FREIE UNIVERSITÄT BERLIN

Methods to detect Evolutionary Constraints: Application to HIV

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
Maureen Smith

Berlin, 2019

Erstgutachter:	Dr. Max von Kleist
Zweitgutachter:	Prof. Dr. Bernhard Renard
Drittgutachter:	Prof. Dr. Knut Reinert
Tag der Disputation:	05. September 2019

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die von mir eingereichte Dissertation selbständig verfasst habe. Alle Hilfsmittel, wie Publikationen, Bücher oder Webseiten, wurden im Literaturverzeichnis angegeben und Zitate aus fremden Arbeiten sind als solche gekennzeichnet. Diese Arbeit wurde in gleicher oder ähnlicher Form bisher in keinem anderen Promotionsverfahren eingereicht und auch nicht veröffentlicht.

Berlin, 8.4.2019

Maureen Smith

Acknowledgements

In the first instance, I wish to thank my supervisor Max for giving me the opportunity to delve into the exciting field of HIV research, putting trust in me and my work, and being a great supporter and advisor at all times.

I would like to thank Redmond and Roland from Strasbourg, for providing good ideas, fruitful discussions, fantastic collaborations, and of course precious data! I also thank our collaborators from RKI, Karo and Claudia.

I highly appreciate my most awesome office family Kaveh, Sulav, and Nadja, with whom it was a pleasure to work, talk, and laugh with, in all these years. This includes of course also Stefan R., providing the daily dose of coffee. Also to the rest of the lunch crew, Biocomputing colleagues, and friends I would like to acknowledge my gratitude: Nada, Han, Evgenia, Jonny, Iurii, Marjan, Kasia, Stefan K., Patrick, Martin, Victor, Mona, Pooja, Wei, Christian, Vikram, and Anika. You made my time in university unique and unforgettable, and I thank you for all the wonderful moments we had.

I thank my family, for being supportive, loving and always there for me. Thank you, Dani, for proofreading my thesis. Last but not least, I want to thank Mathias, not only for proofreading and many helpful discussions, but most of all for his love, support, encouragement, and patience, especially during stressful times.

Contents

1 Introduction 1

1.1 Motivation . . . 1

1.2 Structure of the Thesis . . . 2

2 Background: HIV 3

2.1 A short History Timeline of HIV . . . 3

2.2 HIV Classification . . . 4

2.3 Virus Structure . . . 5

2.3.1 Genome Organisation and Structure . . . 6

2.4 Replication Cycle . . . 8

2.4.1 Cell Entry . . . 8

2.4.2 Reverse Transcription . . . 8

2.4.3 Nuclear Import and Integration . . . 9

2.4.4 Replication and Nuclear Export . . . 10

2.4.5 Assembly and Maturation . . . 11

2.5 Treatment . . . 12

3 Background: Evolutionary Constraints 15

3.1 Fitness . . . 15

3.2 Constraints . . . 17

3.3 Epistasis . . . 18

3.3.1 Neutrality Functions . . . 18

3.3.2 Multiplicative Model . . . 19

3.3.3 Additive Model . . . 19

3.3.4 Measures of Epistasis . . . 19

Magnitude Epistasis . . . 19

Sign Epistasis . . . 20

4 Methods to infer Evolutionary Constraints 21

4.1 Representation of Biological Sequences and MSA . . . 22

4.2 Mutual Information . . . 23

4.3 Direct Coupling Analysis . . . 24

4.3.1 Derivation of the Maximum Entropy Model . . . 24

The Potts Model . . . 26

Gauge Fixing . . . 26

4.3.2 Parameter Inference . . . 27

Maxmum Likelihood . . . 27

Thermodynamic Potentials for the Inverse Potts Problem . . . 28

Mean Field Approximation . . . 31

Thouless-Anderson-Palmer Equation . . . 33

4.3.3 Scoring Schemes in DCA . . . 33

5 Quantifying Functional Constraints 35
5.1 Mutational Interference Mapping Experiment . . . 35
5.1.1 Relation between Nucleotide Frequency and Relative Kd Values . . 36
5.1.2 Error Correction . . . 37
5.1.3 Statistical Evaluation . . . 38
5.1.4 Quality Criteria . . . 40
5.1.5 Detecting Functional Structures with MIME . . . 41
5.2 Next Generation Sequencing . . . 44
5.2.1 General Prodedure . . . 44
5.2.2 Paired-End Sequencing . . . 46
5.2.3 Quality Scoring . . . 46
5.2.4 FASTQ . . . 47

6 Software 49
6.1 SAM to Counts . . . 49
6.1.1 Preprocessing: From Sequencing Output to SAM Files . . . 50
6.1.2 Implementation . . . 50
Prefilter . . . 51
Merging . . . 52
Counting . . . 52
6.2 MIMEAnTo . . . 53
6.2.1 Implementation . . . 54
Project Initialisation . . . 54
Error Estimation . . . 55
Kd Estimation and Quality Filtering . . . 56
Output . . . 58

7 MIME applied to *in cell* Experiments with the HIV-1 61
7.1 From *in vitro* to *in cell* Experiments . . . 62
7.2 Analysis of *in cell* MIME Data . . . 63
7.2.1 Relation between Nucleotide Frequencies and the Effect on intracellular gRNA Production . . . 65
7.2.2 Relation between Nucleotide Frequencies and the Effect on Packaging 66
7.2.3 Error Correction . . . 67
7.2.4 Statistical Assessment of Effects . . . 68
7.2.5 Estimation of Error Probability κ . . . 69
7.2.6 Quality Criteria . . . 69
7.3 Results . . . 70
7.3.1 Data Description . . . 70
7.3.2 Regulation of intracellular gRNA Production . . . 71
7.3.3 Regulation of gRNA Packaging . . . 73
7.3.4 Discussion . . . 75

8 Combining MIME and DCA Techniques: Accuracy Benchmark 79
8.1 Intruding Effects from Epistasis . . . 79
8.2 Potts Model applied to the MIME Framework . . . 82
8.2.1 Gauge Fixing . . . 83
8.2.2 Regularisation . . . 84
Pseudocounts . . . 84
L_1- and L_2 Regularisation . . . 85
8.2.3 Parameter Determination . . . 85

8.3 Set up the Test Data 86

8.3.1 Sampling Sequence Variants 86

8.3.2 Sampling Kds and Epistasis 87

8.3.3 Deriving Species Frequencies in Equilibrium 88

8.3.4 Adding Noise 89

8.3.5 Output 89

8.3.6 Implementation 89

8.4 Evalutating different Methods 90

8.5 Benchmark Results 91

8.5.1 Estimation of Single Site Effects 92

8.5.2 Estimation of Epistatic Effects 96

8.6 Discussion 99

9 Discussion 101

Zusammenfassung 105

Bibliography 107

Chapter 1

Introduction

1.1 Motivation

The pandemic spread of HIV started almost 30 years ago and has claimed the lives of several million people since then. The virus attacks the immune system of its host. Untreated, the HIV infection leads to drastic opportunistic diseases, advancing to AIDS after a few years. At this stage, the immune system is severely damaged and causes the decease of the infected person. Fortunately, the number of new infections per year decreases thanks to extensive HIV/AIDS-awareness campaigns and potent treatment regimen. Whilst to date no vaccination or cure for an HIV infection has been found, there exist antiviral treatments, which allow infected people to reach a close to normal life expectancy.
Still, the pandemic can not be eradicated, with annually new infections world wide running into millions, especially in developing countries [1]. The reason is, on the one hand, lacking HIV education, hence renouncing of prevention, and on the other hand, unavailable treatment options. Also in developed countries the virus continues to spread, despite of pervasive awareness- and treatment programs. The highest risk of further transmission is early after infection [2]. This can be explained with the elevated virus load in the human body in the early (acute) phase after transmission, in combination with the initial unawareness of the own infection.

Usually, medication lowers the risk of viral transmission [3]. However, HIV is a rapidly mutating virus with a fast replication cycle, which allows it to quickly adapt to selective pressure. This entails the development of resistant mutations, leading to reduced or no effect of an antiviral drug. The combination of drugs needs to be changed accordingly, to prevent treatment failure.
The situation gets critical, if multi-drug resistances occur and hardly any drug combination is working. Yet, mutational pathways of reaching resistant mutations include bottlenecks and can only be acquired with a certain genomic background [4]. It remains an ongoing challenge to learn more about the constraints for the virus to develop drug resistance.
The currently available antiviral drugs target different processes of the HIV life cycle by binding to crucial enzymes, and other involved proteins, or by competing with nucleosides to stop the reverse transcription process [5]. Recently, new therapeutic strategies are advancing: Despite the encoding for essential proteins, RNAs comprise a multitude of regulatory elements and functional structures controlling virtually all biological processes. Hence, regulating motifs in RNAs are promising drug targets, and first achievements in this direction have been attained [6, 7]. Nevertheless, processes of the regulatory mechanisms in the HIV life cycle, as well as detailed structural information of the RNA are still underexplored.

In this work, we present methods for the qualitative and quantitative inference of evolutionary constraints, including the Mutational Intereference Mapping Experiment (MIME)

[8] and direct coupling analysis (DCA) [9, 10]. Based on the former, we provide the software MIMEAnTo to predict functional elements in RNA, published in [11]. Furthermore, results of the adaption of the MIME framework to *in cell* experiments are presented, which we published in [12]. We determined regulatory motifs in the 5′ UTR of HIV-1, which are essential for viral RNA production in cells, as well as the packaging process of nascent virions, respectively. Lastly, we attempt to improve the prediction of functional regions with MIME by incorporating methods used in DCA. We set up a benchmark with different scenarios of mutational effects (disrupting function) including pairwise epistasis (evolutionary constraints), and could indeed see improvements for a number of cases. Yet, these are preliminary results, which encourage us to further address the approach in more detail.

1.2 Structure of the Thesis

Foremost, in Chapter 2, we introduce the current knowledge regarding HIV biology, treatment, and epidemiology to highlight the importance of ongoing HIV research. The subsequent Chapter 2 provides definitions of fitness, evolution, and evolutionary constraints. With this background, we describe state of the art methods to infer epistasis and evolutionary constraints in multiple sequence alignments in Chapter 4, with focus on direct coupling analysis (DCA). The following Chapter 5 covers a method to identify and quantify the commitment of positions within a functional sequence: the Mutational Interference Mapping Experiment (MIME). Based on this method, we present our implementation of the MIME data analysis in Chapter 6, and an adaption of the MIME framework for *in cell* experiments in Chapter 7. We introduce an approach combining MIME with DCA related methods in Chapter 8, with the aim to improve the prediction of mutational effects. The presented work is discussed in a conclusive summary in Chapter 9, including an outlook for future aspects.

Chapter 2

Background: HIV

In this chapter, we give an introduction of HIV and its biological background. First we explore a short timeline since its suspected origin until today and the classification of HIV strains. We look into more detail of the virus biology, including structure, genome organisation, and replication cycle, and present different treatment strategies.

2.1 A short History Timeline of HIV

It is known that the human immunodeficiency virus (HIV) has its origin in Africa and emerged by several independent cross-species transmissions of the Simian Immunodeficiency Virus (SIV) from primates to humans [13]. In 1999, Gao et al. [14] found an SIV strain in chimpanzees (SIVcpz) which was almost identical to HIV-1 and concluded that the virus may have been transmitted from chimps to humans. They also discovered that the transmittable and human infecting SIVcpz strain evolved by the junction of two different SIV strains, obtained by two different smaller monkey species eaten by the chimpanzees [15].

There are different theories of how the zoonotic transmission happened. The most commonly accepted explanation is the hunting theory: The virus could be transmitted through blood contact when the chimpanzees were hunted, killed and eaten, or kept as pets [13]. Normally, SIV would have been fought by the human immune system, but on a few occasions the virus adapted itself within its new host and evolved to HIV-1 [16]. Latest research results suggest that the virus crossed species from chimpanzees to humans in Kinshasa, the Democratic Republic of Congo, in the 1920s [17].

HIV and AIDS remained unknown until the early 1980s, as infections happen unperceivedly and without any symptoms. The virus quietly und unnoticedly reached different continents (North America, South America, Europe, Africa, and Australia) [16]. Although few cases of AIDS were retrospectively documented before, current data suggest that an epidemic spread started in the 1970s [16]. In the beginning of the 1980s the virus finally caught widespread medical awareness: Several cases of serious diseases, such as a rare lung infection called Pneumocystis carinii pneumonia (PCP), severe immune deficiency, and unusually aggressive cancer (Kaposi's Sarcoma), of previously healthy patients were linked and clustered, described by various names in different parts of the world [18, 19]. The unexplained cases were often connected to homosexual communities, and hence corresponding names emerged, including GRIDS (gay-related immunodeficiency syndrome), "gay cancer", etc. In September 1982, the Centers for Disease Control (CDC) in the USA introduced the term **AIDS (Acquired Immune Deficiency Syndrome)** for the first time, describing it as a disease causing a defect in cell mediated immunity and other opportunistic infections [20]. Soon the disease turned out to affect all people from all communities. Increasing evidence revealed that it is not only a sexually transmitted disease, but can also be forwarded by shared needles among drug users, contaminated

blood contact, and transmission from an infected mother to her child [16, 21–23].
In May 1983, the Pasteur Institute in France discovered a new retrovirus called Lymphadenopathy-Associated Virus (LAV), that was susptected to be the cause of AIDS [24]. Also the National Cancer Institute announced that they found a retrovirus related to AIDS one year later, calling it Human T-Cell Lymphotropic Virus (HTLV)-3 [25]. At a joint conference, they clarified that the two viruses are identical [26]. The International Committee on the Taxonomy of Viruses agreed in May 1986 on the official name **Human Immunodeficiency Virus (HIV)** instead of HTLV-3/LAV [27].
In 1986, a second type of the virus has been found in West Africa [28]: The rare and not pandemic strain of HIV-2 was found to be closely related to an SIV strain appearing in sooty mangabeys.

Since the start of the pandemic, HIV has infected globally an estimated number of 77.3 million people and around 35.4 million people died from AIDS [1]. Fortunately, the annual number of AIDS related deaths decreased since the peak in 2004 (1.9 million) by 51% (940K deaths in 2017) [1]. Today, estimated 36.9 million people are living with an HIV infection, yet, with a decreasing number of new infections every year [1].

2.2 HIV Classification

HIV belongs to the family of Retroviridae and is associated to the group of Lentiviruses [29, 30]. Retroviruses have genomic RNA, which is reverse transcribed into DNA and inserted into the host's DNA to be replicated [31]. Lentiviruses are causing slowly progressing, chronical degenerative diseases [30].
HIV infects primarily vital CD4$^+$ T cells in humans, helper cells of the immune system, which will be destroyed by various mechanisms. The decline of CD4$^+$ T cells to a critical level leads to failure of the cell-mediated immune system. This advance phase of HIV infection allows for the development of opportunistic infections and tumors that rarely affect people who have a functioning immune system [32]. This late stage of symptoms is referred to as acquired immunodeficiency syndrome (AIDS). However, antiretroviral treatment (ART) allows many patients to live with an HIV infection without advancing to AIDS.
Besides the CD4$^+$ T cells, HIV infects macrophages [33], and dendritic cells[34].

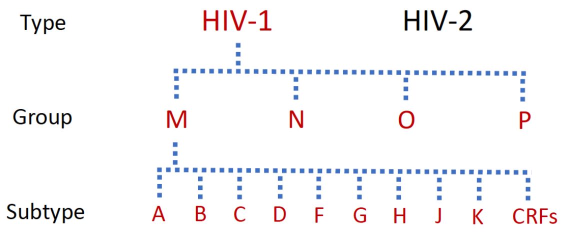


FIGURE 2.1: Types and strains of HIV.

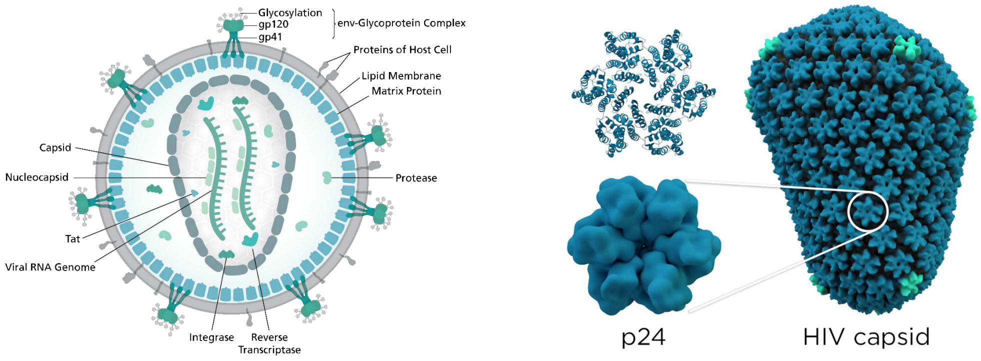
There are two types of HIV, characterised on the basis of genetic differences and its suspected pandemic origin: HIV-1, which is the most common type affecting around 95% of the infected population, and HIV-2, which is mainly found in regions of West Africa and is less virulent and infective than HIV-1 [13, 35]. HIV-1 and HIV-2 are genetically different by 55% [35]. Hence, not all detection tests for one of the types are reliable for the other one [35, 36]. Both types can be classified into distinct groups and subtypes on the basis of differences in the envelope region (cf. Section 2.3.1). This indicates that each group arose out of independent SIV transmission events from non-human primates to humans [13].
The predominant type HIV-1 comprises the following groups: Group M is the oldest, "major" strain and responsible for the majority of the global HIV pandemic. The other groups

O ("outlier"), N ("non-M or O"), and P (reported in 2009) are very rarely spotted in West-Central Africa and Cameroon repsectively [13]. Group M diversified into nine subtypes (A-D, F-H, J, K) and more than 40 different circulating recombinant forms (CRFs), based on differences in the whole genome (see Figure 2.1).

Subtype B accounts only for 12% of the world wide HIV epidemic, but is prevalent in the western countries (America, Western Europe, Australia). This fact explains why most of the clinical research for HIV and available treatments were developed against subtype B, although subtype C dominates with almost 50% of all infected people, most prevalent in Southern Africa and India [13]. Studies suggest that some of the subtypes are more virulent than others and might carry resistance against different treatments [36]. Still, antiretroviral drugs (ARV) showed efficacy on different kinds of subtypes, altough mainly developed for subtype B [37].

2.3 Virus Structure

One HIV virion measures a diameter of approximately 100 to 130 nm in its immature and 130 to 145 nm in its mature state respectively [38–40]. Only mature virions, i.e. all structural components are assembled, are able to infect new cells [41]. In Figure 2.2a the structure of the mature virion is shown.



(A) Structure of HIV-1 virion

(B) Capsid of HIV-1 virion

FIGURE 2.2: **A**: Structure of a mature HIV-1 virion. Graphic taken from [42].**B**: Structure of the capsid of HIV-1, which is composed of viral capsid proteins (p24). Graphic taken from [43].

The outer membrane or **envelope** of the virus is composed of a lipid bilayer originating from a infected host cell, acquired during the budding process (cf. Section 2.4.5). The surface is studded with viral envelope proteins (Env). Each consist of three transmembrane glycoproteins (gp41) with surface proteins gp120 [44, 45], allowing the virus to attach and fuse with target cells (cf. Section 2.4.1).

The envelope is surrounding a symmetrical **matrix** of viral structural proteins (p17), which build the inner membrane of the virus particle [46].

The conical **capsid**, which is typical for retroviruses, is composed of viral capsid proteins (p24) [47]. A close up is shown in Figure 2.2b. The capsid encloses the viral **genome**, two copies of the positive sense genomic RNA (gRNA), which are tightly attached to **nucleocapsid** proteins (p7). Furthermore, the capsid encompasses the essential **enzymes** reverse transcriptase and integrase [44].

2.3.1 Genome Organisation and Structure

The genome of HIV consists of two identical single stranded RNA molecules of roughly 10,000 nucleotides. Each of the sequences contains nine open reading frames (ORF) encoding 15 (16 for HIV-2) proteins and several structural elements, depicted in Figure 2.3.

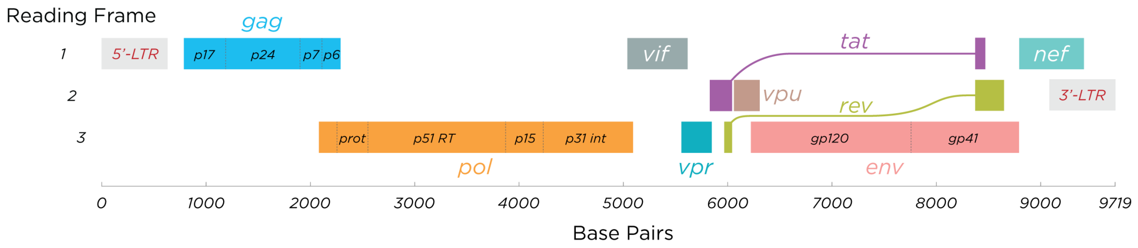


FIGURE 2.3: Genome organisation in HIV-1 on the three reading frames. Figure taken from [42].

Structural Proteins
The three polyproteins Gag, Pol and Env encoded by their respecitve genes are common for all retroviruses [48].
The ***gag*** and ***env*** genes encode **structural components** of the viral particle: The precursor Gag protein p55 is important for the viral assembly at the plasma membrane, hence its name assemblin. It is processed by the viral protease into four capsid proteins, namely matrix (MA/p17), capsid (CA/p24), nucleocapsid (NC/p7) and p6 protein, which form the core of the virion, as explained above [49]. The *env* gene encodes the glycoprotein gp160. Cleavage of this precursor results in two components for the outer membrane envelope, the surface protein (gp120) and transmembrane protein (gp41).

Enzymes and Gene Regulating Proteins
The ***pol*** region provides three essential **enzymes**: protease (PR), reverse transcriptase (RT), and integrase (IN), products of the protease processed Gag-Pol precursor protein [48, 49].
The proteins Tat and Rev are essential **gene regulating** factors. The ***tat*** gene comprises transcriptional transactivators p14 and p16, facilitating the HIV gene expression by binding to the TAR element in the 5′ long terminal repeat (LTR) [49]. The ***rev*** gene expresses a phosphoprotein, the regulator of expression of viral proteins, promoting the nuclear export of the viral RNA to the cytoplasm by binding to the Rev response element (RRE) [48, 49].

Accessory Proteins
The expressed proteins beyond these essential structural and regulatory proteins and enzymes are called **accessory proteins**. They are not necessary for viral propagation, but appear to be conserved in different strains. The genes ***vif***, ***vpr***, and ***nef*** are found in all HIV and SIV strains, whereas ***vpu*** is unique for HIV-1 and a few closely related SIV strains. A duplication of the *vpr* gene resulted in the emergence of an additional gene in HIV-2 and few SIV types [49], ***vpx***.
Vif encodes the viral infectivity factor, i.e. it is promoting the infectivity of the virus. The viral protein R (Vpr) is found inside the nucleus of the virion and targets different regulatory functions, e.g. the import of pre-integration complexes, the induction of cellular differentiation, or cell growth arrest. The viral protein U (Vpu) plays a role in degradation of CD4 in the endoplasmic reticulum, and the release of virus particles from the plasma membrane of HIV-1 infected cells [49]. The *nef* (negative factor) gene, located within the 3′ LTR, is one of the first genes to be expressed within primate lentivirus infected cells and

an important virulence factor *in vivo*, however dispensable *in vitro* [49]. The Vpr homologue Vpx in HIV-2 also improves the efficiency of lentiviral infection by influencing the replication of the virus.

In addition to genes encoding for the above described proteins, the viral genome occupies non-coding regions comprising various structural and functional elements that interact in complex ways to modulate gene expression.
After reverse transcription the viral DNA is flanked by **LTRs**, a repetitive sequence to steer the integration of the provirus into the host genome. Furthermore, the 5′ LTR contains the promotor for the entire gene expression of the virus, once the viral DNA has been integrated. The 3′ LTR contains a polyadenylation signal as well as the gene for the accessory protein Nef [50]. The gRNA of the virus is preceded by an untranslated region (**UTR**) controlling various steps of the HIV life cycle including transcription, translation, reverse transcription, export and virus packaging [51–54]. It forms different structures according to the executed function, with an example shown in Figure 2.4, building the following structural elements.

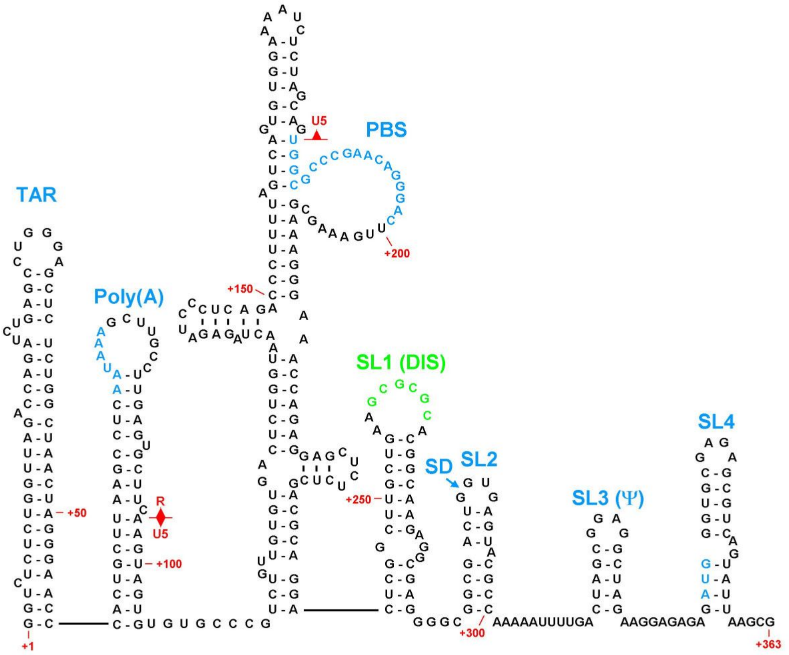


FIGURE 2.4: Structure model of the 5′ UTR in HIV-1: the TAR element, the polyA stem loop, the PBS complex, and the Psi domain including SL1 (DIS), SL2 (SD), the major packaging signal SL3, and SL4 with the gag start codon. Graphic taken from [55].

