, Thomas J Magliery, and Peter G Schultz. Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. *Proceedings of the National Academy of Sciences*, 110(11):4261–4266, 2013.

[142] Peng Wang, Chang-ming Shih, Hai Qi, and Yue-heng Lan. A stochastic model of the germinal center integrating local antigen competition, individualistic t–b interactions, and b cell receptor signaling. *The Journal of Immunology*, page 1600411, 2016.

[143] Armin A Weiser, Nicole Wittenbrink, Lei Zhang, Andrej I Schmelzer, Atijeh Valai, and Michal Or-Guil. Affinity maturation of b cells involves not only a few but a whole spectrum of relevant mutations. *International immunology*, 23(5):345–356, 2011.

[144] Ivonne Wollenberg, Ana Agua-Doce, Andrea Hernández, Catarina Almeida, Vanessa G Oliveira, Jose Faro, and Luis Graca. Regulation of the germinal center reaction by foxp3+ follicular regulatory t cells. *The Journal of Immunology*, 187(9):4553–4560, 2011.

[145] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. na, 1932.

[146] Huafeng Xu, Aaron G Schmidt, Timothy O'Donnell, Matthew D Therkelsen, Thomas B Kepler, M Anthony Moody, Barton F Haynes, Hua-Xin Liao, Stephen C Harrison, and David E Shaw. Key mutations stabilize

antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. *Proteins: Structure, Function, and Bioinformatics*, 83(4):771–780, 2015.

[147] Sule Yavuz, Akif S Yavuz, Kenneth H Kraemer, and Peter E Lipsky. The role of polymerase η in somatic hypermutation determined by analysis of mutations in a patient with xeroderma pigmentosum variant. *The Journal of Immunology*, 169(7):3825–3830, 2002.

[148] Jingshan Zhang and Eugene I Shakhnovich. Optimality of mutation and selection in germinal centers. *PLoS Comput Biol*, 6(6):e1000800, 2010.

List of Figures

List of Tables

Résumé

Le système immunitaire adaptatif est capable de produire une réponse spécifique contre presque tous le pathogènes qui agressent notre organisme. Ceci est du aux anticorps qui sont des protéines sécrétées par les cellules B. Les molécules qui provoquent cette réaction sont appelées antigènes : pendant une réponse immunitaire, les cellules B sont soumises à un processus d'apprentissage afin d'améliorer leur capacité à reconnaître un antigène donné. Ce processus est appelé maturation d'affinité des anticorps.

Nous établissons un cadre mathématique très flexible dans lequel nous définissons et étudions des modèles évolutionnaires simplifiés inspirés par la maturation d'affinité des anticorps. Nous identifions les éléments constitutifs fondamentaux de ce mécanisme d'évolution extrêmement rapide et efficace : mutation, division et sélection. En commençant par une analyse rigoureuse du mécanisme de mutation dans le Chapitre 2, nous procédons à l'enrichissement progressif du modèle en ajoutant et analysant le processus de division dans le Chapitre 3, puis des pressions sélectives dépendantes de l'affinité dans le Chapitre 4.

Notre objectif n'est pas de construire un modèle mathématique très détaillé et exhaustif de la maturation d'affinité des anticorps, mais plutôt d'enquêter sur les interactions entre mutation, division et sélection dans un contexte théorique simplifié. On cherche à comprendre comment les différents paramètres biologiques influencent la fonctionnalité du système, ainsi qu'à estimer les temps caractéristiques de l'exploration de l'espace d'états des traits des cellules B.

Au-delà des motivations biologiques de la modélisation de la maturation d'affinité des anticorps, l'analyse de ce processus d'apprentissage nous a amenée à concevoir un modèle mathématique qui peut également s'appliquer à d'autres systèmes d'évolution, mais aussi à l'étude de la propagation de rumeurs ou de virus. Notre travail théorique s'accompagne de nombreuses simulations numériques qui viennent soit l'illustrer soit montrer que certains résultats demeurent extensibles à des situations plus compliquées.

Mots clés Marches aléatoires sur des graphes, Hypercube, Temps d'attente, Marches aléatoires branchantes, Graphes expanseurs, Processus de Galton-Watson multi-type, Réaction du centre germinatif, Paysage évolutif

Abstract

The adaptive immune system is able to produce a specific response against almost any pathogen that could penetrate our organism and inflict diseases. This task is assured by the production of antigen-specific antibodies secreted by B-cells. The agents which causes this reaction are called antigens: during an immune response B-cells are submitted to a learning process in order to improve their ability to recognize the immunizing antigen. This process is called antibody affinity maturation.

We set a highly flexible mathematical environment in which we define and study simplified mathematical evolutionary models inspired by antibody affinity maturation. We identify the fundamental building blocks of this extremely efficient and rapid evolutionary mechanism: mutation, division and selection. Starting by a rigorous analysis of the mutational mechanism in Chapter 2, we proceed by successively enriching the model by adding and analyzing the division process in Chapter 3 and affinity-dependent selection pressures in Chapter 4.

Our aim is not to build a very detailed and comprehensive mathematical model of antibody affinity maturation, but rather to investigate interactions between mutation, division and selection in a simplified theoretical context. We want to understand how the different biological parameters affect the system's functionality, as well as estimate the typical timescales of the exploration of the state-space of B-cell traits.

Beyond the biological motivations of antibody affinity maturation modeling, the analysis of this learning process leads us to build a mathematical model which could be relevant to model other evolutionary systems, but also gossip or virus propagation. Our method is based on the complementarity between probabilistic tools and numerical simulations.

Keywords: Random walks on graphs, Hypercube, Hitting times, Branching random walks, Expander graphs, Multi-type Galton-Watson process, Germinal center reaction, Evolutionary landscape