Starting from 5′ to 3′, the target sequence for viral transactivation response (**TAR**) is located in the beginning of the UTR, providing binding sites for the Tat protein and proteins of the host cell important for transcription. The region forms a hairpin loop with a bulge in the stem, which is important for its function.
It is followed by the polyadenylation site (5′ **PolyA**), which forms a stem loop and is involved in genome packaging and dimerisation [56]. The primer binding site (**PBS**) binds to a specific primer which is required for initiation of reverse transcription.
The packaging signal (**Psi**) domain is starting upstream the *gag* gene, partly overlapping with it. It consists of four stem loops (SL) which are involved in genome packaging [51, 57, 58].

The highly conserved **SL1** contains the dimer initiation site (**DIS**), a palindromic loop, which binds to a second gRNA strand to build a dimer by forming a kissing stem loop during encapsidation [59]. The **SL2** contains the splice donor (**SD**) site, involved in regulation of the splicing process to balance the level of synthesis for the viral proteins [60]. Great importance has been awarded to **SL3**, historically considered as the major packaging signal (also referred to as **Psi** element) [61–63]. However, several studies revealed a crucial role of SL1 for packaging *in vitro* [8, 12, 52, 64]. In Chapter 7, we will refer to the major packaging domain in *in cell* experiments [12], exposing additional motifs within the 5′ UTR involved in the viral packaging process. The **SL4** stem loop starts right behind the start codon for the Gag protein. Compared to the other SL elements, SL4 has a low binding affinity to the NC domain during the packaging process. Hence, the authors of [65] suggest that SL4 is involved in the genome recognition by stabilising the structure of the Psi domain.

Another important structural domain is the Rev response element (RRE) within the *env* gene. The segment is about 350 nucleotides long and binds to the Rev protein to be exported from the nucleus to the cytoplasm [66].

2.4 Replication Cycle

The replication cycle of HIV, illustrated in Figure 2.5, starts with the virion binding and fusing with the surface of the targeted cell. The pre-integration complex, containing the viral gRNA and proteins, is released into the cell. The capsule is degraded and the RNA can be transcribed by the viral reverse transcriptase into viral DNA. After transportation of the DNA inside the nucleus, it is integrated into the host genome by the enzyme integrase. The DNA of the host including the provirus is transcribed by the cell's transcription machinery. Multiple copies of the viral mRNA are formed. Some of them are translated into viral proteins and some are packed into the new emerging virus particles at the surface of the host cell, together with the translated proteins. After the release of the immature virion, the viral protease inside of the virion cleaves the packed polyproteins and a mature, infectious virus is assembled. In the following, the individual steps are explained in more detail.

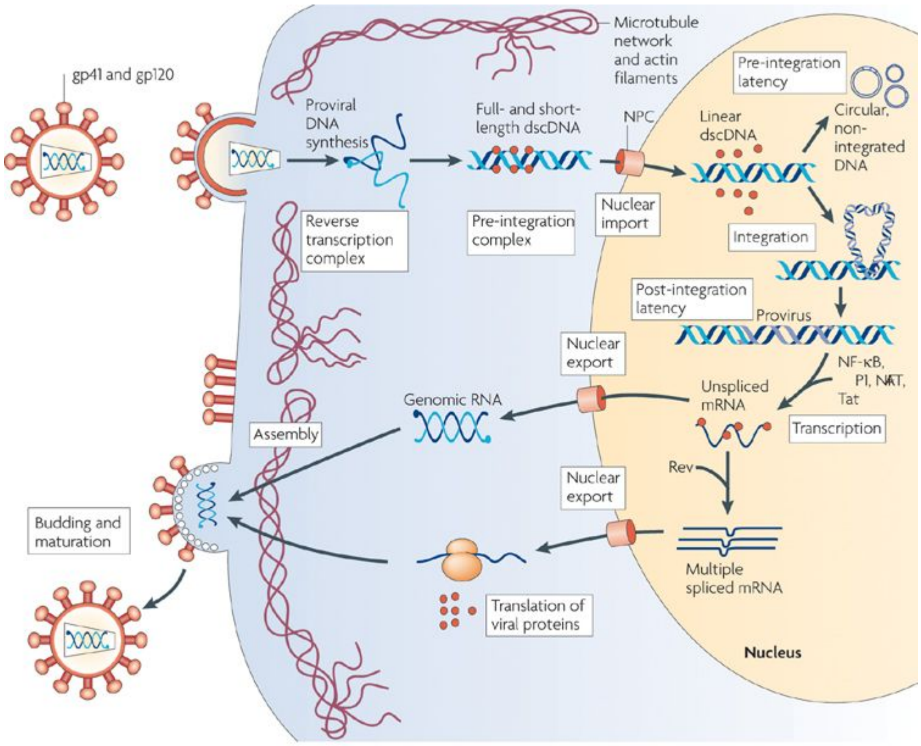
2.4.1 Cell Entry

The first phase of the life cycle of an HIV virion is the entry of the targeted cell, depicted in detail in Figure 2.6. To bring virion and target cell into close proximity, quite unspecific attachment mechanisms exist, which increases the efficiency of an infection [68]. Binding to the surface of the cell is mediated by the viral envelope complex (Env) with gp120 binding to the CD4 receptor of the host cell, and subsequently to a transmembrane chemokine co-receptor [68, 69].

The stable attachment of Env to CD4 and the co-receptor causes conformational changes allowing the fusion peptide of p41 to enter the cell membrane. This mediates the fusion of the viral membrane with the cell membrane, and, consequently, the entry of the viral capsid into the host cell [68].

2.4.2 Reverse Transcription

The mechanisms happening between membrane fusion and establishment of the reverse transcription complex (RTC) in the host cell are poorly understood [68, 70, 71]. When exactly the RTC is constructed, and hence the reverse transcription starts, still remains unclear. It is assumed that the gRNAs together with viral proteins are released into the



Nature Reviews | Microbiology

FIGURE 2.5: Replication cycle of HIV. Graphic taken from [67].

host cell after the uncoating procedure, where the viral capsid is disassembled [72]. Subsequently the RTC is constructed comprising structural proteins MA, NC, the enzymes RT and IN, and the accessory protein Vpr, starting with the reverse transcription process [70, 71]. However, some research teams report an onset of reverse transcription within the capsid directly after the viral entry or even within the unfused virion [72].

The heterodimer reverse transcriptase removes the nucleocapsid proteins from the RNA and produces complementary DNA copies (cDNA). It uses both RNAs as templates and switches frequently between the strands to be transcribed during the copy process. If the initial virion contained two distinct gRNAs new recombinants emerge during this process. Furthermore, because the reverse transcription process is highly error prone (mutation rate per base around 10^{-5}), the initial HIV genome gets modified by multiple mutations. This causes a vast variety of heterogeneous HIV species ("quasispecies"), consequently allowing the virus to evade the human immune response and attain resistance against antiviral drugs [71].

The reverse transcriptase also occupies ribonuclease and DNA polymerase activity: the RT degrades the transcribed RNA and produces a complementary (positive sense) copy of the cDNA such that a double stranded DNA molecule can be formed [71].

2.4.3 Nuclear Import and Integration

Shortly after the final steps of reverse transcription the RTC forms the pre-integration complex (PIC). The double stranded viral DNA is imported into the nucleus, with MA and Vpr carrying nuclear localisation signals (NLS). It is assumed that the PIC interacts with the cellular machinery that executes the host's own nuclear import of karyopherin [73]. Also the enzyme integrase is part of the PIC and plays an important (but not very clear) role for the import [71, 73, 74]. It is known that MA interacts with the core domain of the

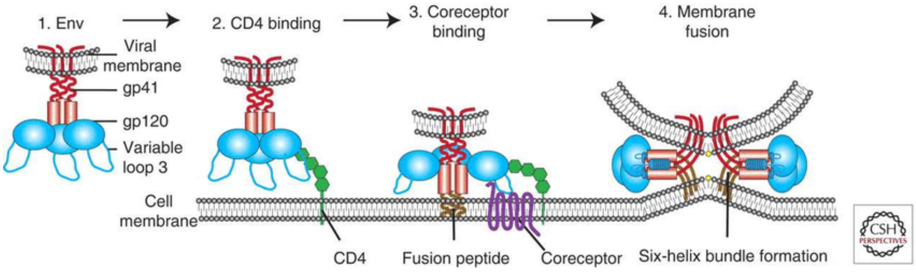


FIGURE 2.6: Fusion of HIV with the target cell. After attaching to the host cell, the viral Env protein binds to the CD4 receptor, and the co-receptor of the cell. Conformation changes allow the fusion glycoprotein gp41 to penetrate the cell membrane, which mediates the fusion of the viral membrane and the cell membrane. Graphic taken from [68].

integrase [74]. Moreover, [75] showed that the viral DNA itself possesses a strand overlap, obtained during reverse transcription, that mediates the import into the nucleus.
After the transfer, the double stranded viral DNA is inserted into the host cell genome by integrase in several steps: clipping nucleotides from the $3'$ end of both strands, staggered cleavage of the target host DNA and integrating the viral strands. Cellular repair enzymes mend arisen gaps at the attachment sites [71].

2.4.4 Replication and Nuclear Export

The integrated viral DNA, referred to as "provirus", is transcribed like host own genes with the necessary support of cellular transcription factors, but encoding the full information of structural, regulatory, and accessory proteins used to direct virus replication [71]. One of the most important transcription factor, nuclear factor kappa B, promotes the expression of genes which largely participate in the host immune response. As a consequence, those cells that are actively fighting an infection are most likely to be infected and subsequently killed by HIV [76].
A large number of viral mRNAs are transcribed (more than 30). They are further processed in three ways [71]:

- The mRNA is spliced multiple times into small fragments (1.7 to 2.0 kb) to be translated into the proteins Rev, Tat (encouraging new virus production), and Nev.
- Partially spliced mRNA (∼5 kb) is translated into the proteins Env, Vif, Vpu, and Vpr.
- The mRNA remains unspliced and serves as mRNA for Gag and GagPol polyprotein precursors, and is packed into new virions as full-length gRNA.

The small, fully spliced viral mRNAs can be transported to the cytoplasm for translation with the export mechanism of the cell for its own mRNAs. After the viral Rev protein is translated in the cytoplasm, it enters the nucleus, binds to the RRE domain present in unspliced and partially spliced RNA, and mediates the nuclear export of these [71].
Gag is expressed as a 55 kDa polyprotein precursor ($Pr55^{Gag}$), and is the key player for the viral assembly step. A frame shift during translation of Gag enables the production of the 160 kDa GagPol polyprotein precursor, which contains the enzymes PR, RT, and IN. GagPol yields a rate of ∼5% of the Gag production [71, 77].

2.4.5 Assembly and Maturation

After protein synthesis and nuclear export, the assembly process, depicted in Figure 2.7, is initiated at the plasma membrane of the cell.

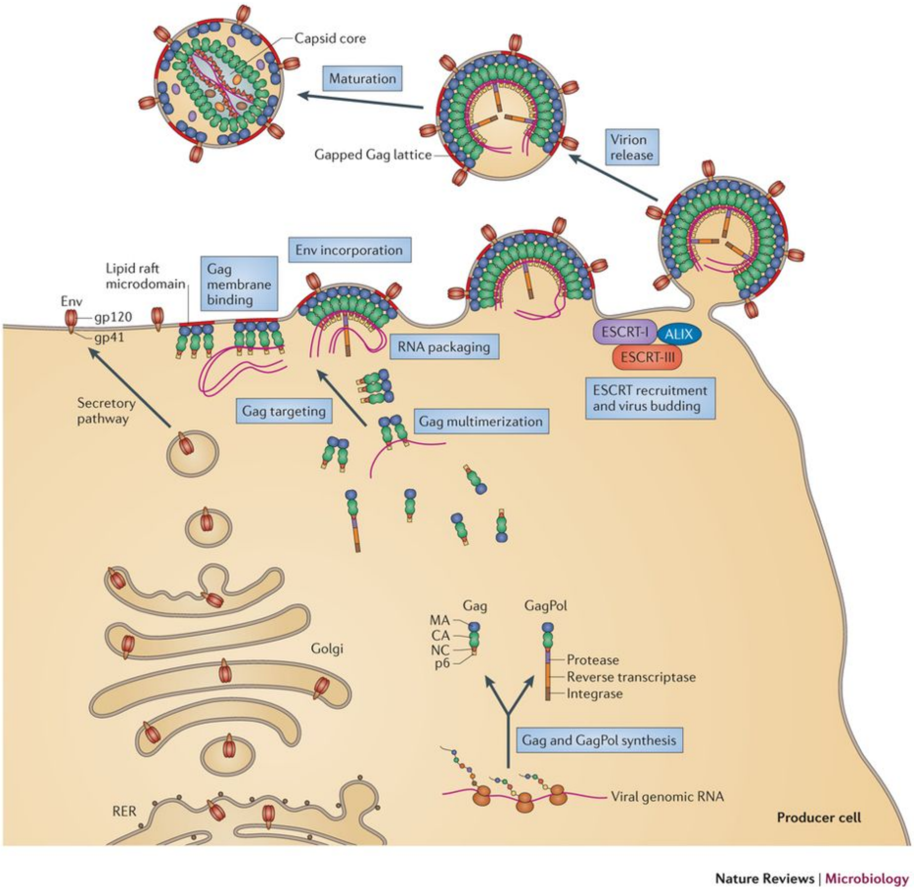


FIGURE 2.7: Assembly of virions. Graphic taken from [77].

Pr55^{Gag} is the main driver of the viral assembly. After translation in the cytosol it interacts with the gRNA-dimers. The untranslated gRNA molecules are dimerised after transfer to the cytosol through intermolecular base-pairing [78], which is crucial for packaging into assembled virions. The main determinant in Gag for packaging of the gRNA-dimer is the NC domain, which binds to the packaging signal in the 5′ UTR of the gRNA [77, 79, 80]. However, studies showed that the Gag-RNA interaction changes through assembly and maturation processes [77, 81]. They revealed that after arrival to the cell membrane and during virus assembly the molecules are linked at several sites throughout the genome. The NC-RNA-interaction plays also a role in Gag multimerisation, however, the main driver for the Gag assembly is the CA domain in Gag.

The transport of Gag to the plasma membrane is mediated by the MA element. Here, Gag targets lipid raft microdomains of the membrane to start the assembly of the virus [82–84]. For building the immature virus particle, Gag molecules are connected and packed with the MA domain building the inner matrix of the viral membrane and the C terminus towards the virus centre [71, 77]. Also the GagPol polyprotein precurser is packed into the nascent virus via Gag-Gag interactions [71].

The viral Env glycoprotein is synthesised in the rough endoplasmatic reticulum (ER) of the cell from the translated Env precursor protein gp160 and transferred through the Golgi. From here it is transported within vesicles to the cell membrane [71]. The specific

mechanism how Env is incorporated into the nascent virus is not completely resolved [77]. One of the theories promotes a central role for the MA domain of Gag interacting directly with the gp41 part of Env. This assumption is supported by several research results [85–88]. Thus, the viral assembly engages Gag as the central determinant building Gag-RNA, Gag-Gag, Gag-lipid and Gag-Env interactions.
After the immature Gag lattice is assembled at the membrane, the release of the virus particle is achieved through Gag taking over the cellular ESCRT machinery, which is responsible for membrane budding and fission processes [77]. The so called late- or "L"-domains, located in the p6 N-terminus of Gag, trigger the membrane scission [71, 77].
The maturation process starts shortly after budding off. It is a highly ordered multistep process involving a cascade of conformational switches and subunit rearrangements [44, 77]. The viral enzyme protease in the immature virion cleaves the polyprotein precursors Gag and GagPol into the actual mature proteins, including the structural elements of the virus (MA, CA, NC), and the viral enzymes (PR, RT, IN) respectively. The cleavage of the polyprotein precursors is accompanied by morphologic changes of the virion. The immature virus contains the Gag molecules in a radial composition, whereas the cleaved CA proteins form a conical capsid around a complex built by the RNA, NC, and RT. The matrix proteins remain at the inner leaflet of the viral membrane [44, 77]. The mature virus is now capable of infecting new host cells.

2.5 Treatment

The life cycle of HIV involves many crucial steps, each of them being a potential target for antiviral drug development. One of the first targets was the reverse transcription process. In 1987, the first antiviral therapy (ART), Zidovudine (AZT), was approved for clinical treatment of HIV-1 patients [89], only four years after the discovery of HIV being the cause for AIDS.
AZT is a nucleoside analogue which inhibits reverse transcription when incorporated into the nascent DNA strand. Shortly after, several other RT targeting drugs were developed. However, due to the high mutation rate of RT, resistance mutations could emerge rapidly after beginning of the treatment. Especially the fact that the therapy initially started with only one medicine reinforced the development of resistant virus strains [71].
Current HIV regimen include three different drugs of at least two drug classes [5]. The four major classes acting on essential steps of the HIV life cycle are briefly listed below with respective drug names.

- **RT inhibitors**: **NRTIs** (nucleoside reverse transcriptase inhibitors) are nucleoside analogues averting the reverse transcription process by competing with nucleosides and stopping reverse transcription
  →zidovudine, abacavir, lamivudine, emtricitabine, tenofovir
  NNRTIs (non-nucleoside reverse transcriptase inhibitors) bind to an allosteric site of the enzyme to inhibit reverse transcription
  →efavirenz, nevirapine, etravirine, rilpivirine

- **PIs** (protease inhibitors): blocking protease to prevent the cleavage of Gag and Gag-Pol precursor proteins resulting in uninfective immature virus particles
  →lopinavir, indinavir, nelfinavir, amprenavir, ritonavir, darunavir, atazanavir

- **Entry inhibitors**: including fusion inhibitors which block HIV Env from merging with the host CD4 cell membrane to prevent the fusion of viral and cell membrane, and inhibitors blocking receptors and co-receptors on the cell surface
  →maraviroc, enfuvirtide

- **INSTIs** (integrase nuclear strand transfer inhibitors): interfering with the enzyme integrase to prevent integration of the viral DNA into the host DNA
  →raltegravir, elvitegravir, dolutegravir

The combination therapy is also known as HAART (highly active antiretroviral treatment). Potent treatment results are achieved by including a "backbone" of two NRTIs in combination with a INSTI, an NNRTI, or a PI [90]. The treatment with HAART allows to manage the chronic disease by keeping the number of virions in the body (viral load) low, maintaining a functional immune system and preventing the development of serious, often lethal opportunistic infections [91, 92]. Additionally, the low viral load decreases the probability of transmission to an uninfected person [92]. Although drug side effects occur, HIV patients can live an almost normal life with a close to ordinary life expectancy [92]. Yet, due to its high mutation– and replication rate, HIV manages to escape drug treatment and develop multiresistance resulting in complete treatment failure.
Furthermore, common therapy regimen do not work optimally for HIV-2. HIV-2 is intrinsically resistant to NNRTIs. It still remains a critical challenge to clearly define the best strategy to treat HIV-2 [93]. Thus, thorough investigation of viral dynamics and evolution is required to improve current treatment strategies.

Chapter 3

Background: Evolutionary Constraints

In this chapter, we give a few definitions of evolutionary genetics:
First, we introduce the term fitness and look at fitness landscapes. These landscapes comprise evolutionary paths, which are useful to infer evolutionary constraints. In the subsequent section we will explore biological constraints in more detail. These constraints include among others genetic interactions, so called epistasis, which we consider in the last section.

3.1 Fitness

In population genetics, the **fitness** of a certain genotype (the genetic variant) or phenotype (observable characteristics) of an organism denotes a quantitative measure of the ability to reproduce in a given environment [94, 95]. The fitness may be broken down into various components (traits), but essentially it denotes surviving, and propagating its genetic material in the gene pool (the set of all genetic information in a population). The fitness of an individual is quantified by its average contribution in the gene pool of the successive generation. Hence, it depends on the background population. Additionally, the fitness of one phenotype can be different in distinct selective environments, i.e. the fitness is also influenced by external conditions.
In the perspective of evolutionary biology, fitness is assigned to the genotype instead of an individual organism and is measured by its reproductive success or "replication rate" [96]. This can be either absolute fitness, or more common, the relative fitness in relation to a reference sequence (e.g. wild type, or the genotype with the highest fitness). Here, fitness is considered as the phenotype of the genotype.
In every new generation of the organisms, the genotypes do not retain the initial sequence, but undergo genetic changes, including mutation, and recombination of distinct genotypes. Hence, the fitness of an organism can be broken down into fitness measures of genetic fragments (e.g. genes) with respect to the average genetic background.
After several generations of replication, the fittest genotype of the regarded trait is very likely most prevalent in the gene pool, resembling the natural selection process, i.e. the Darwinian evolution theory [96].

Fitness landscapes are used to visualise the fitness of different genotypes [96, 97]. An example for a protein sequence with two residues is shown in Figure 3.1. Here, the plain of the surface, x- and y-axis, represent different genetic variants of two distinct genetic loci, and the vertical z-axis shows the respective fitness of the genotype with the combination of the two variants. Genotypes that are similar are close on the axis. The more differences in the genome, the more they are apart in the landscape. This enables the identification of local fitness peaks, i.e. genetic variants with comparatively high fitness, and fitness

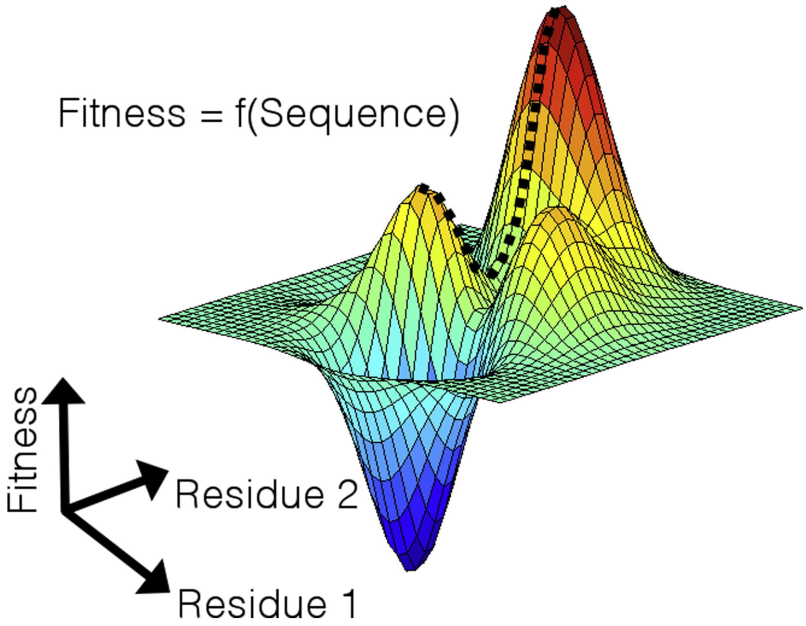


FIGURE 3.1: Fitness landscape showing the replicative fitness of the sequence space comprising two loci (here protein residues) for all combinations of variants (20 amino acids). Figure taken from [98].

valleys, comprising genotypes with low fitness. The landscape can be evaluated by its density of peaks (and valleys), giving the ruggedness, or contrarily, the smoothness. Together with other properties of the fitness landscape, the ruggedness contains reference to the speciation process in the evolution of the gene pool.
The fitness landscape represents a network of all possible trajectories between all genotypes, moving along the path of genotypes that differ only at a single position [99, 100], depicted in Figure 3.2. As mentioned above, the evolutionary path of a population is

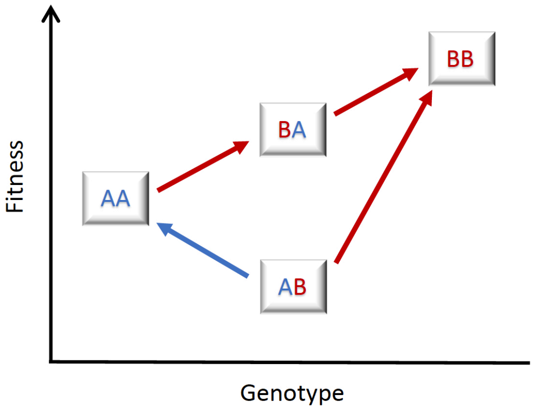


FIGURE 3.2: Fitness landscape for a sequence of length 2 with two possible symbols per positions, A and B. The trajectory of the evolutionary path moves from low fitness to high fitness. The sequences are connected if they differ at only one position.

usually restricted to paths of increasing fitness, especially under conditions with strong selection and low mutation [101]. In the example in Figure 3.2, we have a sequence of length 2. Each position can comprise one of two symbols, A and B. The red arrows depict the paths towards the variant "BB", which is a local/global maximum. Only the sequence

variant "AB" is less fit than "AA". The pathway starting from "AB" might propagate towards the maximum "BB", or to "AA" (blue arrow), and then subsequently towards "BB". Since, fitness is depending on the background and the environment, any change of the conditions can induce a shift of the fitness landscape and hence the evolutionary paths.
Fitness valleys constitute constraints for the evolutionary development. If the trajectory from a genotype with medium fitness to a peak contains a valley, it is less likely that the gene pool evolves towards the peak. In our example, that would be the case, if both "AB" and "BA" have lower fitness than "AA" and "BB".
In the next section, we are going to elaborate on evolutionary constraints in more detail.

3.2 Constraints

Constraints in a biological point of view are factors which limit evolutionary change. These factors may be intrinsic (genetic variants) or extrinsic (selective pressure among the species) causes to prevent a phenotype to adapt and reproduce. They restrict the number of evolutionary paths, or phenotypes that may emerge in the future [102].
As discussed in several publications [103, 104], the term **evolutionary constraint** is very manifold and inconsistently defined.
In essence, evolutionary constraints can be interpreted as specific factors, which prevent the transition of an established trait to a more advantageous state or selective optimum [104, 105].
Various types of constraints were discussed in literature, which describe either the regarded trait, the level of variation or the cause of the constraint [103]. Several reviews discuss and try to disentangle the wide range of terminologies and definitions [104, 105]. Evolutionary constraints can be divided into different groups of abstractions, but they are highly connected.
In the following we briefly summarise different categories.

On the lowest abstraction level are **genetic constraints**. They are determined by the amount or patterns of genetic variation which may limit the adaption of a trait [104]. This includes epistatic interactions (cf. Section 3.3) and pleiotropy, i.e. the influence of one single gene on multiple phenotypic traits.

Developmental constraints are caused by a variety of factors including for example structural restrictions, biological scaling in relation to size (allometry), and similarities in early stages of embryonal development of vertebrates [105]. Constraints due to developmental force are closely connected to genetic constraints, as they largely determine the required genetic interactions [104]. On the other side, they influence phenotypic development by conditioning, which heritable phenotype variation is exposed to environmental factors and hence, natural selection [106].

Selective and functional constraints, also referred to as adaptive trade-offs [105], are given, when the selection of one trait or function effects, or even prevents, another trait to evolve towards adaption [104]. Thus, the selective constraint is in regard to a specific trait, which is limited by selection of correlated functions. This can be either by preference of another trait, or selective pressure of different factors acting on this particular trait. The selection can either be external or internal [107]: External selection corresponds to the classical Darwinian definitions of an organism interacting with and adapting to its environment. Internal selection is referring to the interaction of internal components enabling the response to external factors. Again, these constraints are in close coherence with genetic interactions constraining evolution.

Phylogenetic and historical constraints denote the influence of the preceding developments in the past, and how ancestry stirs evolution in particular directions by constraining the possible variation. The illustrative example in [104] makes the concept more clear. Regarding different animals swimming in the sea: they are all well adapted to move in water. Yet, they developed different methods to swim depending on the ancestral origin. The penguin uses the inherited wings for swimming and a whale bred a horizontal fluke out of hind-legs. These constraints are highly affected by developmental constraints.

In the following section we explore details of genetic constraints, i.e. epistasis.

3.3 Epistasis

The term **epistasis**, originally describing the interaction of gene loci, covers a variety of different definitions and interpretations among different scientific fields. In essence, epistasis gives the deviation from an expected outcome of independent effects.
The expression "epistatic" was initially used by Bateson [108] in 1909 to explain models of natural selection beyond dominant and recessive alleles on individual loci. Back then, it described the ability of a pair of gene alleles at one locus ("allelomorphic pair") to mask the phenotypic effect of another locus, meaning one of the loci is dominating the other locus. Over time, with the progression of genetic research, this definition has been extended to more complex gene interaction models [109, 110]:
Epistasis is given

- whenever an effect is masked by the absence or presence of an attribute,
- when an effect of one allele is modified by one allele (or more) of another locus,
- or, if at least two loci interact such that a new phenotype emerges.

Also quantitative genetics started to adopt the idea of epistasis, though these definitions depart even more from the original meaning. Fisher et al. used "epistacy" to describe gene interactions generating statistical deviation from additive effects on quantitative phenotypes [111].
In general, the quantitative interaction analysis defines two components. A quantitative phenotypic measure and a neutrality function, representing the phenotype without interaction. The epistasis is given by the deviation of the expected neutral phenotype [112]. In molecular– and evolutionary biology the techniques to quantify effects of genes and gene interactions on measurable attributes advanced. These measurable attributes include for example pigmentation, catalysis, or growth rate.
The latter is one type of **fitness epistasis**, which describes interaction of distinct genetic sites affecting the fitness of the genotype.
Here, we are considering the measure as the decrease/increase of the mutant population relative to wild type after one generation. In the following, we will introduce different neutrality functions and types of interactions that may occur.

3.3.1 Neutrality Functions

To describe the variety of neutrality functions, we consider a haploid organism (one set of genes), with two genotype variants for each locus, wild type and mutant, indicated with a small and capital letter respectively. The quantitative phenotype is a fitness function f, i.e. $f(ab)$ gives the fitness of the wild type and $f(AB)$ denotes the neutral fitness of the double mutant for loci a and b, without interaction.

3.3.2 Multiplicative Model

The definition of the multiplicative neutrality function is given by

$$f(AB) \cdot f(ab) = f(Ab) \cdot f(aB). \tag{3.1}$$

If the fitness measure is given in relation to the wild type, we have $f(ab) = 1$ and thus

$$f(AB) = f(Ab) \cdot f(aB). \tag{3.2}$$

The epistasis is given by

$$E_{ab} = \frac{f(AB)f(ab)}{f(Ab)f(aB)}. \tag{3.3}$$

3.3.3 Additive Model

If two co-occurring mutations have the same effect as the sum of the individual effects, they are purely additive without any interaction. The interaction can be formulated as

$$f(AB) + f(ab) = f(Ab) + f(aB) \tag{3.4}$$

The epistatic effect of a double mutant is attained with

$$E_{ab} = f(AB) + f(ab) - f(Ab) - f(aB). \tag{3.5}$$

The multiplicative model can also be expressed as additive model by transforming the fitness into the log scale:

$$\begin{aligned} \log E_{ab} &= \log\left(\frac{f(AB)f(ab)}{f(Ab)f(aB)}\right) \\ &= \log f(AB) + \log f(ab) - \log f(Ab) - \log f(aB). \end{aligned} \tag{3.6}$$

3.3.4 Measures of Epistasis

The various types of epistatic interactions are illustrated in Figure 3.3 and are explained in the following.

Magnitude Epistasis

If the magnitude of the effect, yielded by the double mutant, deviates from the purely additive effect of the two individual mutations, an epistatic interaction is given. **Positive epistasis** is given, if the fitness is higher than expected, i.e. two beneficial mutations have amplification in fitness when occurring in combination, and two deleterous mutations have less fitness loss than the sum of the individual mutations. In conversion, if the phenotype is less fit than expected, **negative epistasis** is given.

Synergistic epistasis describes the enhancement of the either positive or negative additive fitness magnitude. On the contrary, **antagonistic epistasis** is a reduction of the magnitude for a pair of beneficial, or deleterious mutations, i.e. less severe effects than expected. The interpretation of the actual effect on fitness depends on the context of the two single mutations [112, 114]. An example for antagonistic interactions is a mutation at one particular site, which is highly deleterious and masks the effect of additional mutations.

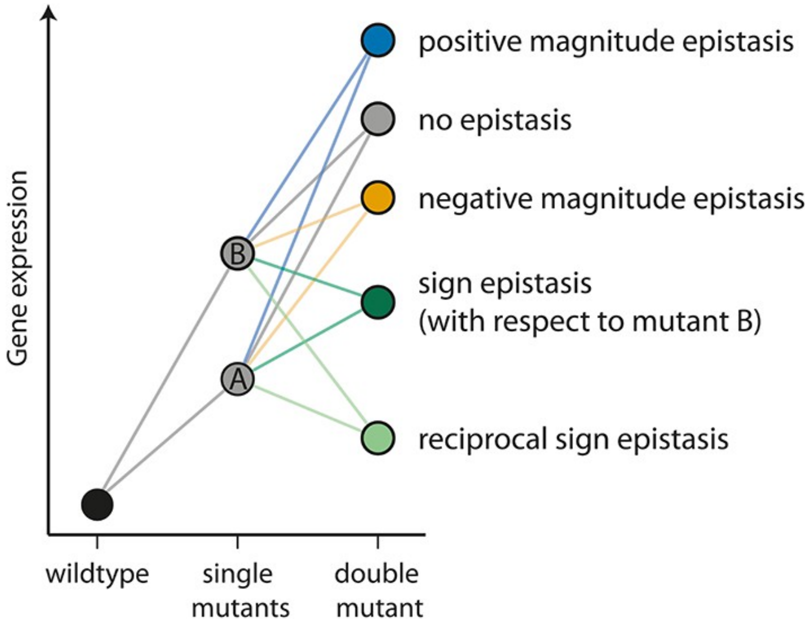

positive magnitude epistasis
no epistasis
negative magnitude epistasis
sign epistasis
(with respect to mutant B)
reciprocal sign epistasis
Gene expression
B
A
wildtype
single mutants
double mutant

FIGURE 3.3: Different types of epistasis between two point mutations. Figure taken from [113].

Sign Epistasis

The concept of sign epistasis states that the effect of a mutation depends on the genetic background of the population, which means the mutation can be beneficial for some of the genotypes and deleterous for others [115]. Thus, the sign of the effect of a mutation is under epistatic control. An example would be a mutation that is deleterious if it appears individually, but enhances the effect of a beneficial mutation.
A more extreme case is the **reciprocal sign epistasis**, where the sign of the effect changes for both mutations when ocurring together, i.e. both mutations have a negative effect on its own, but are beneficial (positive) in co-occurrence [99].

Chapter 4

Methods to infer Evolutionary Constraints

This chapter gives an overview of current techniques for the inference of evolutionary constraints using closely related biological sequences. These sequences are aligned in a multiple sequence alignment (MSA) of a biologically relevant length, with each column representing the distribution of symbols at one position of the sequence. Hence, an MSA contains information about the sequence conservation and allows to assess correlations between two residues of the sequence, which give some indication of evolutionary constraints. Co-evolution in sequences is an essential component to understand evolutionary relationships and comprehend functional developments.

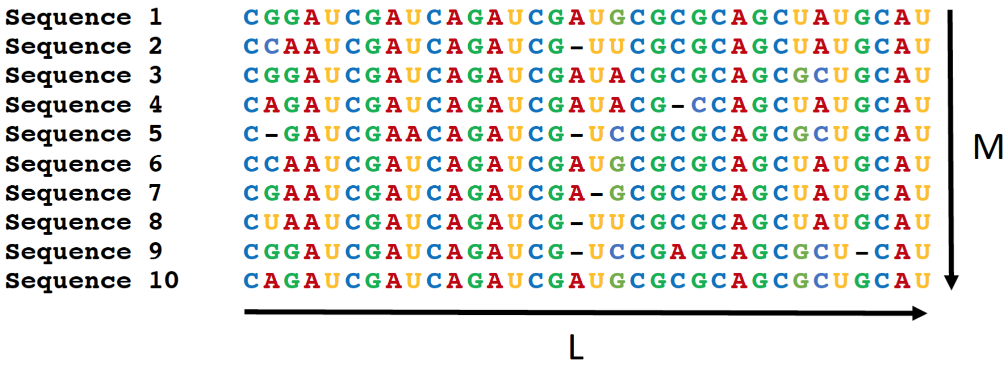
A great variety of methodologies have been established in the last decades considering MSAs with the goal to predict functional and structural constraints within polymeric molecules [116].
One of the first approaches to detect covarying sequence pairs in a biological context was introduced by Korber et al. in 1993 [117], applying techniques based on information theory. Finding interdependent amino acid frequencies by mutual information of two sequence residues facilitated the identification of co-evolving mutations. These may give implications for structural or functional relationships between those residues. Improvements of the basic method accounted for and corrected for false positives caused by sampling biases and phylogenetic effects [118–120]. Another approach, sometimes known as so called McLachlan-based substitution correlation (McBASC), uses correlation measures to extract substitution patterns of two positions of an MSA[121–123]. This information is used to predict residues in close proximity within three-dimensional protein structures. Other methods use phylogenetic approaches detecting patterns of simultaneous amino acid substitutions [124, 125].
As we will explain in the following sections, an important obstacle is the distinction of direct couplings from indirect relations, which occur when more complex substitution patterns involving more than two positions exist. Initially proposed by Lapedes et al. [126], the idea was seized in the recently developed direct couping analysis (DCA) [9, 127, 128], as well as the method of protein sparse inverse covariance estimation (PSICOV) [129]. Both methods construct a global statistical model of the MSA to infer direct couplings. An alternative approach predicts pairwise protein-protein interactions with an parameter-free Bayesian network [130].

In this chapter, we present dependency measures based on information theory. First, we introduce the input data, an MSA of biological sequences. After briefly discussing mutual information, the main focus is put on direct coupling analysis. For the latter, the derivation of the maximum entropy model is described and the resulting Potts model is explained broadly, including its origin in statistical physics. Different approaches for the inference of the model parameters are presented, and various scoring schemes for the evaluation are discussed.

4.1 Representation of Biological Sequences and MSA

Each biological sequence $\sigma = \{\sigma_1, \dots, \sigma_L\}$ is composed of L symbols from the alphabet $\mathcal{A}$ containing a set of q different symbols. A protein sequence would consist of 20 different amino acids (or 21 symbols including a gap) and a DNA or RNA sequence consists of 4 nucleotides (or 5 characters including a gap).
The MSA $\mathcal{M}$ (an example shown in Figure 4.1) contains the aligned sequences, resulting in a matrix of residues with M sequences as rows and L columns for each sequence position:

$$\mathcal{M} = \{\mathcal{M}_i^a\}, \text{ with } i = 1, \dots, L \text{ and } a = 1, \dots, M. \tag{4.1}$$



```
Sequence 1     CGGAUCGAUCAGAUCGAUGCGCGCAGCUAUGCAU
Sequence 2     CCAAUCGAUCAGAUCG-UUCGCGCAGCUAUGCAU
Sequence 3     CGGAUCGAUCAGAUCGAUACGCGCAGCGCUGCAU
Sequence 4     CAGAUCGAUCAGAUCGAUACG-CCAGCUAUGCAU
Sequence 5     C-GAUCGAACAGAUCG-UCCGCGCAGCGCUGCAU
Sequence 6     CCAAUCGAUCAGAUCGAUGCGCGCAGCUAUGCAU
Sequence 7     CGAAUCGAUCAGAUCGA-GCGCGCAGCUAUGCAU
Sequence 8     CUAAUCGAUCAGAUCG-UUCGCGCAGCUAUGCAU
Sequence 9     CGGAUCGAUCAGAUCG-UCCGAGCAGCGCU-CAU
Sequence 10    CAGAUCGAUCAGAUCGAUGCGCGCAGCGCUGCAU
```

FIGURE 4.1: MSA of RNA sequences with $M = 10$ rows and $L = 34$ columns.

For simplicity of the notation the q symbols are translated into consecutive numbers $1, \dots, q$.

The empirical residue frequencies for single positions and pairwise positions contained in the MSA $\mathcal{M}$ are given by

$$f_i(A) = \frac{1}{M} \sum_{a=1}^{M} \delta_{A,\mathcal{M}_i^a} \tag{4.2}$$

$$f_{ij}(A,B) = \frac{1}{M} \sum_{a=1}^{M} \delta_{(A,\mathcal{M}_i^a)} \cdot \delta_{(B,\mathcal{M}_j^a)} \tag{4.3}$$

with $1 \leq i, j \leq L$ and $1 \leq A, B \leq q$, and the Kronecker-symbol δ being 1 if the symbols are equal and 0 otherwise.
In the MSA in Figure 4.1, for example, the marginal frequency in the second column for nucleotide G is $f_2(\text{G}) = 0.4$ and in the third column for nucleotide A is $f_3(\text{A}) = 0.4$. The pairwise frequency is $f_{2,3}(\text{G}, \text{A}) = 0.1$.

4.2 Mutual Information

In information theory, mutual information (MI) is a local correlation measure of how dependent two random variables are. It quantifies the amount of information one random variable reveals about another variable, hence how much information they share. MI is closely linked to the variability or uncertainty of random variables, the so called entropy, defined by Claude Shannon [131]:

$$\begin{aligned} MI(X,Y) &\equiv H(X) - H(X|Y) \\ &\equiv H(Y) - H(Y|X) \\ &\equiv H(X,Y) - H(X|Y) - H(Y|X) \\ &\equiv H(X) + H(Y) - H(X,Y) \end{aligned} \tag{4.4}$$

with $H(X)$ and $H(Y)$ the marginal entropies of the random variables X and Y respectively, $H(X|Y)$ and $H(Y|X)$ giving the conditional uncertainties of one variable with knowledge of the other one and $H(X,Y)$ is the joint entropy of X and Y. The relation is depicted in the Venn diagram in Figure 4.2. The mutual information measures how much the knowledge about one variable reduces the uncertainty about the other one. If two variables are absolutely independent, the knowledge about one of the variables gives no information about the other one, thus the MI is zero. With maximal MI, the uncertainty about the random variable vanishes, the two variables are deterministically connected and the information conveyed for one of them includes all information about the other variable.

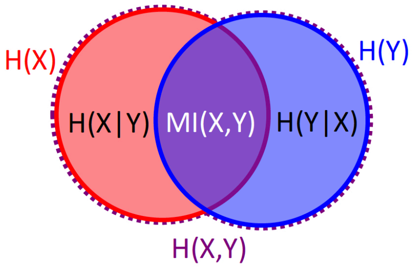


FIGURE 4.2: Venn diagram depicting the set-theoretical relationship of mutual information ($MI(X,Y)$, purple intersection area) and entropy H of correlated random variables X and Y. The red and blue circles show the marginal entropies $H(X)$ and $H(Y)$ respectively, substracting the intersecting purple MI result in the conditional entropies $H(X|Y)$ and $H(Y|X)$. The joint entropy $H(X,Y)$ is indicated by the purple dashed line.

Given the discrete random variables X and Y, the mutual information is denoted by

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y) \ln\left(\frac{P(x,y)}{P(x)\cdot P(y)}\right), \tag{4.5}$$

with the joint probability distribution function $P(x,y)$ of X and Y and the respective marginals $P(x)$ and $P(y)$ for X and Y. The MI is zero if and only if the two variables are independent: $MI(X,Y) = 0 \longleftrightarrow P(x,y) = P(x)\cdot P(y)$. Moreover, MI is nonnegative, i.e. $MI(X,Y) \geq 0$, and symmetric, i.e. $MI(X,Y) = MI(Y,X)$.

For our case, the MI for two positions i and j within the MSA can be written as

$$MI_{ij} = \sum_{A=1}^{q} \sum_{B=1}^{q} f_{ij}(A,B) \ln \left(\frac{f_{ij}(A,B)}{f_i(A) \cdot f_j(B)} \right), \tag{4.6}$$

where $f_i(A)$ and $f_j(B)$ denote the frequencies of symbols A at position i and B at position j, $f_{ij}(A,B)$ gives the observed co-occurrence of symbols A and B at positions i and j within the MSA.

4.3 Direct Coupling Analysis

Direct coupling analysis (DCA) is the generic name for a collection of several methods analysing sequence data, like proteins or RNA. The main idea is to use statistical modelling to extract the **direct** relation of two positions within an MSA of homologous sequences without confounding effects from other positions.
Common methods to derive the correlation in an MSA, for example mutual information (cf. Section 4.2), use only **local** measures, i.e. only the frequencies of two particular residues are taken into account. Pairs are considered in isolation and **indirect** associations might be inferred, simplified shown in Figure 4.3. Transitive pair correlations can confound the true result.

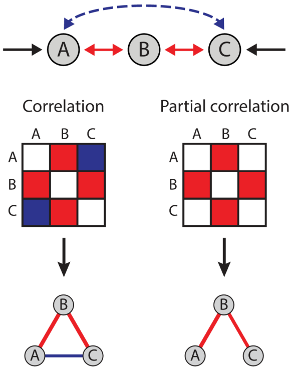


FIGURE 4.3: Correlation analysis. Pearson's correlation yields interactions between all three compounds A, B, and C. The partial correlation reveals the true correlations between A and C. Figure taken from [132], based on [133].

In DCA, co-evolving nucleotides are identified by using a **global** statistical model: P describes the probability of occurrence of the nucleotide sequence as a joint probability distribution in the sequence space. The aim is to derive the simplest possible probability model explaining the observed data, which will be explained in detail in the next section.

4.3.1 Derivation of the Maximum Entropy Model

DCA aims at modelling the sequence variability in the input MSA via a generalised Potts model, which is derived by using the **principle of entropy maximization** [134]: information and information entropy are linked. The more information in the system, the lower its entropy and vice versa.
The principle of maximum entropy states that, subject to precisely stated prior data the probability distribution which best represents the current state of knowledge is the one

with the largest entropy.
Thus, we search for a probability distribution P for sequences σ:

1. fulfilling the normalisation constraint

$$\sum_{\sigma} P(\sigma) = 1 \tag{4.7}$$

2. being compatible with the measured single and pair frequencies

$$P_i(A) = f_i(A) \tag{4.8}$$
$$P_{ij}(A,B) = f_{ij}(A,B) \tag{4.9}$$

3. maximising the information entropy

$$\text{maximise } H = -\sum_{\sigma} P(\sigma) \ln P(\sigma) \tag{4.10}$$

Lagrange multipliers are used to find the stationary point of a function subject to equality constraints and thus the Lagrangian for the function is given by

$$\begin{aligned}\mathcal{L}(P(\sigma),\alpha,\beta(a),\gamma(a,b)) = & -\sum_{\sigma} P(\sigma)\ln P(\sigma) \\ & + \alpha \left(\sum_{\sigma} P(\sigma) - 1 \right) \\ & + \sum_{i=1}^{L} \beta_i(\sigma_i)\left(P_i(\sigma_i) - f_i(\sigma_i)\right) \\ & + \sum_{i=1}^{L}\sum_{j=1}^{L} \gamma_{ij}(\sigma_i,\sigma_j)\left(P_{ij}(\sigma_i,\sigma_j) - f_{ij}(\sigma_i,\sigma_j)\right).\end{aligned} \tag{4.11}$$

The stationary point is found by setting the functional derivative of $\mathcal{L}$ with respect to the unknown density $P(\sigma)$ to zero: $\frac{\partial \mathcal{L}}{\partial P(\sigma)} = 0$.

$$\frac{\partial \mathcal{L}}{\partial P(\sigma)} = 0 = -\ln P(\sigma) - 1 + \alpha + \sum_{i=1}^{L} \beta_i(\sigma_i) + \sum_{i=1}^{L}\sum_{j=1}^{L} \gamma_{ij}(\sigma_i,\sigma_j) \tag{4.12}$$

$$\ln\left(P(\sigma)\right) = \alpha - 1 + \sum_{i=1}^{L} \beta_i(\sigma_i) + \sum_{i=1}^{L}\sum_{j=1}^{L} \gamma_{ij}(\sigma_i,\sigma_j) \tag{4.13}$$

$$\begin{aligned} P(\sigma) &= \exp\left(\alpha - 1 + \sum_{i=1}^{L} \beta_i(\sigma_i) + \sum_{i=1}^{L}\sum_{j=1}^{L} \gamma_{ij}(\sigma_i,\sigma_j)\right) \\ &= \frac{1}{Z}\exp\left(\sum_{i=1}^{L} \beta_i(\sigma_i) + \sum_{i=1}^{L}\sum_{j=1}^{L} \gamma_{ij}(\sigma_i,\sigma_j)\right) \end{aligned} \tag{4.14}$$

The Lagrangian multiplier α is not a free parameter because it can be fully determined for given β and γ by the normalisation constraint $Z \equiv \exp(1-\alpha)$. Because of the symmetry of the Lagrangian multipliers $\gamma_{ij}(a,b) = \gamma_{ji}(b,a)$ and $\delta_{\sigma_i,a} \cdot \delta_{\sigma_i,b} = 1 \iff a = b$, the free parameters can be reduced to

$$h_i(a) := \beta_i(a) + \gamma_{ii}(a,a) \text{ and} \tag{4.15}$$

$$g_{ij}(a,b) := 2\gamma_{ij}(a,b) \text{ for } i < j \tag{4.16}$$

yielding the **maximum entropy probability distribution**

$$P(\sigma) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L} h_i(\sigma_i) + \sum_{1 \le i < j \le L} g_{ij}(\sigma_i, \sigma_j) \right), \tag{4.17}$$

for a system of $q = 2$ states known as the Ising model and for the multivariate case with $q > 2$ known as the **Potts model**.

The Potts Model

In statistical physics, the Potts model is a generalisation of the Ising model: it describes a system of L interacting variables or spins with q different states on a lattice. The Lagrangian multipliers can be interpreted as local **magnetic fields** h acting on individual spins within the system and the **coupling strength** g encoding pairwise interactions. The model, also known as Boltzmann distribution, is typically of the form

$$P(\sigma) = \frac{1}{Z} e^{-\frac{\mathcal{H}}{kT}} \tag{4.18}$$

where $\mathcal{H}$ is the **Hamiltonian**

$$\mathcal{H} = -\sum_{i=1}^{L} h_i(\sigma_i) - \sum_{1 \le i < j \le L} g_{ij}(\sigma_i, \sigma_j) \tag{4.19}$$

specifiying the energy of a spin configuration (or sequence) $\sigma \equiv \{\sigma_1, \ldots, \sigma_L\}$. The temperature is incorporated into the couplings and fields such that $kT = 1$. Z is the normalising **partition function**, the sum of the potential functions for each sequence σ

$$Z = \sum_{\sigma} \exp\left(\sum_{i=1}^{L} h_i(\sigma_i) + \sum_{1 \le i < j \le L} g_{ij}(\sigma_i, \sigma_j) \right). \tag{4.20}$$

Collectively, magnetic fields and couplings are the parameters of the **Potts problem**. If the couplings and fields are given, the **forward problem** is to compute the single site **magnetisations** and **pair correlations** under the Boltzmann distribution.
The **inverse problem** is seen from the reverse direction: with existing statistical observables the unknown fields and couplings are derived.

Gauge Fixing

The parameters h and g of the model have to be fitted.
In order to get a **unique** solution for the model, the number of independent constraints has to match the number of free parameters to estimate. The problem here is, that dependencies of constraints due to

$$1 = \sum_{A \in \mathcal{A}} P_i(A) \qquad \text{and} \qquad P_i(A) = \sum_{B \in \mathcal{A}} P_{ij}(A,B) \quad \forall\, i,j = 1,\ldots,L$$

lead to a number of $\binom{L}{2}q^2 + Lq$ free parameter to estimate versus $\binom{L}{2}(q-1)^2 + L(q-1)$ independent constraints.
To ensure uniqueness of the estimated parameters and the probability distribution, the number of free parameters has to be reduced (so called **gauge fixing**).
There are different solutions of creating dependencies of parameters, for example

- setting parameters corresponding to the last symbol in the alphabet q to zero: $g_{ij}(q,*) = g_{ij}(*,q) = h_i(q) = 0$
- zero-sum gauge: $\sum_{a=1}^{q} g_{ij}(A,B) = \sum_{a=1}^{q} g_{ij}(B,A) = \sum_{a=1}^{q} h_i(A) = 0$

The different variants for gauge fixing are not equally efficient, depending on the scoring scheme. For example, the zero-sum gauge is suitable for non-gauge invariant scoring functions like the (APC) Frobenius norm, because it minimises the sum of squares of the pairwise parameters.
Some scoring schemes do not require any gauge fixing, like the $L1$ and $L2$ regulariser in maximum likelihood-based estimation. Here, the regulariser selects for a unique representation among all parametrisations of the optimal distribution.

4.3.2 Parameter Inference

Parameters h and g have to be determined obeying the constraints in Equations (4.8) and (4.9). The main problem is the determination of Z, which requires the computation of the Hamiltonians of the whole sequence space, consisting of q^L sequences, hence taking exponential time. The calculation of exact results is thus infeasible, especially for large data sets.
Different strategies have been proposed to approximate the model parameters, nicely reviewed in [10]. In the following we will present a variety of approaches to ascertain the parameters, with emphasis on the mean field approximation and its modifications.

Maxmum Likelihood

The inverse Potts problem involves the reconstruction of unknown parameters via statistical inference. Basis for many methods to infer these parameters is the **maximum likelihood** framework.
We suppose that the variables $X = (X_1, \ldots, X_n)$ are drawn i.i.d. from the density function $P(x;\theta)$. For fixed parameter set θ the joint density of X is equal to the product of the single densities:

$$P(x;\theta) = \prod_{i=1}^{n} P(x_i;\theta), \tag{4.21}$$

the **probability** of observing the data $x_1, \ldots, x_n$ with given parameter θ. The same function viewed as a function of θ at fixed data x refers to the **likelihood** $L(\theta;x)$ of the parameters θ taking certain values given the observed data x [10].
The maximum likelihood estimator (MLE) finds the values for the model parameter θ^{ML} that maximise the likelihood function $L(\theta;x)$, hence the parameter set that make the data most probable:

$$\theta^{\text{ML}} = arg \max_{\theta} L(\theta; x_1, \ldots, x_n). \tag{4.22}$$

With finite sampling, it is improbable to determine θ exactly. However, the MLE possesses a number of attractive limiting properties with infinitely increasing sample size,

which allow to converge to the parameters θ [135]. Consistency is one of these properties. It means that if a sufficiently large number of data were generated by the common density function $P(x;\theta)$, the sequence of estimators θ^{ML} converges in probability to the original parameters θ with arbitrary precision.
To avoid the computation with very small numbers, because the likelihood scales exponentially with the number of samples, it is more convenient to maximise the logarithm of the likelihood. This leads to the same parameter estimates since the logarithm is a strictly monotonic function. Applied to the inverse Potts problem, where M configurations $S = \{s_1^1, \dots, s_L^1, \dots, s_1^M, \dots, s_L^M\}$ are sampled from the Boltzmann distribution, cf. Eq. (4.18), the log-likelihood is given by

$$\begin{aligned} L_S(h,e) &= \frac{1}{M} \ln L(h,e;S) \\ &= -\ln Z + \sum_{i=1}^{L} \sum_{A} h_i(A) \cdot f_i(A) + \sum_{1 \le i < j \le L} \sum_{A,B} g_{ij}(A,B) \cdot f_{ij}(A,B). \end{aligned} \tag{4.23}$$

The log-likelihood depends on the parameters of the Potts model, as well as the magnetisations and the correlations between pairs of spins. The sample averages provide sufficient statistics to determine the model parameters. Hence, the observed pairwise frequencies f_{ij} and single site frequencies f_i can be used to determine the parameters h and g.

Interpreting the log-likelihood in the context of thermodynamics, it can bee seen as the entropy of the Potts model [10]. The first term of the equation is known as the free energy and the latter two terms are the sample average of the energy. Approaches to approximate the model parameters utilise this property, thus we consider this relation in more detail.

Thermodynamic Potentials for the Inverse Potts Problem

In statistical physics, the introduction of thermodynamic potentials helps to solve certain problems. Thermodynamics explain the behaviour of physical systems by giving the relation between the temperature, entropy, internal energy (the total energy of the system), and pressure. The thermodynamic state of such a system is described by potentials including the Gibbs and Helmholtz free energies. They describe quantitative measures of the stored energy in a system, depending on different variables and constraints (i.e. fixed variables), respectively.
Applying a **Legendre transformation** (see Infobox 4.1) allows to convert between these potentials and to change the independent variables, hence to perform calculations with the more convenient potential. For the Ising and the Potts model, the MLE of the system parameters corresponds to such a transformation of thermodynamic potentials [10].

Regarding the forward problem, i.e. magnetic fields h and couplings g are given, the derivatives of the **Helmholtz free energy**

$$\mathcal{F}(h,e) = -\ln Z(h,e) \tag{4.24}$$

allow to determine observables, i.e. the first and second moments of the distribution. In the context of thermodynamics, we will refer to magnetisations m and pair correlations χ. Thus, we can note here that all necessary information on the marginals can by derived from the partition function Z from Eq. (4.20) [9]:

$$\frac{\partial \mathcal{F}(h,e)}{\partial h_i(A)} = \frac{\partial \ln Z}{\partial h_i(A)} = -m_i(A) \tag{4.25}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{F}(h,e)}{\partial g_{ij}(A,B)} = \frac{\partial^2 \ln Z}{\partial g_{ij}(A,B)} &= \frac{\partial^2 \ln Z}{\partial h_i(A)\partial h_j(B)} \\ &= -\chi_{ij}(A,B) + m_i(A)m_j(B) \equiv -C_{ij}(A,B)\end{aligned} \tag{4.26}$$

The derivative with respect to g can also be generated by a partial differentiation with respect to h. It implies the connected correlations (covariance) C, with $C_{ij} \equiv \chi_{ij} - m_i m_j$.

Infobox 4.1: Legendre transformation

Aim: find transformation with $f(x) \underset{y=\frac{\partial f}{\partial x}}{\longrightarrow} \tilde{f}(y)$

Legendre transformation:

1. $y = \frac{\partial f}{\partial x}$
2. $\tilde{f}(y) = x(y) \cdot y - f\left(x(y)\right)$

Properties:

- all information for (re–)transformation are given in the function (no external information necessary)
- transformation is own inverse transformation
  $T^2 = \mathbb{1},\ T = T^{-1}$
- simple differentials
  $df = \frac{\partial f}{\partial x} dx = y\, dx$
  $d\tilde{f} = d(xy - f) = dx\, y + x\, dy - df = x\, dy$
- allows to convert problem into a representation with more controllable variables, which is more convenient to handle (particularly useful in thermodynamics)

Example:
Equation:

$$f(x) = (x-a)^2$$

$$\longrightarrow y = \frac{\partial f}{\partial x} = 2(x-a)$$

Transformation:

$$\tilde{f}(y) = x(y) \cdot y - f\left(x(y)\right) = \left(\frac{y}{2} + a\right) \cdot y - \left(\frac{y}{2}\right)^2 = ay + \frac{y^2}{4}$$

$$\longrightarrow z = \frac{\partial \tilde{f}}{\partial y} = a + \frac{y}{2}$$

Inverse transformation:

$$\tilde{\tilde{f}}(z) = y(z) \cdot z - \tilde{f}\left(y(z)\right) = 2(z-a) \cdot z - \left(2a(z-a) + \frac{\left(2(z-a)\right)^2}{4}\right)$$

$$= (z-a)^2 \underset{z=x}{=} (x-a)^2 = f(x)$$

The inverse problem, i.e. pair correlations χ and magnetisations m are given, can be solved by considering the Legendre transform of the Helmholtz free energy with respect to both fields and couplings:

$$\mathcal{S}(m,\chi) = -\sum_{i=1}^{L}\sum_{A} h_i(A)m_i(A) - \sum_{1\leq i<j\leq L}\sum_{A,B} g_{ij}(A,B)\chi_{ij}(A,B) - \mathcal{F}(h,e). \tag{4.27}$$

It is also recognised as the entropy function and corresponds to the negative maximum likelihood of the model parameters in Eq. (4.23).
The fields and couplings can be derived by differentiation

$$h_i(A) = -\frac{\partial\mathcal{S}(m,\chi)}{\partial m_i(A)} \tag{4.28}$$

$$g_{ij}(A,B) = -\frac{\partial\mathcal{S}(m,\chi)}{\partial\chi_{ij}(A,B)}. \tag{4.29}$$

The inverse transformation of Equation (4.27) is given by

$$\mathcal{F}(h,e) = -\sum_{i=1}^{L}\sum_{A} h_i(A)m_i(A) - \sum_{1\leq i<j\leq L}\sum_{A,B} g_{ij}(A,B)\chi_{ij}(A,B) - \mathcal{S}(m,\chi). \tag{4.30}$$

The thermodynamics of the inverse problem can be reduced to a single Legendre transform, because as mentioned above, the derivatives of the Helmholtz potential w.r.t g can be attained with the partial differentiation w.r.t. h (cf. Eq. (4.26)). The reduced Legendre transformation thus yields the **Gibbs free energy** [10]:

$$\mathcal{G}(m,e) = \sum_{i=1}^{L}\sum_{A=1}^{q-1} h_i(A)m_i(A) - \mathcal{F}(h,e), \tag{4.31}$$

which depends now on the single site marginal m and couplings g. Note, that the Potts variables A only run until $q-1$, as the q^{th} variable is set to zero for gauge fixing (Section 4.3.1).
The local field can be found by the first derivative of the Gibbs potential

$$h_i(A) = -\frac{\partial\mathcal{G}(m,e)}{\partial m_i(A)}. \tag{4.32}$$

Using the following inverse function theorem

$$\left[\frac{\partial(h_1,\ldots,h_N)}{\partial(m_1,\ldots,m_n)}\right]_{ij} = \left[\left(\frac{\partial(m_1,\ldots,m_N)}{\partial(h_1,\ldots,h_n)}\right)^{-1}\right]_{ij} \tag{4.33}$$

and taking the determination of the covariance matrix in Equation (4.26) into account leads to the inference of the couplings with the second derivatives of the Gibbs potential:

$$\frac{\partial^2\mathcal{G}(m,e)}{\partial m_i(A)\partial m_j(B)} = \frac{\partial h_i(A)}{\partial m_j(B)} = (C_{ij}^{-1})(A,B). \tag{4.34}$$

This result turns out to be central to many methods for the inverse Ising and Potts problem [10]. As soon as the Gibbs potential (Eq. (4.31)) can be evaluated, the couplings are easily to derive. Plugging these coupling estimates and single site marginals into

Eq. (4.32) allows the determination of the magnetic fields h and thus the complete reconstruction of the parameters for the model.

Mean Field Approximation

Mean field (MF) theory, also known as self-consistent field theory, has its origin in statistical physics where problems with high dimensional probability distributions need to be approximated. Certain dependencies between random variables are neglected to derive a closed set of equations for these variables resulting in tractable approximations for the computation of high dimensional sums and integrals in probabilistic models [136]. There exist different methods to approach the mean field theory [10]:

In the **variational approach**, the true intractable distribution is approximated by a tractable one, which is factorised in the sites, thus making the different spin variables statistically independent from each other. The probability has to minimise a certain distance measure, e.g. the Kullback-Leibler distance (the relative entropy).

The **Field Theoretic approach** replaces discrete expectations over the random variables by integrals of auxiliary field variables to be able to use better performing approximation methods for these integrals, like Laplace transformation or saddle-point methods.

In this section we want to focus on the approach introduced by Plefka et. al [137]. Here they used the method of Taylor expansion around zero coupling, the so-called **small coupling expansion**, of the Gibbs free energy in a Sherringtion-Kirkpatrick (SK) model. The aim is to approximate the unknown distribution (i.e. the parameters) from which the known magnetisations arise. Another derivation was given by Georges and Yedida [138] and was extended by Morcos et al. [9] for the Potts model with $q > 2$.
In this approach they introduce an additional parameter $\alpha \in [0, 1]$, which affects couplings g, and allows to interpolate between independent variables (for $\alpha = 0$) and the original model (for $\alpha = 1$).
The Gibbs potential, cf. Eq. (4.31), with the perturbed Hamiltonian $\mathcal{H}(\alpha)$ is given by

$$\begin{aligned} \mathcal{G}(\alpha) &= -\ln\left[\sum_{\sigma} e^{-\mathcal{H}(\alpha)}\right] + \sum_{i=1}^{L}\sum_{A=1}^{q-1} h_i(A)m_i(A) \\ &= -\ln\left[\sum_{\sigma} e^{\sum_{i=1}^{L} h_i(\sigma_i) + \alpha\sum_{1\le i<j\le L} g_{ij}(\sigma_i,\sigma_j)}\right] + \sum_{i=1}^{L}\sum_{A=1}^{q-1} h_i(A)m_i(A). \end{aligned} \tag{4.35}$$

The Gibbs potential is approximated with the Taylor expansion. More precisely, the independent site case ($\alpha = 0$) of the potential is expanded by the Taylor (or Maclaurin) series:

$$\mathcal{G}(\alpha) = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{d^n\mathcal{G}(\alpha)}{d\alpha^n}\,|_{\alpha=0}\,\alpha^n. \tag{4.36}$$

It turns out that the zeroth and first order terms of the expansion

$$\mathcal{G}(\alpha) = \mathcal{G}^{(0)}(\alpha) + \mathcal{G}^{(1)}(\alpha) = \mathcal{G}(0) + \frac{d\mathcal{G}(\alpha)}{d\alpha}\,|_{\alpha=0}\,\alpha + \mathcal{O}(\alpha^2) \tag{4.37}$$

reveal the **mean field theory** and will be explained below. The second order term corresponds to the TAP free energy, what we refer to in Section 4.3.2.

The first part of Eq. 4.37 denotes the Gibbs potential for a non-interacting system. In statistical mechanics, the free energy equals the average energy (average Hamiltonian) minus the entropy. If $\alpha = 0$, the average energy is removed during the Legendre transformation, so the negative entropy of the marginals for the uncoupled spins remains:

$$\begin{aligned}\mathcal{G}^{(0)}(\alpha) = \mathcal{G}(0) &= \sum_{i=1}^{L}\sum_{A=1}^{q} m_i(A)\ln m_i(A)\\ &= \sum_{i=1}^{L}\sum_{A=1}^{q-1} m_i(A)\ln m_i(A) + \sum_{i=1}^{L}\left(1-\sum_{A=1}^{q-1} m_i(A)\right)\ln\left(1-\sum_{A=1}^{q-1} m_i(A)\right).\end{aligned} \tag{4.38}$$

The terms for $A = q$ are removed to include only independent variables (cf. Section 4.3.1).

The first order of Eq. (4.36) results in the average coupling term in the Hamiltonian (the mean field), and for $\alpha = 0$, hence

$$\mathcal{G}^{(1)}(\alpha) = \frac{d\mathcal{G}(\alpha)}{d\alpha}\,|_{\alpha=0} = -\sum_{1\le i<j\le L}\sum_{A,B} g_{ij}(A,B)m_i(A)m_i(B). \tag{4.39}$$

If Equations (4.38) and (4.39) are plugged into Eq. (4.35) the first order approximation of the Gibbs potential is found. Recalling Eq. (4.32) the local fields h can be determined by the first derivative of the Gibbs potential with respect to the marginals m providing the self-consistency equation

$$\frac{\partial\mathcal{G}(\alpha)}{\partial m_i(A)} = h_i(A) = \ln\left(\frac{m_i(A)}{m_i(q)}\right) - \sum_{1\le i<j\le L}\sum_{B\neq q} g_{ij}(A,B)m_i(B). \tag{4.40}$$

The second derivative (cf. Eq. (4.34)) yields the inverse covariance matrix and can be solved for couplings g:

$$\frac{\partial^2\mathcal{G}(\alpha)}{\partial m_i(A)m_j(B)} = \frac{\partial h_i(A)}{\partial m_j(B)} = (C^{-1})_{ij}(A,B) = -g_{ij}(A,B). \tag{4.41}$$

This last equation provides the solution for the problem of parameter inference in only one single step, instead of iterative schemes for parameter fitting. This approach of mean field approximation allows to fit the one- and two-site marginal of the model $P(\sigma)$ with the empirical frequencies $f_i(A)$ and $f_{ij}(A,B)$ from data samples. In order to determine the couplings g, the empirical covariance matrix

$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B) \tag{4.42}$$

has to be inverted. The matrix inversion with a time complexity of $\mathcal{O}(L^3)$ is much simpler to compute than the maximum likelihood where the parameters have to be fitted iteratively with an exponentially large number of steps to calculate the partition function and its derivatives.

To use only the zeroth and first order to determine the parameters is called the **naive mean field** approach.

Thouless-Anderson-Palmer Equation

As already implied in the last section, the second order expansion of the Gibbs potential (cf. Eq. (4.36)) leads to the TAP equations. TAP was introduced in 1977 by Thouless, Anderson, and Palmer [139] derived for the Sherrington-Kirkpatrick (SK) model in high temperature (which means $\alpha = 1$):

$$\begin{aligned} \mathcal{G}^{(2)}(\alpha) &= \frac{1}{2}\frac{d^2\mathcal{G}(\alpha)}{d\alpha^2} \\ &= \frac{1}{2}\sum_{1\leq i<j\leq L}\sum_{A,B} g_{ij}(A,B)^2\left(1-m_i(A)^2\right)\left(1-m_j(B)^2\right). \end{aligned} \tag{4.43}$$

The added term, the so called Onsagar correction term, can be interpreted as the effect of fluctuations of a spin variable on its magnetisation resulting from their impact on neighbouring spins [139].
Adding this correction leads to a self-consistency equation for the local field, according to Eq. (4.32), of

$$\begin{aligned} h_i(A) =& \ln\left(\frac{m_i(A)}{m_i(q)}\right) - \sum_{1\leq i<j\leq L}\sum_{B\neq q} g_{ij}(A,B)\cdot m_i(B) \\ &+ \sum_{1\leq i<j\leq L}\sum_{B\neq q} g_{ij}(A,B)^2\cdot m_i(A)\cdot\left(1-m_j(B)^2\right). \end{aligned} \tag{4.44}$$

The differentiation with respect to m_j (cf. Eq (4.34)) yields

$$(C^{-1})_{ij}(A,B) = -g_{ij}(A,B) - 2m_i(A)m_j(B)g_{ij}(A,B)^2. \tag{4.45}$$

Solving for couplings g we get

$$g_{ij}(A,B) = \frac{\sqrt{1-8m_i(A)m_j(B)(C^{-1})_{ij}(A,B)}-1}{4m_i(A)m_j(B)}. \tag{4.46}$$

This can be used to retrieve the local fields h in Eq. (4.44).

4.3.3 Scoring Schemes in DCA

For DCA in a biological context, different scoring schemes have been derived for the prediction of sequence structures, reviewed in [132]. In [9, 127] the **Direct Information (DI)** was introduced, which is equivalent to the MI score (Section 4.2). Instead of using the local frequencies of co-occurring symbols $f_{ij}(A,B)$, the inferred site specific probabilities

$$P_{ij}(A,B) = \frac{1}{Z_{ij}}\exp\left(g_{ij}(A,B)+h_i(A)+h_j(B)\right) \tag{4.47}$$

are inserted, yielding the DI score

$$DI_{ij} = \sum_{A,B} P_{ij}(A,B)\ln\left(\frac{P_{ij}(A,B)}{f_i(A)\cdot f_j(B)}\right). \tag{4.48}$$

This score is invariant with respect to the choice of gauge fixing (cf. Section 4.3.1), the resulting score is always the same.

The **Frobenius norm (FN)** of the coupling terms g_{ij} is given by

$$\left\| g_{ij} \right\|_2 = \sqrt{\sum_{A,B} g_{ij}(A,B)^2}. \tag{4.49}$$

Here, the choice of the gauge fixing method is important. The most appropriate variant is the zero-sum gauge, which minimises the FN [127].
To overcome a bias due to phylogeny and undersampling the authors of [118] introduce the **average product correction (APC)** of a norm rank.
This correction for the Frobenius norm was used in several studies [128, 140], after Jones et. al. [129] introduced the idea of using the APC of the 1-norm for couplings.
In [141], this scoring alternative was also used for the inference of RNA structures and was termed **evolutionary couplings (EC)**:

$$EC_{ij} = \left\| g_{ij} \right\|_2 - \frac{\| g_{i\star} \|_2 \left\| g_{\star j} \right\|_2}{\| g_{\star\star} \|_2}. \tag{4.50}$$

The inferred interactions indicate sites in close proximity within the sequence structure. These physical contact predictions can be used to derive secondary and tertiary structures of the molecule, as illustrated in Figure 4.4. For example in [141], the authors used the tool Nucleic Acid Simulation Tool (NAST) [142], which is a coarse-grained modeller for sequence structures. Feeding NAST with 2d and 3d contact predictions as input, allows the algorithm to simulate structure folding, resulting in coarse-grained structure models. Final candidates are picked according to the lowest folding energy and are used to ascertain all-atom structures.

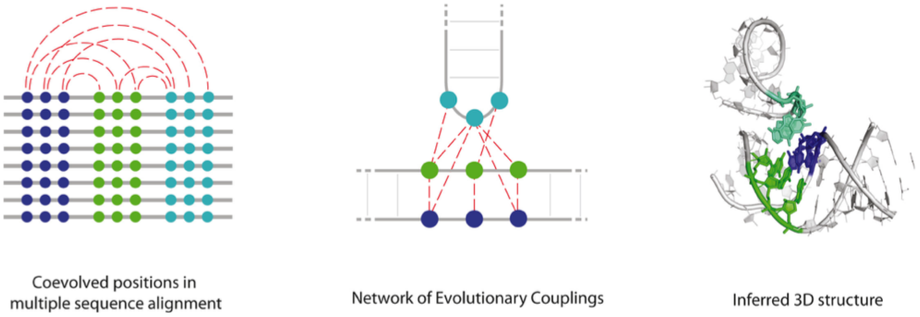

Coevolved positions in multiple sequence alignment
Network of Evolutionary Couplings
Inferred 3D structure

FIGURE 4.4: Inferring RNA structures with evolutionary couplings. Interacting nucleotides in close proximity within a sequence structure may co-evolve in order to retain a functional RNA. These interactions can by detected in MSAs with DCA methods and allow to reveal the RNA secondary and tertiary structure. Figure taken from [141].

Chapter 5

Quantifying Functional Constraints

In the chapter before, we discussed methods to infer evolutionary constraints from multiple sequence alignments of homologous sequences. The sequences originate from a common ancestor and possess information about domains which are crucial for the viability of the regarded organism. Thus, these domains remain conserved since changes lead to impaired structures and functionality, and consequently to a lower fitness, and might not be competitive with the established species.
Within the MSA, only viable sequences are included. The more dominant a geno– or phenotype is, i.e. the higher the contribution to the gene pool, the more likely it is to observe a certain species, and hence to find it within the MSA. This makes it a challenging task to quantify and to comparatively assess functionally important sites without this sampling bias. Moreover, intermediate forms are overpowered and never or hardly observed, but may give insights on which configurations or which particular positions are the bottleneck to keep a vital organism.
Several approaches have been presented to characterise functional domains and structure-function-relations with mutation experiments, and quantifying RNA-RNA or RNA-Protein interactions [143–146]. Most of these classical approaches are time-consuming or require substantial experience and equipment.

In this chapter, we first present the method of Smyth et al. [8], the Mutational Interference Mapping Experiment (MIME). This method enables the detection of regions and structures crucial for certain functions of non-coding RNAs. With a single experiment it is possible to quantify the impact that a mutation has on these functions for each nucleotide.
The samples, which are conducted in MIME experiments, are sequenced with next generation sequencing (NGS). The succeeding data preparation and analysis, performed with the software tools presented in Chapter 6, considers properties and challenges due to the next generation sequencing process.
Therefore, we briefly explain NGS, focusing on the Illumina sequencing techniques.

5.1 Mutational Interference Mapping Experiment

The mutational interference mapping experiment (MIME) [8], is a time- and cost-efficient experimental method to detect domains and structures important for RNA function at single site resolution. In MIME, a target RNA is randomly mutated, selected by function, physically separated, and sequenced using NGS, as depicted in Figure 5.1.

The mutation frequencies in the functionally selected vs unselected pools contain information about the function and structural commitment of each single nucleotide within the analysed RNA. The analytical approach allows the recovery of quantitative parameters and permits the identification of base pairing partners directly from the sequencing data.

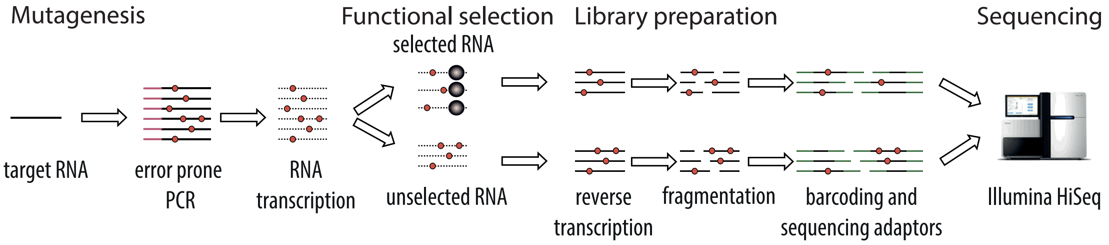


FIGURE 5.1: Experimental pipeline for MIME: random introduction of mutations into an RNA target, physical separation of functional from non-functional RNA, and quantification of RNA mutations in each population using NGS. Figure adapted from [8].

The input for the analysis of the MIME-generated data consists of base counts at each position, respectively for the pool of functionally selected- and the non-selected sequences. The nucleotide occurrences are derived after mapping the NGS reads to the reference, the original wild type sequence. These counts are translated into a quantitative effect on the function associated with each particular mutation m at each nucleotide position i, relative to the functionality of the wild type. To assess the statistical significance of that effect, a resampling-like procedure is applied that can be obtained from the given input data.
Since the experimental procedure (sequencing and reverse transcription) may introduce a substantial number of falsely detected mutations, these errors need to be quantified. Control experiments, conducted without mutagenesis, allow to determine the error rates. The relative effects are corrected for these errors and a signal-to-noise ratio can be derived.
In a last step, the results are evaluated according to several quality criteria, including the statistical significance and signal-to-noise ratio.

In a first *in vitro* application, the binding domain of the HIV-1 Pr55$^{\text{Gag}}$ protein to the 5′ region of the viral genomic RNA could be identified with MIME at single nucleotide resolution and the RNA structural elements promoting this interaction [8]. Subsequently, MIME could be adapted to *in cellulo* experiments [12], defining elements within the 5′ region of the HIV-1 genomic RNA that regulate viral genomic RNA production, as well as motifs required for gRNA packaging into virions, explained more fully in Chapter 7.
In the following section we describe the mathematical background of the MIME framework, as an example with the affinity of an RNA binding to a certain protein, as presented in [8].

5.1.1 Relation between Nucleotide Frequency and Relative Kd Values

The basic reaction scheme underlying the competitive binding experiment that separates RNA by binding affinity is shown in Fig. 5.2. In the graphic, the differentially colored $\star$ symbols indicate the presence of a particular mutation at a specific nucleotide position in the RNA. The mass-action kinetics describing this experiment are given by:

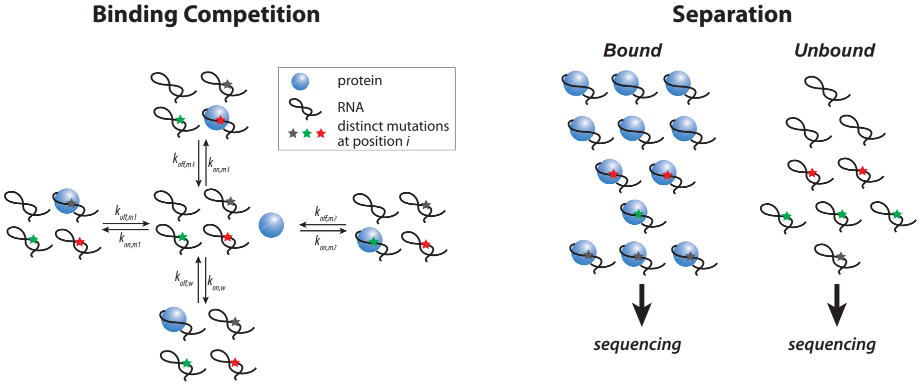


FIGURE 5.2: Reaction scheme underlying the competition experiment that selects RNA by binding affinity (left). Bound- and unbound RNAs are separated, prepared for sequencing and sequenced to obtain mutation frequencies (right). Graphic taken from [8], Supplementary Figures.

$$\frac{\mathrm{d}}{\mathrm{d}t}S_w^b(i) = S_w^u(i) \cdot B \cdot k_{\mathrm{on},w}(i) - S_w^b(i) \cdot k_{\mathrm{off},w}(i) \tag{5.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S_w^u(i) = -S_w^u(i) \cdot B \cdot k_{\mathrm{on},w}(i) + S_w^b(i) \cdot k_{\mathrm{off},w}(i) \tag{5.2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S_m^b(i) = S_m^u(i) \cdot B \cdot k_{\mathrm{on},m}(i) - S_m^b(i) \cdot k_{\mathrm{off},m}(i) \tag{5.3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S_m^u(i) = -S_m^u(i) \cdot B \cdot k_{\mathrm{on},m}(i) + S_m^b(i) \cdot k_{\mathrm{off},m}(i) \tag{5.4}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}B &= S_w^b(i) \cdot k_{\mathrm{off},w}(i) + S_{m_1}^b(i) \cdot k_{\mathrm{off},m_1}(i) + \ldots + S_{m_3}^b(i) \cdot k_{\mathrm{off},m_3}(i) \\ &\quad - B \cdot \left(S_w^u(i) \cdot k_{\mathrm{on},w}(i) + S_{m_1}^u(i) \cdot k_{\mathrm{on},m_1}(i) + \ldots + S_{m_3}^u(i) \cdot k_{\mathrm{on},m_3}(i)\right),\end{aligned} \tag{5.5}$$

where $S_w^b(i)$ denotes the concentration of bound RNA carrying a wild type base at nucleotide position i and $S_w^u(i)$ denotes the concentration of unbound wild type RNA. The subscript m represents the three possible mutations $m \in \{m_1, m_2, m_3\}$ and indicates the presence of one of the mutations at nucleotide position i. The parameter B denotes the free protein, and $k_{\mathrm{off}}(i)$, $k_{\mathrm{on}}(i)$ describe the respective rates of dissociation and association.
In a steady state condition, i.e. if there is sufficient time to obtain a binding equilibrium, the left hand side (the rate of change) of the equations becomes zero. Solving one equation for unbound protein B and substituting it, we get:

$$K_m(i) = \frac{Kd_m(i)}{Kd_w(i)} = \frac{S_m^u(i)}{S_w^u(i)} \cdot \frac{S_w^b(i)}{S_m^b(i)} \tag{5.6}$$

which denotes the **relative impact** of a particular mutation m at position i in the RNA sequence on the binding affinity in relation to the wild type. The **Kd value** describes the dissociation constant, i.e. the ratio of dissociation rate and association rate

$$Kd(i) = \frac{k_{\mathrm{off}}(i)}{k_o n(i)}. \tag{5.7}$$

5.1.2 Error Correction

Unfortunately, the numbers of bound and unbound sequences S are not known. Instead we derive NGS reads R, subject to errors X (sequencing and RT-errors). In order to account for this noise, we have to consider the relation between "sequence numbers" S and "reads" R: For any mutant nucleotide m at nucleotide position i, the number of NGS reads $R_m(i)$ in the samples is related with the RNA sequence numbers $S_m(i)$ via

$$S_m(i) = \frac{R_m(i)}{\nu} + \sum_{n \neq m} X_{m \to n}(i) - \sum_{n \neq m} X_{n \to m}(i), \tag{5.8}$$

where $X_{m \to n}(i)$ is a random variable that denotes the number of sequences that are in fact nucleotide m, but were falsely detected as some other base n. Likewise, $X_{n \to m}(i)$ is a random variable indicating the number of sequences that were originally some other nucleotide n, but were falsely detected as m. Parameter ν denotes the normalisation factor, i.e. the relative titration, if applied. It guaranteed that equal amounts of protein-bound and –unbound sequences were used in the NGS machinery.
The error probability $\kappa(i)$ is assumed to be very low (for Illumina $< 10^{-3}$) and may not be vastly different for the distinct types of transitions $n \to m$. Thus, false detections, which do not affect the concentration significantly can be neglected, resulting in

$$S_m(i) \approx \frac{R_m(i)}{\nu} - X_{w \to m}(i) \text{ and} \tag{5.9}$$

$$S_w(i) \approx \frac{R_w(i)}{\nu}. \tag{5.10}$$

Assuming that the noise X is multinomially distributed, i.e. $X_{w \to m} \sim \mathcal{M}(S_w, \kappa_{w \to m})$, the expectation value for the number of mutations m can be estimated with

$$\mathbb{E}(X_{w \to m}(i)) = S_w(i) \cdot \kappa_{w \to m}(i) \approx \frac{R_w(i)}{\nu} \cdot \kappa_{w \to m}(i), \tag{5.11}$$

where $\kappa_{w \to m}(i)$ denotes the probability of falsely detecting a wild type residue at position i as mutation m.
To estimate the error probability $\kappa_{w \to m}(i)$, experiments with unmutated RNA are conducted in parallel. This means, the number of sequences with mutations are zero (for both selected and non-selected samples), i.e. $S_{m_1}(i) = S_{m_2}(i) = S_{m_3}(i) = 0 \; \forall \; i$. Thus, the (falsely) detected mutations of reads R arise from the RT- and sequencing procedure.

Substituting the expected noise of Eq. (5.11) into Eq. (5.9) and plug Eqs. (5.9) and (5.10) into the Kd estimation proposed in Eq. (5.6) yields then

$$\frac{Kd_m}{Kd_w}(i) \approx \frac{\frac{R^u_m(i)}{R^u_w(i)} - \kappa^u_{w \to m}(i)}{\frac{R^b_m(i)}{R^b_w(i)} - \kappa^b_{w \to m}(i)}. \tag{5.12}$$

5.1.3 Statistical Evaluation

Note, that the above described procedure yields one single point estimate for each possible mutation at each nucleotide position within the RNA of length L for the impact on the function. Although it is possible to infer the effect from the simple method above, it is not possible to assess the statistical certainty of these estimates, unless vast numbers of repetition experiments are performed. However, this procedure is expensive and time-consuming.
Instead, the statistical certainty of the relative Kd estimates is inferred with a method based on resampling, inherent in the data.
The basic idea is to determine relative impact of a mutation $K_{m,w}(i,j)$ for each combination of nucleotide positions $i \neq j$, where the first residue i is mutated and the second residue

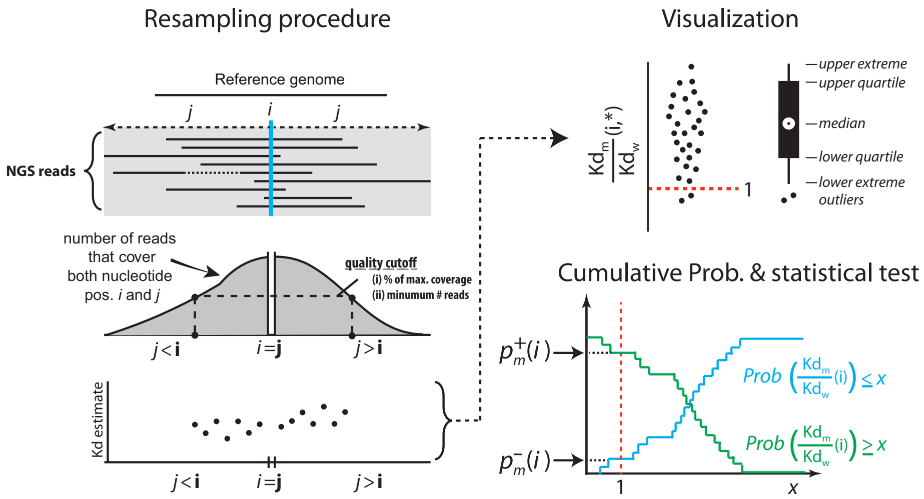


FIGURE 5.3: Schemec of the resampling procedure. Each pair of positions (i, j) yields a Kd estimate (left). The resampling distribution for $\frac{Kd_m}{Kd_w}(i)$ can be visualised (right, top) and statistical tests can be performed in order to detect significantly increasing or decreasing binding affinity of a mutant in relation to the wild type with a certain p-value (right, bottom). Graphics taken from [8], Supplementary Figures.

j is in the wild type configuration, thus having no effect on the relative Kd estimate. This allows to re-estimate the effect of a mutation m at position i N-times, with $N \leq L - 1$. Since less sequence fragments will cover both i and j the further j lies away from i (see Fig. 5.3, left panel), the procedure is highly similar to a classic jack-knife resampling procedure. Here, a re-estimation is performed each time after removing one individual from the pool of samples. The resampling will then give a non-parametric and unbiased probability distribution of the estimate derived in Eq. (5.12)

$$K_{m,w}(i,j) = \frac{Kd_{m,w}(i,j)}{Kd_{w,w}(i,j)} \approx \frac{\frac{R^u_{m,w}(i,j)}{R^u_{w,w}(i,j)} - \kappa^u_{w\to m,w}(i,j)}{\frac{R^b_{m,w}(i,j)}{R^b_{w,w}(i,j)} - \kappa^b_{w\to m,w}(i,j)}. \tag{5.13}$$

The resampling scheme also allows to estimate the probability of falsely detecting a wild type residue at position i as some mutant m (cf. Section 5.1.2) with statistical properties:

$$\mathbb{E}\left(\kappa_{w\to m}(i)\right) \approx \frac{1}{N} \sum_{j \neq i} \frac{R_{m,w}(i,j)}{R_{w,w}(i,j)}, \tag{5.14}$$

where N denotes the number of positions $j \neq i$, if the position pairs (i, j) have sufficient read coverage. Thus, following the resampling scheme for the relative Kd, we can estimate a confidence range for the error probability $\kappa_{w\to m}(i)$.
Note: The coefficient of variation of the error rate distribution is assumed to be small in the selected and unselected sample, respectively. Hence, a more reliable estimate can be inferred (law of large numbers). This justifies the use of the estimate $\mathbb{E}\left(\kappa_{w\to m}(i)\right)$ for the error correction of all Kd estimates instead of the single error resample $\kappa_{w\to m,w}(i,j)$.

5.1.4 Quality Criteria

For each position i and mutation m it is possible to resample a distribution of effects with the scheme explained above. Different criteria can be applied in order to filter these estimates to improve the result.

Evaluable Signal
One of the filters conducts the signal-to-noise ratio, which can be assessed with the data from the control experiments also used for the error correction. For each of the N resamples for position i and mutation m we compute the signal-to-noise ratio by

$$D_{m,w}(i,j) = \frac{R_{m,w}(i,j)}{\nu \cdot \mathbb{E}(X_{w\to m,w}(i,j))} \approx \frac{R_{m,w}(i,j)}{R_{w,w}(i,j) \cdot \kappa_{w\to m,w}(i,j)}$$
$$\approx \frac{R_{m,w}(i,j)}{R_{w,w}(i,j) \cdot \mathbb{E}(\kappa_{w\to m}(i))}. \tag{5.15}$$

If the ratio is below a user-supplied threshold, both in the bound and unbound samples, the corresponding Kd estimate $K_{m,w}(i,j)$ is discarded. If the signal is below the threshold at either the bound or unbound samples, the respective estimate is tagged as either being a lower– or upper estimate of $K_{m,w}(i,j)$. In this case, the value of the median of the estimate distribution $K_{m,w}(i,\star)$ is assigned. This has the following reason: If a mutation strongly decreases the Kd (increases the function), all sequences carrying this mutation may be bound. None, or too little amounts of sequence may remain unbound. Thus, $K_{m,w}(i,j)$ may not be accurately determined and may in fact be lower than estimable.

Coverage
For the resampling procedure, only positions j are evaluated where the total number of sequence fragments covering both i and j has at least a user defined percentage of the maximum coverage (middle panel in Fig. 5.3, left). Secondly, the total number of reads that have to be available (middle panel in Fig. 3, left) has to exceed a "minimum coverage" value. Both criteria together ensure that each resampling is based on a sufficient number of reads and thus provides meaningful estimates.

Number of resamplings
After applying the filters above, a certain number of estimates $K_{m,w}(i,j)$ remain and give rise to an empirical distribution (see Fig. 5.3, left (lower panel) and right (upper panel)). A minimum number of resamplings is required in order to reconstruct the empirical distribution with sufficient confidence (see Fig. 5.3, lower panel on the right).

Statistical test
The statistical test can subsequently be performed on the resampling distribution (Fig 5.3, lower panel on the right): To test whether a mutation at position i significantly increases the Kd/decreases the function, i.e. $\mathcal{H}_0 : K_m(i) \leq 1, \mathcal{H}_1 : Km(i) > 1$, the raw p-value (= probability of type I error/false rejection of the null hypothesis) can be computed according to:

$$p_m^-(i) = \frac{\#K_{m,w}(i,\star) \leq 1}{\#K_{m,w}(i,\star)}, \tag{5.16}$$

where # denotes the number of estimates and $\star$ indicates all N positions j that pass the quality criteria above. To test if mutation m at position i decreases the Kd/increases the function, the p-value is calculated according to:

$$p_m^+(i) = \frac{\#K_{m,w}(i,\star) \geq 1}{\#K_{m,w}(i,\star)}. \tag{5.17}$$

Note, that any threshold (e.g. 2-fold increase/decrease, etc.) can be tested. When several nucleotide positions i are assessed, test corrections need to be performed, e.g. the Benjamini-Hochberg false discovery rate method (BHFDR) [147].
There is a significant impact of mutation m at nucleotide position i, if the corrected p-value does not exceed a certain significance level α (false positives), which is also given by the user.

5.1.5 Detecting Functional Structures with MIME

The MIME framework is not only able to detect single positions which play an important role for the function. It also allows to assess whether this residue is directly involved in the function (e.g. forming the binding site) or indirectly, i.e. as part of the structure within the RNA which is necessary for the function, more specifically, if it forms a base pair with another position.

The main concept for detecting interacting base pairs is illustrated in Fig. 5.4.

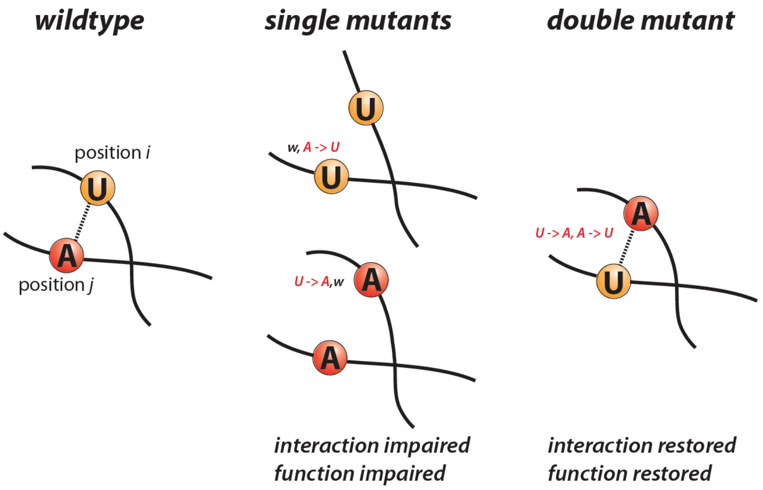


FIGURE 5.4: Two nucleotides interact in their wild type configuration in a way that is important for function (left). Single mutations at either position disrupt this interaction (middle) and thus impair RNA function. When both positions are mutated in a particular way, the interaction, and thus RNA function may be restored. Graphic taken from [8], Supplementary Figures.

If two positions i and j play a role for the function in their wild type configuration, certain single mutations at either position i or j would disrupt the function of the RNA (cf. the methods above). If these two positions are not interacting, the mutations of both positions may still impair the function, probably with an even stronger effect. In the case of correlation of these positions, certain configurations of double mutations of i and j may compensate the disrupting consequence of the single mutations. If the two positions are forming a base pair in the wild type configuration, which is important for the functional structure, the replacement with a base pair in the double mutant constellation may keep the structure and hence the function.
A measure for these interactions is called **epistasis** and is extensively elucidated in Chapter 3.3. The aim is to detect positive sign epistasis, which allows us to infer pairs of interacting sites where mutations at both sites i and j may restore a function that is impaired by the single mutants.
Epistasis E for the mutation pair (m_1, m_2) at the respective nucleotide positions (i, j) can

be computed according to

$$E_{m_1,m_2}(i,j) = \log\left(\frac{K_{m_1,w}(i,j) \cdot K_{w,m_2}(i,j)}{K_{m_1,m_2}(i,j)}\right), \tag{5.18}$$

where $K_{m_1,m_2}(i,j)$ denotes the relative Kd value for a RNA that harbours mutation m_1 at nucleotide position i and mutation m_2 at nucleotide j. $K_{m_1,w}(i,j)$ denotes the relative Kd of an RNA that exhibits mutation m_1 at position i and has the wild type residue at position j. Likewise, $K_{w,m_2}(i,j)$ denotes the relative Kd of an RNA that has the wild type residue at nucleotide position i and mutation m_2 at nucleotide position j.
If there is no epistasis ($E = 0$), the two residues are independent. An positive epistasis value $E > 0$ means that the double mutant restores the function, whereas an negative epistasis value $E < 0$ means that the double mutant amplifies the impact on the function.

Note, that Eq. (5.18) yields one point estimate for each pair of mutations m_1, m_2 at each pair of nucleotide positions (i,j). To assess the statistical certainty of these epistasis estimates, the resampling procedure, introduced in Section 5.1.3, has to be adapted. The basic idea is as before: In order to attain a probability distribution of the epistasis for positions (i,j), we compute epistasis values for the triplet of positions(i,j,k), where the first- and second residue (i,j) are mutated respectively and the third residue k is in the wild type configuration (see Fig. 5.5). This allows us to recompute the epistasis value for the pair

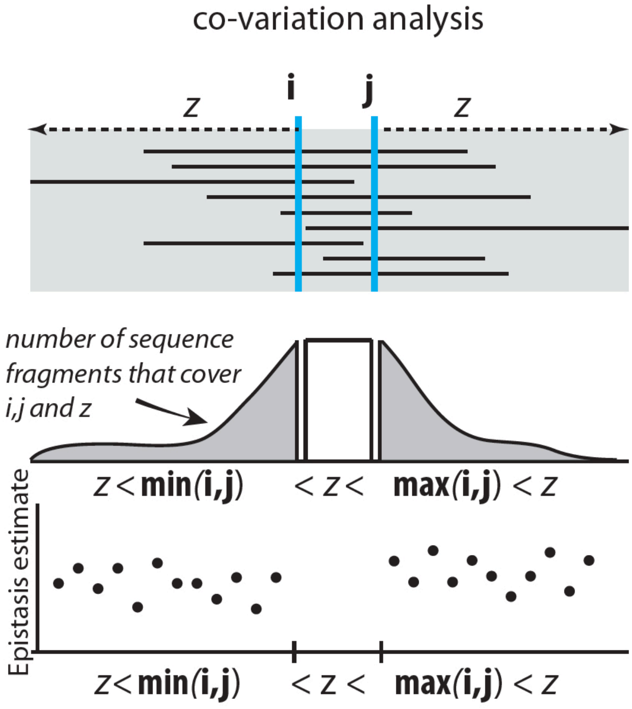


FIGURE 5.5: Resampling scheme for epistasis estimates. In order to resample the epistasis values $E_{m_1,m_2}(i,j)$ for the mutation pair m_1, m_2 at nucleotide positions i, j, values are computed for the triplet of positions (i,j,k). Position i and j are mutated to m_1 and m_2, respectively, and position k is wild type. Each triplet yields one epistasis estimate (bottom). Graphic taken from [8], Supplementary Figures.

(i, j) M times, with $\max(i, j) < k < \min(i, j)$:

$$E_{m_1,m_2,w}(i,j,k) = \log\left(\frac{K_{m_1,w,w}(i,j,k) \cdot K_{w,m_2,w}(i,j,k)}{K_{m_1,m_2,w}(i,j,k)}\right). \tag{5.19}$$

In order to accommodate the resampling method we will have to extend the previously described method (previous section) to be able to compute the relative Kd values for any triplet of positions.
As before, the number of selected/unselected sequences S are not known, but we have the number of NGS reads R instead, which are subject to errors X (sequencing and RT-errors). The RNA sequence number S harboring two mutations at positions i and j respectively and has a wild type residue at position k is related to the RNA read numbers R by

$$\begin{aligned} S_{m_1,m_2,w}(i,j,k) \approx {} & R_{m_1,m_2,w}(i,j,k) - X_{w\to m_1,m_2,w}(i,j,k) \\ & - X_{m_1,w\to m_2,w}(i,j,k) - X_{w\to m_1,w\to m_2,w}(i,j,k). \end{aligned} \tag{5.20}$$

Assuming a multinomial distribution of the noise as before, we get:

$$K_{m_1,m_2,w}(i,j,k) \approx \frac{\frac{R^u_{m_1,m_2,w}(i,j,k)}{R^u_{w,w,w}(i,j,k)} - Y^u_{m_1,m_2,w}(i,j,k)}{\frac{R^b_{m_1,m_2,w}(i,j,k)}{R^b_{w,w,w}(i,j,k)} - Y^b_{m_1m,m_2,w}(i,j,k)}, \tag{5.21}$$

with

$$\begin{aligned} Y_{m_1,m_2,w}(i,j,k) = {} & -\kappa_{w\to m_1,w\to,m_2,w}(i,j,k) \\ & + \frac{R_{w,m_2,w}(i,j,k)}{R_{w,w,w}(i,j,k)} \cdot \kappa_{w\to m_1,\star,w}(i,j,k) \\ & + \frac{R_{m_1,w,w}(i,j,k)}{R_{w,w,w}(i,j,k)} \cdot \kappa_{\star,w\to m_2,\star,w}(i,j,k). \end{aligned} \tag{5.22}$$

The following conditions hold for both the bound and unbound fractions:
$\kappa_{w\to m_1,\star,w}(i,j,z) \approx \kappa_{w\to m_1,\star}(i,j), \kappa_{\star,w\to m_2,\star,w}(i,j,k) \approx \kappa_{\star,w\to m_2,\star}(i,j)$.
The parameters $\kappa_{w\to m_1,\star}$ and $\kappa_{\star,w\to m_1}$ denote unconditional probabilities, namely one position was falsely detected as a mutant, regardless of the second position (indicated by the $\star$), thus the estimated error probabilities from Eq. (5.14) may be used. The same conditions holds for the error probability of detecting 2 mutations: $\kappa_{w\to m_1,w\to,m_2,w}(i,j,k) \approx \kappa_{w\to m_1,w\to,m_2}(i,j)$, which is approximated by

$$\kappa_{w\to m_1,w\to,m_2}(i,j) \approx \kappa_{w\to m_1,w}(i,j) \cdot \kappa_{w,w\to,m_2}(i,j). \tag{5.23}$$

Accordingly, the relative effects can be derived:

$$K_{m_1,w,w}(i,j,k) \approx \frac{\frac{R^u_{m_1,w,w}(i,j,k)}{R^u_{w,w,w}(i,j,k)} - \kappa^u_{w\to m_1,\star}(i,j)}{\frac{R^b_{m_1,w,w}(i,j,k)}{R^b_{w,w,w}(i,j,k)} - \kappa^b_{w\to m_1,\star}(i,j)} \tag{5.24}$$

$$K_{w,m_2,w}(i,j,k) \approx \frac{\frac{R^u_{w,m_2,w}(i,j,k)}{R^u_{w,w,w}(i,j,k)} - \kappa^u_{\star,w\to m_2}(i,j)}{\frac{R^b_{w,m_2,w}(i,j,k)}{R^b_{w,w,w}(i,j,k)} - \kappa^b_{\star,w\to m_2}(i,j)}. \tag{5.25}$$

5.2 Next Generation Sequencing

The samples, which are used in MIME experiments, are sequenced with **Next Generation Sequencing** (NGS). The term NGS comprises a variety of sequencing technologies processing millions of DNA fragments in parallel in one cycle. This allows for high-throughput sequencing with a drastic reduction of costs in comparison to classical sequencing techniques. The individual procedures from the DNA sample to the sequence can be grouped into three steps:

1. Library preparation
2. Clonal, parallel DNA amplification
3. Sequencing

In the following, we will briefly present these steps of the sequencing process, based on Illumina's Next Generation Sequencing Technology [148, 149].

5.2.1 General Prodedure

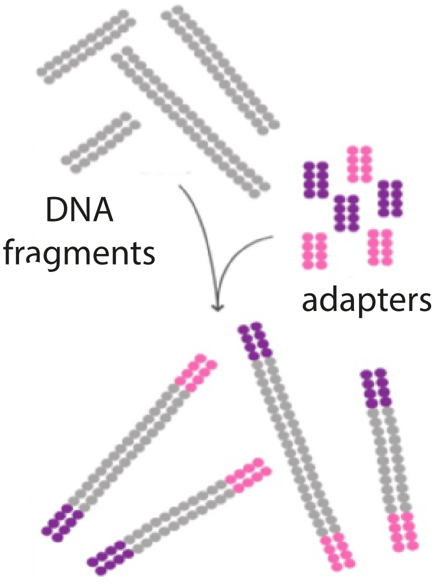


FIGURE 5.6: Library preparation. Genomic DNA is randomly fragmented and sequencing adapters are ligated to both ends. Figure adapted from [148].

Library preparation

The NGS library preparation (seen in Figure 5.6) starts with the random fragmentation of the investigated DNA or reverse transcribed RNA (cDNA) sample. Sequencing adapters, which are necessary for subsequent steps, are ligated to both ends of the fragments. The fragments are amplified with PCR and purified on gel.

Cluster Generation

The library is loaded into a flow cell, which contains a lawn of oligo nucleotides bound to the surface. These oligos are complementary to the library adapters, in order to capture the fragments by hybridisation, and function as primers for the amplification. Each bound fragment is amplified building clonal clusters through bridge amplification, as seen in Figure 5.7.

Sequencing

A flow cell containing millions of unique clusters is now loaded into the sequencer for automated cycles of extension and imaging. Illumina's method of Sequencing-by-Synthesis (SBS) uses the technique of Cyclic Reversible Termination (CRT), illustrated in Figure 5.8. In each cycle, DNA polymerase, and modified nucleotides are added to the flow cell. The four distinct nucleotides are marked with different fluorescent colours and contain a reversible terminator. All four nucleotides are added and synthesised by the DNA polymerase to the complementary DNA strand, which is attached to the flow cell surface.

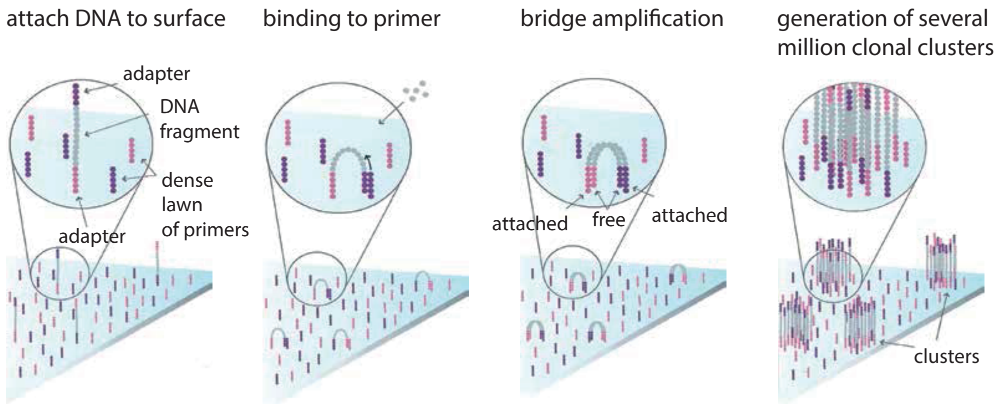


FIGURE 5.7: Cluster amplification. The fragments of the library are loaded into a flow cell and bind to complementary oligos on its surface. The fragments are amplified via bridge aplification building clonal clusters. Figure adapted from [148].

Only one nucleotide per cycle can be incorporated, since the terminator group is blocking the further synthesis.

After non-attached components are washed away, the synthesised fluorescent nucleotides are excited with a laser and the emitted colour of each cluster is captured in an image. Through imaging techniques, the respective nucleotide is recognised for each cluster.

In the end of the cycle, the terminator and fluorescence are removed and the next cycle starts. Nucleotide by nucleotide, the sequence for each cluster is determined at the same time.

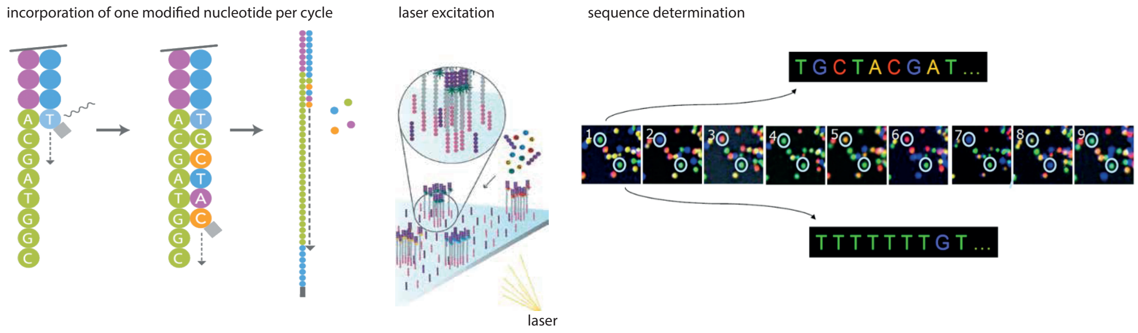


FIGURE 5.8: Base calling. During each sequencing cycle, the four labeled reversible terminators, and DNA polymerase are added to the flow cell. Only one nucleotide is appended to one emerging double strand per cycle, because of the terminator. After laser excitation, the emitted fluorescence from each cluster is captured and and the corresponding base is identified. Cycle by cycle, the sequences for each cluster are determined in parallel. Figures adapted from [148, 150].

5.2.2 Paired-End Sequencing

Depending on the biological issue, NGS facilitates the possibility of choosing between single-read sequencing and paired-end sequencing. The latter involves the sequencing from both ends of the DNA fragments (depicted in Figure 5.9), leading to double the amount of reads in the same time and effort of library preparation. For further data analysis, the reads are aligned to a reference genome (but can also be used for de-novo assembly). The major advantages of aligning the read pairs are higher accuracy of the alignment and the improved possibility of detecting single site nucleotide variants (SNVs), as it is desired in the MIME experiments explained above. Furthermore, it is possible to detect insertions and deletions (indels), as well as to remove PCR duplicates, which often occur during library preparation.

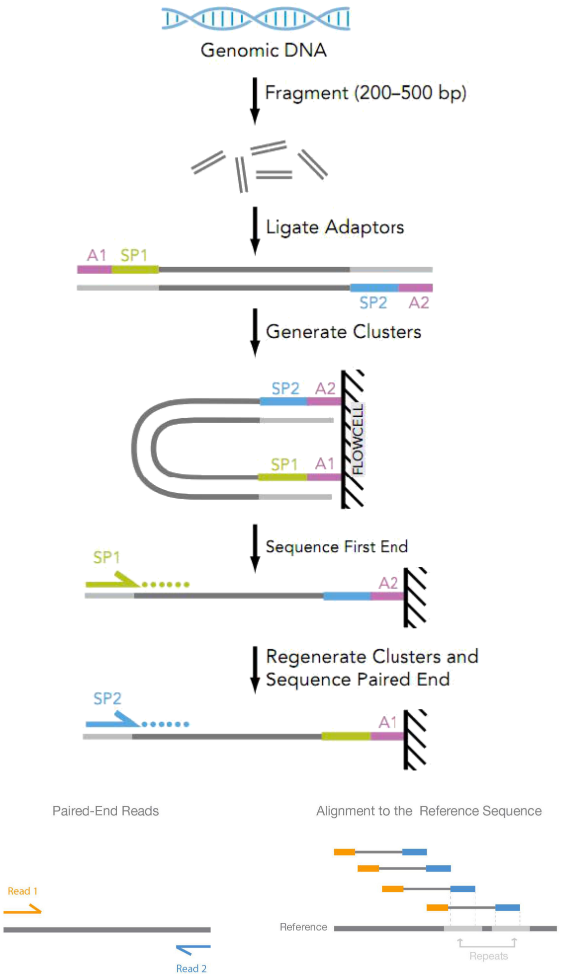


FIGURE 5.9: Paired-end sequencing with NGS. After the clonal clusters are sequenced from the first end, the clusters are regenerated and the complementary sequence is determined starting from the other end. The paired-end reads are aligned to the reference sequence. Figures taken from [149].

5.2.3 Quality Scoring

The determination of the single nucleotides in each cycle of the sequencing procedure, together with all other steps during next generation sequencing and further data processing, have an impact on the quality of the data. A very common score to measure the base calling accuracy of a sequencing platform is the Phred quality score (Q score). Historically, Phred scores were used to measure the quality of base calls with Sanger sequencing, comprising metrics such as peak resolution and shape to derive the quality by comparing results to multivariate lookup tables [151]. Although, the metrics for NGS platforms are different from those for Sanger sequencing, the algorithm for the Phred scoring scheme could be adapted to particular chemical properties [152]. The quality score lookup tables are constructed by exploiting the relevant parameters of a large empirical dataset of known accuracy.

The Q score is logarithmically related to the error probability P of predicting a wrong base [151, 153]: $Q = -10 \times \log_{10}(P)$.
Thus, the Q score denotes the base call accuracy, i.e. the probability of a correct base call. For example, if a Q score of 30 is given, the corresponding error probability is 1 in 1000 times to detect a wrong nucleotide, meaning 99.9% accuracy of the base call. In Table 5.1, Q scores and their corresponding accuracy are given.
A Q score of 30 is often considered a minimum quality score threshold for NGS. The detected sequence is reliable and most likely contains no error. Considering a lower quality score, for instance Q = 20, and hence a given error probability of 1 in 100, would mean an

TABLE 5.1: Quality scores and corresponding base call accuracy

Phred quality score	Probability of incorrect base call	Accuracy of base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

incorrect base every 100 nucleotides, i.e. the sequencing reads most likely contain errors. The range of the scores depend on the sequencing technology and the base calling algorithm. Recent Illumina sequencing reads can reach a quality score up to 41 for raw, unprocessed reads.

5.2.4 FASTQ

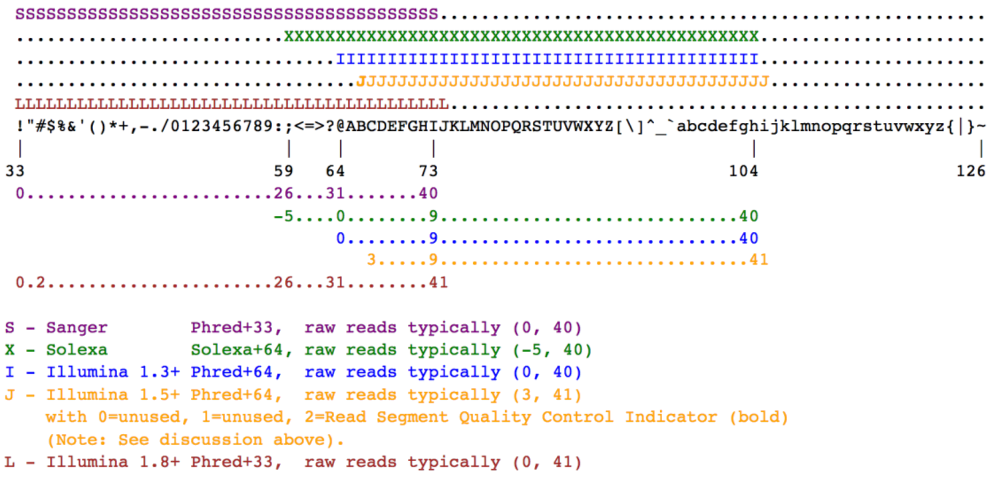
The resulting NGS reads are written into FASTQ files. For each read, the FASTQ file includes four lines containing

- a header starting with symbol "@", containing an identifier that usually gives information about the sequencing run and the cluster
- the nucleotide sequence read by the sequencer
- the symbol "+", (optionally followed by additional descriptions)
- a corresponding quality score for each base call represented by ASCII characters

An example from the Illumina platform looks as follows:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

The encoding of the quality score with ASCII characters in a FASTQ file depends on the sequencing platform, illustrated below [154].



```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0.........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
0.2.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

The example read above has a Phred+33 quality score (Illumina 1.8+): The stretch of As in the middle denote a good quality around 32 and is flanked by decreasing quality (symbols below 30).

Chapter 6

Software

In Chapter 5, the Mutual Interference Mapping Experiment (MIME) [8] has been presented. Briefly recapitulated, in MIME RNA is randomly mutated, selected by function, physically separated, and sequenced using NGS.
For the evaluation, the sequencing data has to be mapped to the reference sequence. With the resulting mapping files it is possible to extract the nucleotide frequencies for each position, which will be used to conduct the data analysis of the mathematical framework presented in the previous chapter.

This pipeline requires several computational steps, which are implemented in efficient and flexible software. The first tool described here uses mapped reads in SAM format to generate count data files containing position-wise nucleotide frequencies. These files are the input for the second software, MIMEAnTo [11], allowing to analyse the target RNA with regard to its function.

6.1 SAM to Counts

In order to evaluate the resulting MIME data after sequencing to infer functional important regions within the assayed RNA, they have to be preprocessed in a workflow, illustrated in Figure 6.1. The fragmented RNA sequences conducted in MIME experiments are in FASTQ format after sequencing. They need to be demultiplexed, trimmed, and mapped to the reference sequence to enable the inference of position-wise nucleotide frequencies. The particular steps are described below.

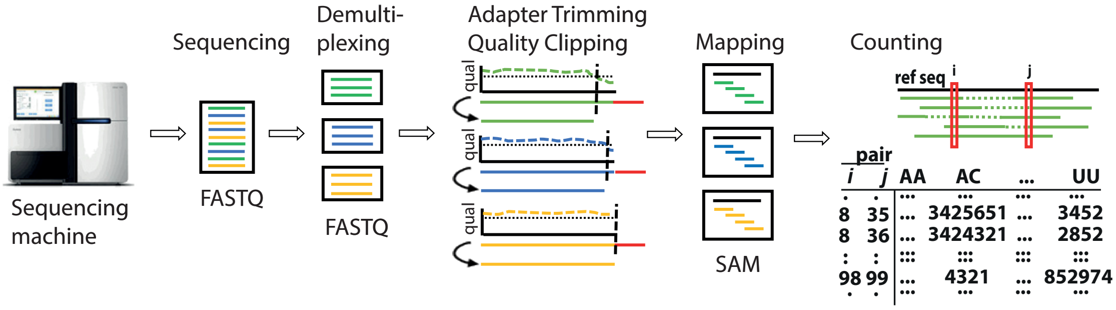


FIGURE 6.1: Pipeline of preprocessing the sequencing data. The sequencing output is in FASTQ format (two files for the read pairs) and contains all samples. The samples need to be separated according to their barcode. Sequencing adapters have to be removed, accompanied by quality clipping. After mapping to the reference sequence, the resulting SAM files serve as input for the counting routine. The resulting output are text files containing position (–pair, or –triplet) –wise nucleotide frequencies.

6.1.1 Preprocessing: From Sequencing Output to SAM Files

The output of the sequencing machine are genome fragments (reads) in FASTQ format, introduced in the NGS Section 5.2. In the MIME protocol, the different samples are tagged with barcodes during library preparation. The sequenced fragments for all samples are pooled in one large file and need to be demultiplexed according to their barcoding. An example demultiplexing tool is Novocraft's "NovoBarcode".

The sequencing procedure requires certain adapters, which were ligated to the genome fragments. They were sequenced along with the fragments and have to be removed. At the same time, quality clipping is necessary in order to get clean reliable data and to achieve a solid result. A tool for adapter trimming would for example be "Trim Galore!" (Bioinformatics, Babraham Institute), which also filters nucleotides according to a given quality threshold. For instance, a threshold with Phred+33 quality scoring of 30 would correspond to a base call accuracy of 99.9% (cf. Section 5.2.3).

The fragmented, reverse transcribed RNA is sequenced from both sides of the donor sequence, resulting in paired-end reads, which are placed in two separate FASTQ files. The information, which paired reads belong together, and the relative direction on the reference sequence are known. The trimmed read pairs are aligned to the reference sequence with a mapping tool which allows multiple mismatches, such as "NovoAlign" from Novocraft, as multiple mismatches are expected in the randomly mutated data. The resulting output files are in Sequence Alignment/Map Format (SAM).

6.1.2 Implementation

The original preprocessing program was based on python scripts created in connection with the MIME method [8]. However, the routines required several hours for big data sets, especially for processing covariational information. Hence, we considered it worthwhile to implement the software in a more efficient way using C++14, also adjusting and extending further filtering options. The complexity of the program is $\mathcal{O}(M \cdot \binom{r}{d} + \binom{L}{d})$, with M giving the number of paired-end reads in the SAM files, r the approximate read length, L the length of the reference sequence and d the dimension for the count routine (considering single sites = 1d, position pairs = 2d, position triplets = 3d). The first term qualifies the $\binom{r}{d}$ increments for M reads, with M usually running into millions. The second term describes the output routine, and is rather negligible.

With our implementation, we were able to achieve an immense speed up. To exemplify, with a number of around 2.5 million read pairs of an average length of 85 nucleotides each and a total length of 535 nucleotides of the reference sequence, we could reduce the runtime for the 1d case from 36 minutes with the python script to 7 seconds with the new implementation. For the 2d case, the time could be reduced from 32 hours to 39 seconds. The programs were launched on a MacBook Pro (2015) with 2.5 GHz Intel Core i7 and 16 GB RAM.

The program is called from the command line with

```
sam2counts <reffile> <samfile1> <samfile2> <outfile> <dimension> <qualiThreshold>
```

and requires the following parameters

reffile	(string)	path to reference sequence file in fasta format
samfile1	(string)	path to first SAM file of the mapped paired reads
samfile2	(string)	path to second SAM file of the mapped paired reads
outfile	(string)	path to output file containing the counts in tsv format
dimension	(int)	either 1, 2, or 3 for the counting dimension
qualiThreshold	(int)	threshold for the quality score

Despite quality clipping during the data preparation, the quality threshold was deemed necessary in this routine. The reason is, that it is usually assumed that the quality of a sequence drops towards the ends of the fragments. Hence, nucleotides with bad quality are trimmed until a good quality is reached. However, it may also happen that a drop of quality occurs within the sequence and these nucleotides of bad quality remain in the read. Therefore, we check for the quality of all nucleotides, since this requires only one additional operation per iteration.

The program can be divided into three procedures. The read pairs are imported and prefiltered, the cohesive sequences of a read pair are merged to one read with certain rules, and finally, the reads are counted for each position and nucleotide and written into the output file. The procedures will be described in particular in the following subsections.

Prefilter

After importing the reference sequence, the read pairs are imported from the two SAM files and processed one by one. Each alignment line of a SAM file contains 11 mandatory fields of essential mapping information, and a variable number of optional fields. An example of an alignment line is given below:

```
FCC1073ACXX:6:1101:17307:2200#AGTCAAAT/1  0   HIV1_535  22  70  94M6S  *  0  0
TCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTCAAAGTAGTGTGTGCCCGTCTTGTGGT
abbcccceggggggiiiiihhhhiiiihhiihhhgfghghfffiiiffhiiiifihicfhffhhiiiheggdgeeeceeadb_bddccbbbaaaBBBBBBB
PG:Z:novoalign  AS:i:64  UQ:i:64  NM:i:0  MD:Z:94
```

Only those fields, which are used for further filtering are explained here.
In a first filtering step, both sequences are checked for their mapping status. We considered multiple conditions to assess mapped reads:

- The second field in a line contains a bit flag. If the fourth bit is set, the fragment is unmapped.
- The fourth field gives the left most mapping position. To be mapped, the positions must be > 0.
- The so called CIGAR string in the sixth field contains important mapping information. If these are unavailable, indicated by a "$*$", we regard the read as unmapped.

Read pairs are only considered if both sequences are mapped. Side note: The example above complies with none of these criteria (indicated in red), hence this read is mapped.
For further processing, the CIGAR string, the mapping positions, the nucleotide sequence (10th field), and the qualities in ASCII representation (11th field) are extracted.
The CIGAR string includes different operations to describe the mapped read. Regions with alignment matches are denoted by "M", which does not necessarily mean that the nucleotides agree with the reference sequence. An alignment match includes sequence matches and mismatches. Deletions are denoted by the character "D", insertions by "I", and soft clipped regions, which can not be mapped, are indicated by the character "S". The preceding numbers in front of the respective characters denote the amount of consecutive nucleotides to which this operation applies. In the example above, the CIGAR string contains 94M6S, i.e. 94 positions are matching the alignment and six positions could not be mapped. For construction of the ultimate read, only matching regions are considered.

Merging

The matching sequence pair is merged to one single read for the count routine. However, these reads may contain regions or single positions which are ignored, because they do not fulfil the following requirements.
First of all, a symbol in the sequence other than the four bases A, C, G, and T is invalid. This is also the case for nucleotides where the quality score is smaller than the given threshold.
The pair of sequences can either be distinctly mapped to the reference such that there is a region in between, which is not covered by the reads, as illustrated in Figure 6.2, top.

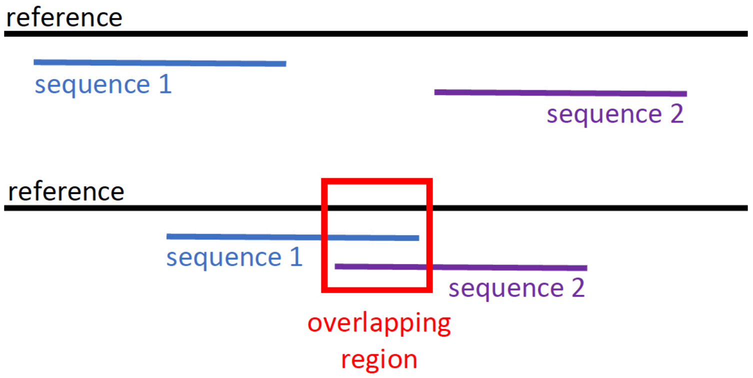


FIGURE 6.2: Mapping of paired reads to the reference sequence. The upper mapping shows the two sequences which are aligned distinctly such that a region in between is not covered. Below, the two fragments are aligned such that they cover partly the same region.

In this case, the nucleotides of the matching positions of both aligned sequences are simply added to the merged read. In case indels occur in between the matching regions, they are also treated as not covered regions.
However, the mapped sequences may also overlap, as seen in Figure 6.2, bottom. For the overlapping region, different cases are contemplated to decide for the incorporated nucleotide. If the two nucleotides are equal, which may be either the wild type of the reference sequence or a mutation, the respective nucleotide is added to the read. If they differ, but one of the sequences contains the wild type, the other one is considered as sequencing error and the wild type is chosen for the read. This applies also, if the one wild type nucleotide is given and the second symbol is invalid. If the two nucleotides disagree, but both differ from the reference (including being invalid) the position is neglected. An overview of the different cases is given in Figure 6.3.

Counting

The constructed new read is evaluated according to the given dimension parameter.
In the case of single site variation (dimension = 1), the respective nucleotide count is incremented by one for each incorporated position. After iterating through all reads, the nucleotide occurrences (#) of all positions are written into the given output file, seperated by tabulator:

```
pos #A #C #G #T
```

For the 2d case, co-occurring nucleotides for each position pair are counted. The resulting output contains the frequencies in the following format:

reference	T	T	T	T	T	T	T	T
sequence 1	T	T	T	T	A	A	A	A
sequence 2	T	C	_	N	A	C	_	N
merged read	T	T	T	T	A	×	A	×

FIGURE 6.3: Different cases to handle overlapping reads. If at least one of the two mapped sequences carries the wild type base at the aligned position, it is incorporated into the merged read for this position. If one of the sequences carries a different base than the reference, it is only considered as true mutation in the merged read if both sequences carry the same base, or if an insertion or deletion (represented by "_") is present in the other sequence. If both sequences carry different bases or one of them is invalid (indicated by "N"), the position is ignored in the merged read.

```
pos1 pos2 #AA #AC #AG #AT #CA #CC ... #GG #GT #TA #TC #TG #TT
```

As example, the column "#AG" contains the number of reads which span the two respective positions and comprise the nucleotide "A" at the first position and the nucleotide "G" at the second position.
Similarly, the reads for a dimension of 3 are counted for each triplet of positions and nucleotide co-occurrences and written into the output file:

```
pos1 pos2 pos3 #AAA #AAC #AAG #AAT #ACA ... #TGT #TTA #TTC #TTG #TTT
```

6.2 MIMEAnTo

The **MIME An**alysis **To**ol (MIMEAnTo [11]) processes MIME generated data, allowing to analyse RNA with regard to its function. The mathematical concepts of the workflow are broadly explained in Chapter 5.1.
The input consists of base counts at each position, obtained from the next-generation sequencing reads after mapping to the reference sequence and counting, as described in the previous section. Each input data set, corresponding to one experimental set up, consists of the counts for the selected and unselected samples of the mutant library, and counts for the selected and unselected samples of the control library conducted without mutations.
The counts are translated into "raw" quantitative effects of mutations on a certain function relative to the wild type configuration. These effects are computed for all three possible mutations at each sequence position. Since the NGS reads from the MIME experiment are confounded by errors, introduced during library preparation and sequencing, error correction and statistical assessment are essential.

MIMEAnTo has a wizard-like graphical user interface, guiding through the data analysis procedure in three steps:

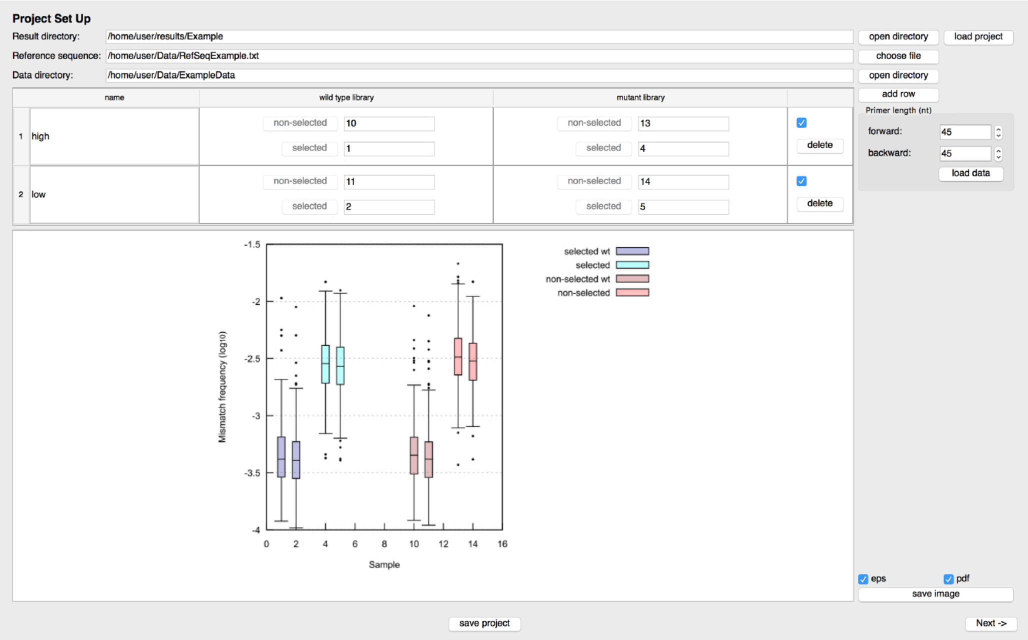
- data input and assessment
- correction of errors due to reverse transcription and sequencing
- quantification of raw effects and quality filtering

The results can be plotted as publication ready graphics.

6.2.1 Implementation

MIMEAnTo is a cross-platform software, implemented in C++ using the boost library, and the Qt framework for the graphical user interface. The plots are generated with gnuplot. The program is available as standalone binary executable for different operating systems or can be compiled using the source files (download from `https://github.com/maureensmith/MIMEAnTo`).
All interim settings will be saved in a text file (project.txt) in the given result directory. This file will be created after initialising the project in the GUI and updated during the workflow. It can be used to reload the project.

Project Initialisation



Project Set Up

Result directory: /home/user/results/Example

Reference sequence: /home/user/Data/RefSeqExample.txt

Data directory: /home/user/Data/ExampleData

	name	wild type library		mutant library		
1	high	non-selected	10	non-selected	13	☑
		selected	1	selected	4	delete
2	low	non-selected	11	non-selected	14	☑
		selected	2	selected	5	delete

Primer length (nt)

forward: 45

backward: 45

☑ eps ☑ pdf

FIGURE 6.4: Step I of the analysis pipeline: Initialising/loading the project and check of the mutation rates. Screenshot of MIMEAnTo.

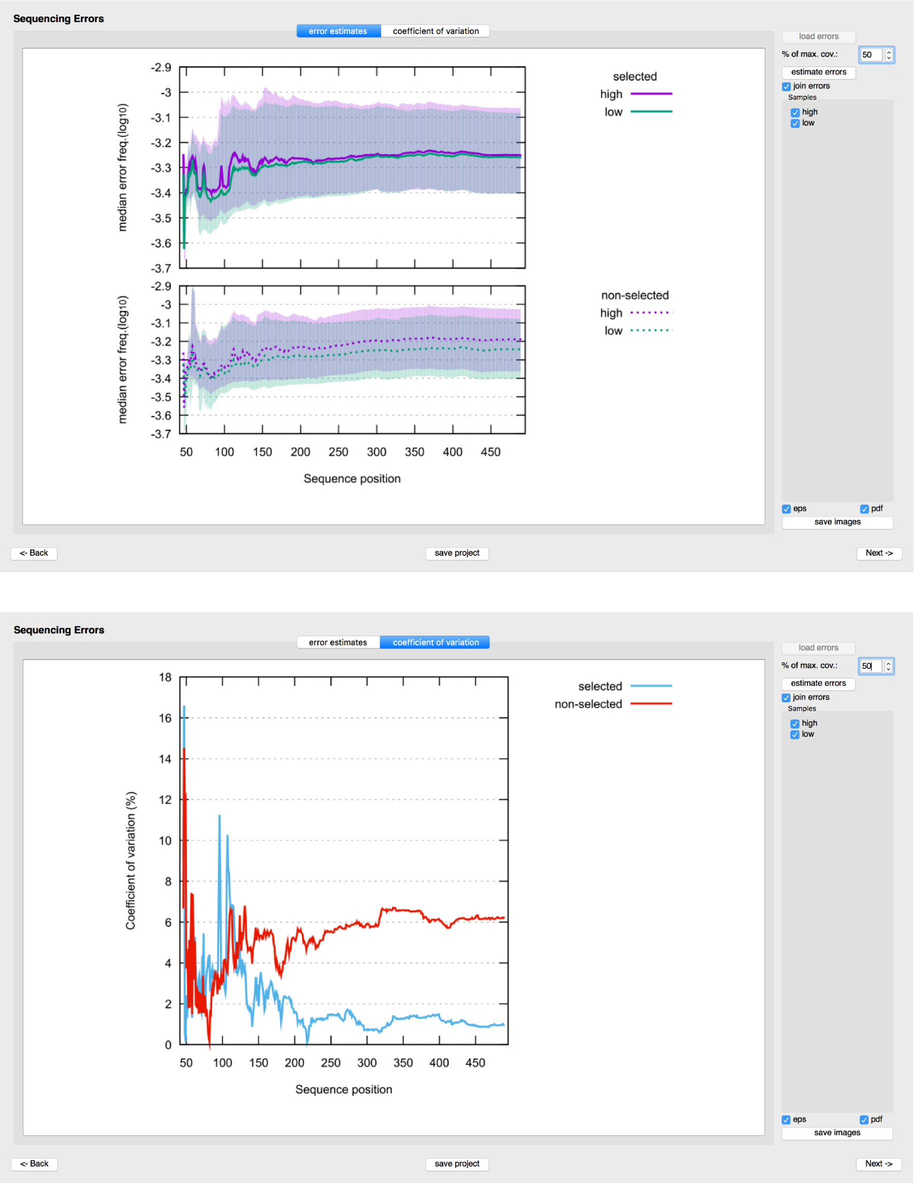
In the first step of the analysis-pipeline (corresponding UI in Figure 6.4), the project is initialised with the following mandatory parameters in the above-placed rows:

- a **result directory**, where all intermediate and final results are saved
- a **reference sequence**, either in fasta format, or a text file containing "position,base-number", where the nucleotides are provided by numbers following the convention: A = 1, C = 2, G = 3, U = 4
- the **data directory** where the reference sequence and the subdirectories ("1d" and "2d") with the respective count files are stored

Furthermore, the **sample data sets** have to be added to the data table. Each sample set has to be named, e.g. indicating the different experiment conditions like a particular protein concentration. As mentioned above, the sample set is composed of the count files for selected and unselected samples from a wild type– and mutant library, respectively. The associated barcodes have to be entered to the respective fields. In addition, the **forward and backward primers** can be clipped from the examined sequence.

The creation (or loading) of the project invokes a first plausibility check. A boxplot showing mismatch frequencies summarised over all positions for each sample library is created using the provided 1d data. The plot allows the user to assess whether the data looks reasonable at first sight. For example, mutant libraries should have a higher mismatch frequency than their corresponding wild type counterpart.

Error Estimation



% of max. cov.: 50

☑ join errors

Samples

☑ high

☑ low

☑ eps ☑ pdf

FIGURE 6.5: Step II of the analysis pipeline: Error estimation (top) and assessment of the differences between errors of selected and non-selected samples (bottom). Screenshots of MIMEAnTo.

The evaluated data contains noise, which has been introduced during library preparation and sequencing. In order to approximate the true mutation frequency in the mutant libraries more precisely, the errors have to be corrected according to Equation (5.14) derived in the mathematical framework of MIME in Chapter 5.1:

$$\mathbb{E}(\kappa_{w\to m}(i)) \approx \frac{1}{N}\sum_{j\neq i}\frac{R_{m,w}(i,j)}{R_{w,w}(i,j)}.$$

The error estimate for each position and mutation is resampled with the covariational (2d) data of the non-mutated control samples for each of the selected and unselected sample pool. The single error estimates can be quality filtered using the parameter **percentage of maximum coverage**. This threshold reassures a considerable depth for the resampling: Position pairs are only regarded for the error calculations if they reach a minimum read coverage.
The error plot on the first tab in MIMEAnTo, seen in Figure 6.5 on the top, shows the mean mutation rates (+IQR) per position for each sample of the project. The second tab, in Figure 6.5 on the bottom, comprises the coefficient of variation between all samples belonging to either the selected or non-selected pool, i.e. giving a measure of how similar the errors among the samples are. On the basis of the coefficient of variation, the user can decide to compute a single error estimate for all samples in one pool for further evaluations, or to consider the errors independently. Joining the errors of a pool with a low coefficient may achieve a better statistical foundation for the error estimate.

In addition to error correction, the estimated error rates will be used in subsequent analysis to asses a signal-to-noise ratio, i.e. the approximated true mutation signal vs the expected noise.

Kd Estimation and Quality Filtering

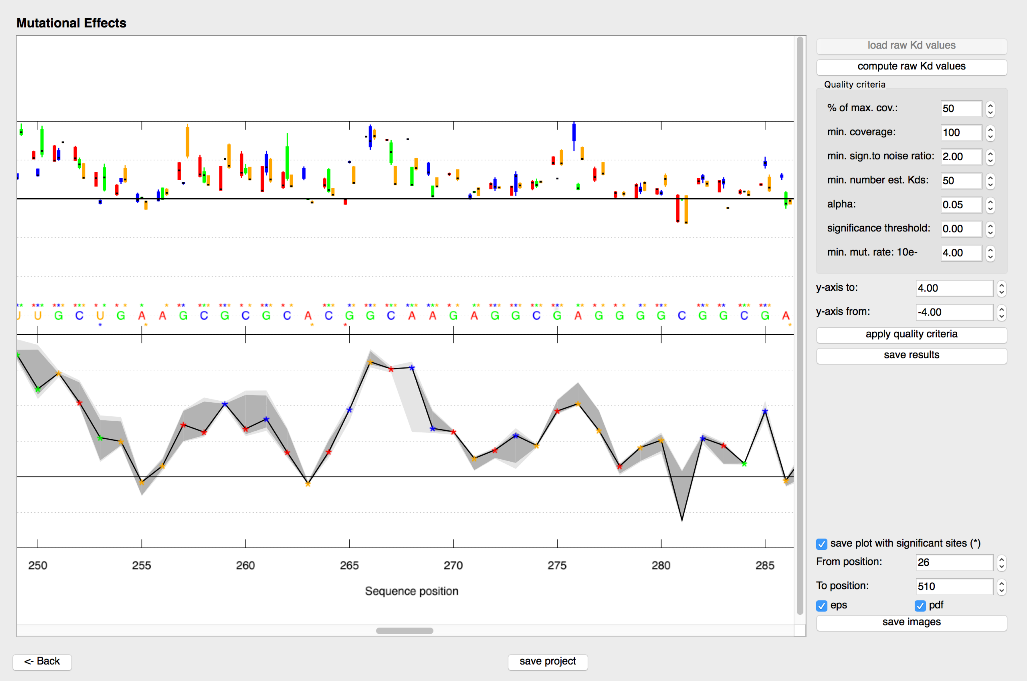
In the third step, relative effects of mutations can be calculated (or imported, if already computed). The raw Kd values for each position pair (i, j) and mutation m at position i are calculated and error corrected according to Equation (5.13) in Chapter 5.1:

$$\frac{Kd_{m,w}(i,j)}{Kd_{w,w}(i,j)} \approx \frac{\frac{R^u_{m,w}(i,j)}{R^u_{w,w}(i,j)} - \kappa^u_{w\to m,w}(i,j)}{\frac{R^b_{m,w}(i,j)}{R^b_{w,w}(i,j)} - \kappa^b_{w\to m,w}(i,j)}. \tag{6.1}$$

The effect of mutation m at sequence position i is resampled by considering the covariational data. The number of reads spanning positions i and $j \neq i$, comprising mutation m at position i and the wild type nucleotide at position j are denoted by $R_{m,w}(i,j)$. $R_{w,w}(i,j)$ gives the number of reads with the wild type residue for both positions. The error correction term is given by the expected error probability of falsely detecting a wild type at position i as a mutant m ($\mathbb{E}(\kappa_{w\to m}(i))$, calculated in the previous step.

Several features describing the data are calculated along with the effects and can be used to filter the sample set to improve the quality of the results. Moreover, a p-value for each estimate is computed applying the significance test described Eq. (5.16) and (5.17) in Chapter 5.1.4. These quality criteria can be set by the user and will be briefly explained below:

% of maximal coverage
Determines the depth of resampling, i.e. only position pairs exceeding a minimum coverage are regarded for Kd estimation, similar to the error estimation before. The minimum coverage is assessed with respect to the position with the maximal read coverage. This quality criteria will be important for filtering when the coverage is in general very high.



Quality criteria

% of max. cov.: 50

min. coverage: 100

min. sign.to noise ratio: 2.00

min. number est. Kds: 50

alpha: 0.05

significance threshold: 0.00

min. mut. rate: 10e- 4.00

y-axis to: 4.00

y-axis from: -4.00

☑ save plot with significant sites (*)

From position: 26

To position: 510

☑ eps ☑ pdf

FIGURE 6.6: Step III of the analysis pipeline: Estimation of the raw effects (Kds) and quality filtering. The upper plot shows the position-wise resampled Kd estimates ($\log_2$) for each of the three possible mutations in boxplot format. The different mutations are colour-coded (A C G U). The lower plot contains the median of the maximum effect ($\log_2$) for each position. Significant effects are denoted with $*$ in the colour of the respective mutation. Screenshot of MIMEAnTo.

minimum coverage
Determines the depth of resampling, i.e. only position pairs exceeding a minimum coverage are regarded for Kd estimation. Here, the minimum coverage is determined in absolute numbers. This criterion will be important for filtering when the coverage is rather moderate.

minimum signal-to-noise ratio
A signal-to-noise ratio $D_{m,w}(i,j)$ is computed for each pair of positions (i, j) and for each mutation m according to Equation (5.15) in Chapter 5.1.4. It is recommended to use a signal-to-noise ratio of at least 2 in MIMEAnTo (the true signal is at least twice the expected error). Otherwise, the noise might be falsely considered as signal. A threshold > 2 should be chosen with caution to find a balance between data quality and quantity.

minimal number of estimable Kds
For the statistical ascertainment, it is important to assure a sufficient sample size: p-values will only be computed for mutation m at position i, if the number of resamplings exceeds the threshold provided here.

alpha
Level of significance. Kd estimates are marked as significant if the p-value is $< \alpha$.

significance threshold
The significance test ascertains position-wise mutations, whether they have an significant effect. The threshold can be adjusted to determine mutations that have a significant, and if desired, strong effect, i.e. if a mutation at the regarded position affects the function more than c-fold.

minimum mutation rate
Filtering resamples with low mutation rate, in case of heterogeneous mutation rates within the samples. The parameter gives the magnitude for the mutation rate, e.g. minMutRate = 4, means that the min. mutation rate is 10^{-4}.

The Kd estimates are plotted after the quality filtering, as seen in Figure 6.6:

1. **All mutational effects**
   The upper plot shows the relative effects, $\log_2\left(\frac{Kd_m}{Kd_w}(i)\right)$, of all three mutations per sequence position i as candlesticks (median + IQR) in different colours for the respective mutations (A C G U). The $*$ above or below the given wild type letter of the position denotes if the effect significantly increases or decreases the Kd.

2. **Maximal mutational effects**
   The lower plot highlights the mutation with the strongest effect (positive or negative) for each position, $\log_2\left(\frac{Kd_{m_{max}}}{Kd_w}(i)\right)$. Again, the $*$ indicates if the effect is significant and the colour-coding identifies the respective mutation.

The plots can also be saved in two distinct files, either for the whole sequence or an user defined region.

Output

Plots All plots of the above described steps can be saved in the subdirectory `path_to_results/plots`. The user can choose to save the plots as eps- and/or pdf-file with an (optional) suffix for the filename.

Tables Tabular result files of the Kd computation are exported as csv-files in the user defined result directory. Optionally, the user can give a suffix for the files, to generate several results with different parameter settings.
Information about all mutational effects are written into the tab seperated file "PositionWiseKdEstimates" containing the following data:

- position
- wild type base w (A = 1, C = 2, G = 3, U = 4)
- max. effect base m_{max} (A = 1, C = 2, G = 3, U = 4)
- for each base (mutation) m = A, C, G, U
  - median $\frac{Kd_m}{Kd_w}$
  - p-value for $\frac{Kd_m}{Kd_w}$
  - #resamplings for $\frac{Kd_m}{Kd_w}$
  - #lower estimates for $\frac{Kd_m}{Kd_w}$
  - #upper estimates for $\frac{Kd_m}{Kd_w}$
  - 5th percentile of $\frac{Kd_m}{Kd_w}$ estimate
  - 95th percentile of $\frac{Kd_m}{Kd_w}$ estimate

The tab separated file "PositionWiseMaxKd" contains the information about the mutation with the maximum effect for each position:

- position
- wild type base w (A = 1, C = 2, G = 3, U = 4)
- max. effect base m_{max} (A = 1, C = 2, G = 3, U = 4)
- median $\frac{Kd_{m_{max}}}{Kd_w}$
- p-value for $\frac{Kd_{m_{max}}}{Kd_w}$
- #resamplings for $\frac{Kd_{m_{max}}}{Kd_w}$
- #lower estimates for $\frac{Kd_{m_{max}}}{Kd_w}$
- #upper estimates for $\frac{Kd_{m_{max}}}{Kd_w}$
- 5th percentile of $\frac{Kd_{m_{max}}}{Kd_w}$ estimate
- 95th percentile of $\frac{Kd_{m_{max}}}{Kd_w}$ estimate

Chapter 7

MIME applied to *in cell* Experiments with the HIV-1

In Chapter 2, we examined the replication cycle of HIV-1 in detail and how the genomic RNA of the virus is involved in each of the steps. We have seen, that the viral gRNA encodes nine (poly–)proteins. In addition to its coding capacity, the HIV-1 gRNA contains numerous regulatory elements. These elements interact in complex ways to control key steps of the HIV-1 life cycle including transcription, translation, export, packaging and reverse transcription. Particularly, the 5′ UTR and the beginning of the Gag coding sequence comprise a high concentration of these regulatory sequences, as presented in detail in Section 2.3.1. This highly structured region contains a series of functional domains, as a reminder shown in Figure 7.1: the trans-activation region (TAR) element for transcription, PolyA for polyadenylation, the primer binding site (PBS) for reverse transcription, SL1 containing the dimer initiation site (DIS) for gRNA dimerisation, SL2 with the splice donor (SD), and SL3 considered hitherto as the major packaging domain.

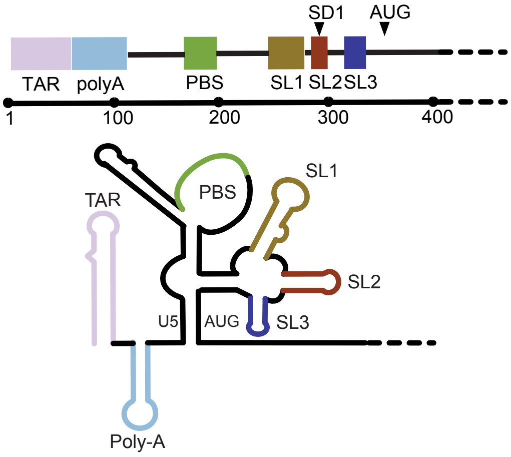


FIGURE 7.1: The HIV-1 5′ UTR folds into a series of structural domains regulating the viral life cycle: TAR, PolyA, PBS and SL1-3. Figure extracted from [12].

The regulatory elements are essential for the viral replication. Hence, they give promising targets for new antiviral drug design [155]. First achievements in this direction have been attained for the case of Hepatitis C virus (HCV) patients, which could be successfully treated by functionally inhibiting regulatory ncRNA [6]. However, the current knowledge about regulatory mechanisms in HIV is limited and needs to be further explored. For the analysis of functionally important RNA, high-resolution and quantitative methods are required.

In Chapter 5.1, we explained the Mutational Interference Mapping Experiment (MIME) in detail. MIME is a powerful tool for the detection of functional constraints on molecular level in a one-step evolutionary process, revealing functionally and regulatory important regions. A first application of MIME was done by the authors in an *in vitro* experiment, mapping the binding site of the human HIV-1 Pr55$^{\text{Gag}}$ protein on the viral gRNA and modelling RNA structure motifs that are crucial for protein binding [8].

In this chapter we present our results for the adaption of MIME for *in cell* experiments [12], defining RNA elements within the 5′ UTR of the HIV-1 gRNA that are important for viral replication in cells. After transferring the biological context of the application from *in vitro* to *in cellulo*, we briefly explain the experimental procedure. The resulting data output is evaluated according to the mathematical background explained in the subsequent section. Finally, we will present the detected regulatory motifs and discuss the results of the analysis.

7.1 From *in vitro* to *in cell* Experiments

The recently deveopled method MIME (cf. Chapter 5.1) has been used for an *in vitro* experiment to determine functional regions within genomic RNA in single site resolution [8]. More precisely, the authors could determine the Pr55$^{\text{Gag}}$ protein binding sites on the viral gRNA of HIV-1 in SL1 and SL3. The complex of Pr55$^{\text{Gag}}$ and the viral gRNA is recognised as a major determinant of the packaging process, where the dimerised viral RNA is packed into new emerging virions. In fact, it has been shown, that with the absence of gRNA, Pr55$^{\text{Gag}}$ binds to any nucleoacid molecule (even without the protein binding domain), suggesting that RNA plays a structural role during virus assembly [156, 157]. Still, the HIV-1 gRNA is selectively packaged into nascent virions, which arouses the curiosity on how the virus can differentiate between RNAs, putting a great importance on the comprehension of signals controlling the packaging process.
We applied MIME to *in cell* experiments to answer the question, whether the motifs binding to Pr55$^{\text{Gag}}$ found *in vitro* also pertain to the native cellular environment. Moreover, we were interested whether additional signals, e.g. binding sites for other proteins or nucleic acids, are relevant for effective packaging.
In addition, the experimental setup allowed to explore the regulatory impact of the 5′ UTR on intracellular gRNA production.

The experimental design will be described in the following, depicted in Figure 7.2: The proviral genome is randomly mutated using error prone PCR, and subsequently cloned into a genomic RNA expression vector. The vector was modified to serve only as expression vector for the mutation libraries by containing:

- restriction sites for the cloning of the mutant libraries
- a substitution in the start codon of gag to prevent Gag expression
- a stop codon to prevent Tat protein expression
- a deletion in env (for biosafety)

The structural and enzymatic precursor proteins, Gag and Gag-Pol, and accessory proteins Tat and Rev, were expressed from a separate packaging vector. The vectors were co-transfected into human 293T cells, leading subsequently to the transcription and packaging of mutant RNAs. The segregated functionally selected and unselected RNA populations in cells and virions, respectively, were extracted and reverse transcribed. After amplification of the viral cDNA and the input DNA plasmids, the material was fragmented, barcoded, and eventually sequenced using Illumina HiSeq2500. The resulting NGS data was analysed using the MIMEAnTo software [11], introduced in Chapter 6.
By varying the functional selection criteria, we obtained two distinct maps of regulatory RNA controlling intracellular gRNA production and gRNA packaging, respectively: For the gRNA production, comprising the steps from the transfected genomic material to the transcribed gRNA in the cells, we found three regulatory motifs. For the packaging of

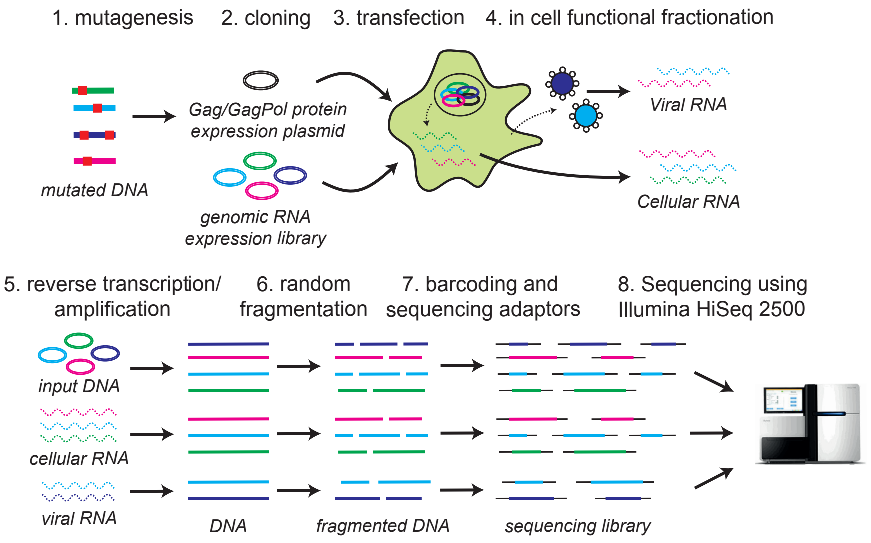

1. mutagenesis
2. cloning
3. transfection
4. in cell functional fractionation
mutated DNA
Gag/GagPol protein expression plasmid
genomic RNA expression library
Viral RNA
Cellular RNA
5. reverse transcription/ amplification
6. random fragmentation
7. barcoding and sequencing adaptors
8. Sequencing using Illumina HiSeq 2500
input DNA
cellular RNA
viral RNA
DNA
fragmented DNA
sequencing library

FIGURE 7.2: *In cell* Mutational Interference Mapping Experiment (*in cell* MIME). The proviral genome was randomly mutated using error prone PCR, and subsequently cloned into a gRNA expression vector. Gag and Gag-Pol were expressed from a separate expression plasmid. The mutant library and Gag/Gag-Pol expression plasmid were transfected into 293T cells. Transcripts of mutant RNAs were functionally sorted by the viral and cellular machinery. Viral RNA present in cells and virus was reverse transcribed. Viral cDNA and the input DNA plasmid was amplified, fragmented, barcoded, sequenced, and analysed using the MIMEAnTo software. Figure extracted from [12].

gRNA, including the steps from the gRNA in cytoplasm to the successfully assembly of virions with packed gRNA, two regulating signals could be detected. Strikingly, a hexamer sequence within 5′ PolyA regulated both gRNA production and packaging, revealing the cellular polyadenylation machinery as a dual regulator of HIV-1 replication.
In the next section, we are going to elaborate the mathematical evaluation leading to these results.

7.2 Analysis of *in cell* MIME Data

Akin to the method explained in Chapter 5, the goal is to infer quantitative effects of each mutation m at each nucleotide position i on gRNA production and packaging, obtained from the generated NGS output.
Mutations with a strong impact on the respective function indicate which regions are involved in the particular process. A schematic depiction of viral gRNA production and packaging in the cellular environment is shown in Fig. 7.3. For simplicity, we focus only on two viral strains here, termed wild type w and mutant m. The analysis easily extends to arbitrarily many viral strains (as in the experiment).
In this example, mutant m and wild type w differ only at a single position i. The mutation has a biophysical impact on at least one of the assayed functions. The following system of ordinary differential equations describes the mass-action kinetics of intracellular and

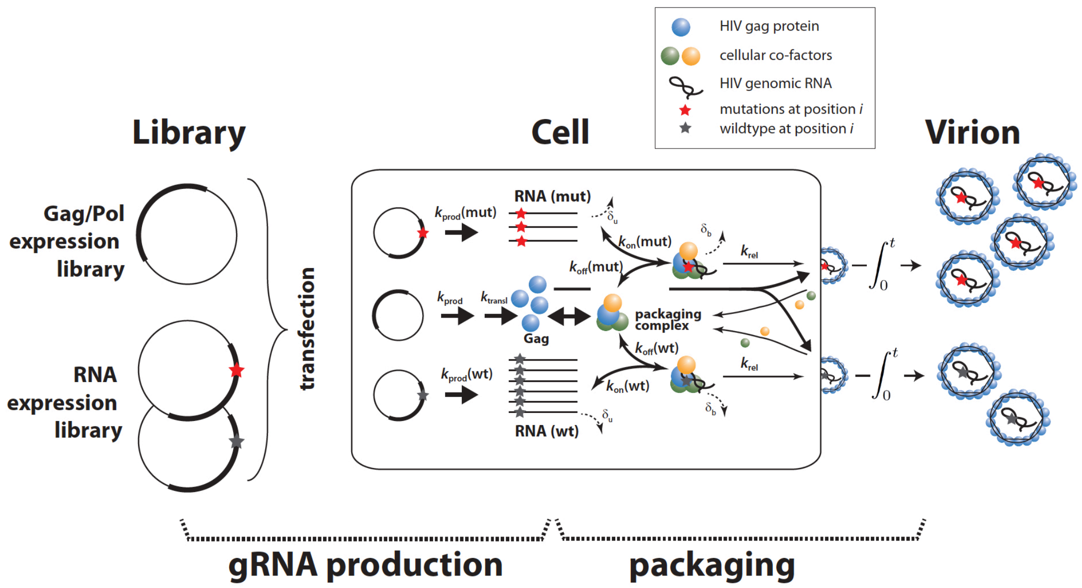


FIGURE 7.3: Dynamical system of RNA populations in cells and virions. Gag expression plasmids are co-transfected with library expression plasmids. The gag expression plasmids produce transcripts that are subsequently translated into Gag polyproteins that may form a packaging complex with viral RNA, or be directly targeted to the cell surface to form nascent virions. Library expression plasmids produce transcripts with rate $k^{\text{prod}}_{w/m}$. This unbound viral RNA may either bind to a packaging complex with rate $k^{\text{on}}_{w/m}$ or be degraded intracellularly. RNA that is bound to the packaging complex may either dissociate with rate $k^{\text{off}}_{w/m}$ or be packed into nascent virions and released with rate k^{rel}. Figure taken from [12], Supplementary Figures.

viral RNA from Fig. 7.3:

$$\frac{\mathrm{d}}{\mathrm{d}t} S^u_w = S^{\text{pl}}_w \cdot k^{\text{prod}}_w - S^u_w \left(P(t) \cdot k^{\text{on}}_w + \delta^u\right) + S^b_w \cdot k^{\text{off}}_w \tag{7.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S^b_w = S^u_w \cdot P(t) \cdot k^{\text{on}}_w - S^b_w \left(k^{\text{off}}_w + k^{\text{rel}} + \delta^b\right) \tag{7.2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S^{\text{virus}}_w = S^b_w \cdot k^{\text{rel}} \tag{7.3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S^u_m = S^{\text{pl}}_m \cdot k^{\text{prod}}_m - S^u_m \left(P(t) \cdot k^{\text{on}}_m + \delta^u\right) + S^b_m \cdot k^{\text{off}}_m \tag{7.4}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S^b_m = S^u_m \cdot P(t) \cdot k^{\text{on}}_m - S^b_w \left(k^{\text{off}}_m + k^{\text{rel}} + \delta^b\right) \tag{7.5}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S^{\text{virus}}_m = S^b_m \cdot k^{\text{rel}}. \tag{7.6}$$

The dynamical system can be explained as follows: The subscripts for the rates and concentrations describe for which species they apply, with w for the wild type and m for the mutated RNA. The number of transfected library expression plasmids are denoted by S^{pl}. They produce transcribed gRNA with rate k^{prod}, which we denote here as the free, unbound RNA S^u. With an association rate of k^{on}, the free gRNA forms a bi-molecular complex with the available packaging protein $P(t)$, and is degraded with rate δ^u. The concentration of the bound RNA is labelled with S^b and releases the complex with rate

k^{off}, or degrades with rate δ^b. The RNA which is bound to the complex gets shuttled to the cellular surface, packed into nascent virions, and released to the extracellular space with rate k^{rel}, resulting in the concentration of RNA in virions S^{virus}.

7.2.1 Relation between Nucleotide Frequencies and the Effect on intracellular gRNA Production

In order to compute the impact of mutations on gRNA production, we deduce the mutation frequencies of the viral RNA present in the cell. A fraction of the produced gRNA is packed and released into viruses, which could in theory distort the true observed amount of produced RNA. For example, a mutation that increases packaging and lowers the pool of cellular RNA, could be misinterpreted as a mutant that decreases the gRNA production. However, the following observation motivated us to assume that the number of viral RNA in cells is not significantly altered by the packaging process, and thus not biasing the prediction of production efficacy: The packaging process is mediated by the Gag-protein (and possibly other cellular factors). However, the number of Gag molecules in virions vastly exceeds the number of RNAs. A dimer of gRNA is packed into a virion together with about 5000 Gag-proteins [158]. Two contradicting scenarios could explain this observed stoichiometry:

1. scenario: The number of RNA molecules is much smaller than the number of available proteins.
2. scenario: The majority of intracellular viral RNAs is not bound to the complex or packaged into virions.

The first scenario implies a very low number close to depletion of viral RNA in the cell. Hence, a very low competition between the different RNA strains would occur. With an excess of proteins the majority of RNAs would be packed into virions, especially due to the mentioned fact that Gag can bind to any RNA for the assembly of new virions. In this case, hardly any RNA could be extracted from the cells, or would yield very low counts. However, we clearly observe plausible signals for RNA in cells. Moreover, the results for packaging in the *in cell* analysis closely resemble the regulatory region found for the Gag binding experiments *in vitro* under competitive conditions. Both facts strongly argue for the second scenario, i.e. there is an excess of free over bound RNA inside the cell, i.e. $S^u >> S^b$ and thus $S^{\text{cell}} \approx S^u$, at least in our assay.

We thus can deduce the number of total cellular RNA (unbound and bound to the packaging complex) S^{cell}, for the wild type w and mutant m respectively:

$$\frac{\mathrm{d}}{\mathrm{d}t} S_w^{\text{cell}} \approx \frac{\mathrm{d}}{\mathrm{d}t} S_w^u = S_w^{\text{pl}} \cdot k_w^{\text{prod}} - S_w^u \cdot \delta^u \tag{7.7}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S_m^{\text{cell}} \approx \frac{\mathrm{d}}{\mathrm{d}t} S_m^u = S_m^{\text{pl}} \cdot k_m^{\text{prod}} - S_m^u \cdot \delta^u. \tag{7.8}$$

$S_{w/m}^{\text{pl}}$ denotes the number of plasmids that carry the wild type or mutant RNA. We assume that the plasmids are stable, hence constant (this assumption is not necessary to derive the result below, but it simplifies the derivations). Eq. (7.7) and (7.8) can be solved analytically and we derive

$$S_w^{\text{cell}}(t) \approx \left(1 - e^{-\delta^u t}\right) \frac{k_w^{\text{prod}} \cdot S_w^{\text{pl}}}{\delta^u} \tag{7.9}$$

$$S_m^{\text{cell}}(t) \approx \left(1 - e^{-\delta^u t}\right) \frac{k_m^{\text{prod}} \cdot S_m^{\text{pl}}}{\delta^u}. \tag{7.10}$$

Computing the fraction of mutants in the DNA library divided by the fraction of mutants in the cell gives

$$\frac{S_m^{\mathrm{DNA}}}{S_w^{\mathrm{DNA}}} \cdot \frac{S_w^{\mathrm{cell}}(t)}{S_m^{\mathrm{cell}}(t)} \approx \frac{S_m^{\mathrm{DNA}}}{S_w^{\mathrm{DNA}}} \cdot \frac{\left(1-e^{-\delta^u t}\right) \cdot k_w^{\mathrm{prod}} \cdot S_w^{\mathrm{pl}}}{\left(1-e^{-\delta^u t}\right) \cdot k_m^{\mathrm{prod}} \cdot S_m^{\mathrm{pl}}} \cdot \frac{\delta^u}{\delta^u} = \frac{S_m^{\mathrm{DNA}}}{S_w^{\mathrm{DNA}}} \cdot \frac{k_w^{\mathrm{prod}}}{k_m^{\mathrm{prod}}} \cdot \frac{S_w^{\mathrm{pl}}}{S_m^{\mathrm{pl}}}. \tag{7.11}$$

We assume that the transfection efficacy is identical between the wild type and the mutant plasmid, and thus, the mutant frequency is identical between the plasmids and the DNA libraries, i.e. $\frac{S_m^{\mathrm{pl}}}{S_w^{\mathrm{pl}}} = \frac{S_m^{\mathrm{DNA}}}{S_w^{\mathrm{DNA}}}$. Therefore, we have

$$K_m^{\mathrm{prod}} = \frac{k_w^{\mathrm{prod}}}{k_m^{\mathrm{prod}}} \approx \frac{S_m^{\mathrm{DNA}}}{S_w^{\mathrm{DNA}}} \cdot \frac{S_w^{\mathrm{cell}}(t)}{S_m^{\mathrm{cell}}(t)} \tag{7.12}$$

for arbitrary time t.

In the equation above, the term K_m^{prod} denotes the decrease of the gRNA production caused by mutation m relative to the gRNA production efficacy in the wild type w and will be used to assess the impact of mutations on the gRNA production efficacy. The term will be increased ($K_m^{\mathrm{prod}} > 1$), whenever a mutation decreases gRNA production.

Depicting only the maximal impact of all three possible mutations $m_{\max}(i)$ for each site i helps to identify regions which are important for gRNA production (as seen in the result Figure 7.6).

The mutation with maximal impact is given by $m_{\max}(i) = arg\max_m |\log_2\left(K_m^{\mathrm{prod}}(i)\right)|$, where mutation m is only considered if it has a significant impact on the production at nucleotide position i. If there is no mutation with a significant impact, all possible mutations at that position are evaluated.

7.2.2 Relation between Nucleotide Frequencies and the Effect on Packaging

If binding to the packaging complex is fast with respect to the other processes, gRNA production, and release, we can assume a quasi-equilibrium in Equations (7.2) and (7.5). Thus, we can set $\frac{\mathrm{d}}{\mathrm{d}t}S_m^b = \frac{\mathrm{d}}{\mathrm{d}t}S_w^b = 0$ and get after rearranging

$$S_w^u(t) = \frac{S_w^b \cdot (k_w^{\mathrm{off}} + \delta^b + k^{\mathrm{rel}})}{k_w^{\mathrm{on}} \cdot P(t)} \tag{7.13}$$

$$S_m^u(t) = \frac{S_m^b \cdot (k_m^{\mathrm{off}} + \delta^b + k^{\mathrm{rel}})}{k_m^{\mathrm{on}} \cdot P(t)}. \tag{7.14}$$

We retain the two assumptions from above: The number of RNAs in the cell S^{cell} is given by the sum of bound and unbound RNAs $S^u + S^b$, and the vast majority of RNAs in the cell is unbound. Hence, $S^u >> S^b$ and thereby $S^{\mathrm{cell}} \approx S^u$. Thus, we get

$$S_w^{\mathrm{cell}}(t) = \frac{S_w^b \cdot (k_w^{\mathrm{off}} + \delta^b + k^{\mathrm{rel}})}{k_w^{\mathrm{on}} \cdot P(t)} \tag{7.15}$$

$$S_m^{\mathrm{cell}}(t) = \frac{S_m^b \cdot (k_m^{\mathrm{off}} + \delta^b + k^{\mathrm{rel}})}{k_m^{\mathrm{on}} \cdot P(t)}. \tag{7.16}$$

In addition, we assume that the quasi-equilibrium is reached quickly, allowing us to approximate the bound RNA by a constant, i.e. $S^b(t) \approx$ constant. It therefore matters to the inference of the relative effect on packaging that sufficient time has passed for this approximation to be accurate. Under these assumption Eqs. (7.3) and (7.6) can be solved analytically, i.e. the RNA accumulating in virions is given by

$$S_w^{\text{virus}}(t) \approx k^{\text{rel}} \cdot t \cdot S_w^b \tag{7.17}$$

$$S_m^{\text{virus}}(t) \approx k^{\text{rel}} \cdot t \cdot S_m^b. \tag{7.18}$$

Dividing the fraction of mutant RNA in the cell by the fraction of mutant RNA that accumulates in virions, we derive

$$\frac{S_m^{\text{cell}}(t)}{S_w^{\text{cell}}(t)} \cdot \frac{S_w^{\text{virus}}(t)}{S_m^{\text{virus}}(t)} \approx \frac{k_w^{\text{on}}}{k_w^{\text{off}} + k^{\text{rel}} + \delta^b} \cdot \frac{k_m^{\text{off}} + k^{\text{rel}} + \delta^b}{k_m^{\text{on}}}, \tag{7.19}$$

meaning the ratio of mutation frequencies in the cells vs virions is constant. Additionally, if the degradation of the RNA packaging complex δ_b is slow, $k_{\text{off}} \cdot k_{\text{rel}} \cdot \delta_b \approx k^{\text{off}} \cdot k^{\text{rel}}$, we have

$$K_m^{\text{pack}} = \frac{k_{\text{on},w}}{k_{\text{off},w} + k_{\text{rel}}} \cdot \frac{k_{\text{off},m} + k_{\text{rel}}}{k_{\text{on},m}} \approx \frac{S_m^{\text{cell}}(t)}{S_w^{\text{cell}}(t)} \cdot \frac{S_w^{\text{virus}}(t)}{S_m^{\text{virus}}(t)}. \tag{7.20}$$

Thus, the ratio of mutant frequencies in the cell versus the mutant frequencies in the virions determines the effect of that mutation on viral genome packaging. The term will be increased ($K_m^{\text{pack}} > 1$), whenever a mutation decreases gRNA packaging.
In order to identify regions that are important for packaging, one may also depict the impact of the mutation $m_{\max}(i)$ that has a maximal impact at position i, akin to the evaluation of maximal effect on gRNA production above (seen in the result Figure 7.10).

7.2.3 Error Correction

The detected nucleotide counts inferred from the NGS output for the DNA library-, the intracellular-, and the viral RNA respectively, are confounded by errors introduced during sequencing and reverse transcription. These errors are corrected analoguously to the derivations in [8], extensively explained in Chapter 5.1.2. The relation of sequences S and NGS reads R containing mutation m at position i is given by

$$S_m(i) \approx \frac{R_m(i)}{\nu} - X_{w\to m}(i) \tag{7.21}$$

and likewise for the wild type

$$S_w(i) \approx \frac{R_w(i)}{\nu}. \tag{7.22}$$

This relation holds for the DNA library, the cellular RNA, and RNA packed into virions. The fraction of RNA that is sequenced is denoted by $0 < \nu < 1$ and $X_{w\to m}(i)$ denotes the number of wild type residues which are falsely detected as mutant m, i.e. the absolute sequencing error. For simplicity, we have skipped the time index t as it is irrelevant for the subsequent arguments.
As in [8], we assume that the absolute sequencing error X is multinomially distributed, i.e. $X_{w\to m}(i) \sim \mathcal{M}\left(S_w(i), \kappa_{w\to m}(i)\right)$. The expectation value for the number of mutations m of the multinomial distribution is trials × success probability. In our context this means that

$$\mathbb{E}\left(X_{w\to m}(i)\right) = S_w(i) \cdot \kappa_{w\to m}(i) \approx \frac{R_w(i)}{\nu} \cdot \kappa_{w\to m}(i), \tag{7.23}$$

where $\kappa_{w\to m}(i)$ denotes the probability of falsely detecting a wild type residue at position i as mutation m.

Using the multinomial model, we then correct the mutation frequencies in the reads for the sequencing errors, yielding the gRNA production efficacy:

$$K_m^{\text{prod}}(i) \approx \frac{\frac{R_m^{\text{DNA}}(i)}{R_w^{\text{DNA}}(i)} - \kappa_{w\to m}^{\text{DNA}}(i)}{\frac{R_m^{\text{cell}}(i)}{R_w^{\text{cell}}(i)} - \kappa_{w\to m}^{\text{cell}}(i)}. \tag{7.24}$$

Analogously, we can derive the effect in packaging

$$K_m^{\text{pack}}(i) \approx \frac{\frac{R_m^{\text{cell}}(i)}{R_w^{\text{cell}}(i)} - \kappa_{w\to m}^{\text{cell}}(i)}{\frac{R_m^{\text{virus}}(i)}{R_w^{\text{virus}}(i)} - \kappa_{w\to m}^{\text{virus}}(i)}. \tag{7.25}$$

The error probability $\kappa_{w\to m}(i)$ can be computed with the data of control experiments, where the RNA is not mutated and will be described in Section 7.2.5.

7.2.4 Statistical Assessment of Effects

To evaluate whether a mutational perturbation of the gRNA production or the packaging process is significant, we apply the same resampling statistics as presented in Section 5.1.3, instead of the single estimates derived above.
In brief, the core idea is to determine the effects $K_{m,w}^{\text{prod}}(i,j)$, respectively $K_{m,w}^{\text{pack}}(i,j)$ for each combination of positions i and j, where the first residue i is mutated and the second residue j is in the wild type configuration. This allows to re-estimate the effect of a mutation m at position i N-times (all pairs of residues with $j \neq i$).

The resultant equations for recomputing mutational effects on gRNA production are

$$K_{m,w}^{\text{prod}}(i,j) \approx \frac{\frac{R_{m,w}^{\text{DNA}}(i,j)}{R_{w,w}^{\text{DNA}}(i,j)} - \kappa_{w\to m,w}^{\text{DNA}}(i,j)}{\frac{R_{m,w}^{\text{cell}}(i,j)}{R_{w,w}^{\text{cell}}(i,j)} - \kappa_{w\to m,w}^{\text{cell}}(i,j)} \tag{7.26}$$

and for the packaging, we derive:

$$K_{m,w}^{\text{pack}}(i,j) \approx \frac{\frac{R_{m,w}^{\text{cell}}(i,j)}{R_{w,w}^{\text{cell}}(i,j)} - \kappa_{w\to m,w}^{\text{cell}}(i,j)}{\frac{R_{m,w}^{\text{virus}}(i,j)}{R_{w,w}^{\text{virus}}(i,j)} - \kappa_{w\to m,w}^{\text{virus}}(i,j)}. \tag{7.27}$$

The resampling gives a non-parametric and unbiased probability distribution of the estimates above.

The statistical test can subsequently be performed on this resampling distribution:
To test whether mutation m at position i increases K^{prod} (decreases gRNA production) more than a constant c, i.e. $\mathcal{H}_0 : \log_2\left(K_m^{\text{prod}}(i)\right) \leq c, \mathcal{H}_1 : \log_2\left(K_m^{\text{prod}}(i)\right) > c$, the raw p-value can be computed according to:

$$P_m^-(i) = \frac{\#\log_2\left(K_{m,w}^{\text{prod}}(i,j)\right) \leq c}{\#K_{m,w}^{\text{prod}}(i,\star)}, \tag{7.28}$$

where # denotes the "number of estimates" and $\star$ indicates that all positions j are evaluated that pass the quality criteria (cf. Section 7.2.6).
To test whether a mutation at position i significantly decreases K^{prod}, the p-value is calculated according to:

$$P_m^+(i) = \frac{\# \log_2 \left(K_{m,w}^{\text{prod}}(i,j) \right) \geq -c}{\# K_{m,w}^{\text{prod}}(i,\star)}. \tag{7.29}$$

The analogous scheme can be used to compute the significance of the effect on packaging.

Note, that one can test any threshold $c \geq 0$ (e.g. 2-fold increase/decrease), whereas the choice $c = 0$ allows to detect all positions that have any effect on the respective function.
In [12], we chose $c = |\frac{1}{N} \sum_i \log_2 \bar{K}_m(i)|$, i.e. the average log-effect over all positions, which allows to test for strong effects.
For K_m^{prod}, we used $c = 0.42$ and for K_m^{pack} we had a threshold of $c = 0.41$, corresponding to ≈ 1.3-fold changes.
All reported p-values were corrected by the false discovery rate (FDR)-based method of Benjamini-Hochberg.
There is a significant impact of mutation m at nucleotide position i, if $P < \alpha$. Here, we used $\alpha = 0.05$.

7.2.5 Estimation of Error Probability κ

For the estimation of the error probability $\kappa_{w\to m}$, control experiments were conducted where the RNA is not mutated, thus $S_{w,m}(i,j) = S_{m,w}(i,j) = 0$.
In the same manner as it is shown in Chapter 5.1.3, the re-sampling scheme allows us to compute statistical properties of the error probability:

$$\mathbb{E}(\kappa_{w\to m}(i)) \approx \frac{1}{N} \sum_j \frac{R_{m,w}(i,j)}{R_{w,w}(i,j)}, \tag{7.30}$$

where N denotes the number of co-varying positions j that have sufficient read coverage, spanning i and j. Thus, along the same lines as the re-sampling scheme for the relative effects on gRNA production and packaging, we can estimate a confidence range for the error probability $\kappa_{w\to m}(i)$.

7.2.6 Quality Criteria

For each pair of reads $R_{m,w}(i,j)$ and $R_{w,w}(i,j)$ we assessed its respective signal-to-noise ratio in the corresponding RNA pool (DNA, cell and virion) according to:

$$D_{m,w}(i,j) \approx \frac{R_{m,w}(i,j)}{R_{w,w} \cdot \kappa_{w\to m,w}(i,j)}. \tag{7.31}$$

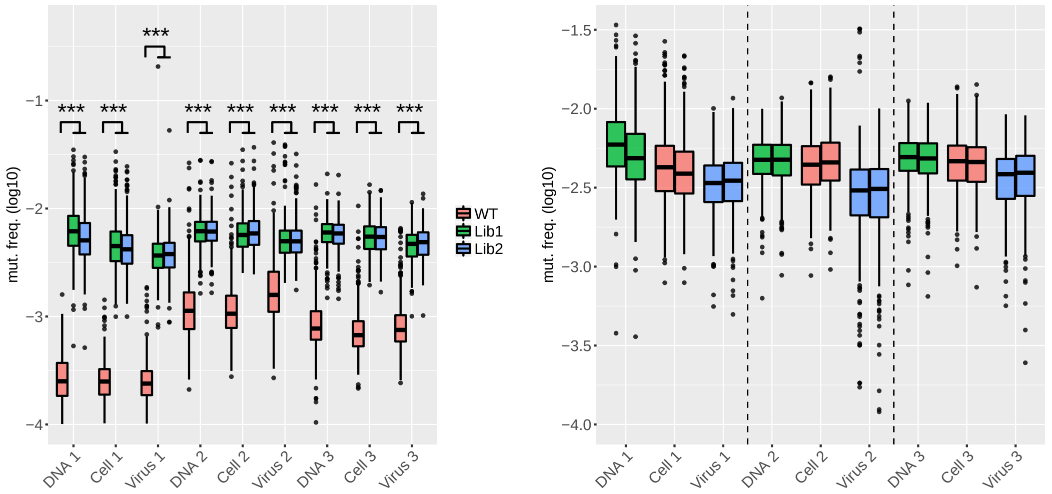
If the ratio is below the user-supplied threshold in both samples (DNA library versus cell and cell versus virion), the corresponding estimate (Eqs. (7.26) or (7.27)) for this position pair is discarded.
If the signal is below the threshold in only one of the samples, the respective estimate is tagged as either being a lower- or upper estimate of the mutations' effect and assigned the value of the median effect estimate on RNA production or packaging respectively, as

explained thoroughly in the Section 5.1.4 in the MIME chapter and Section 6.2.1 in the Software chapter.
For our analysis, we chose a threshold of 2 for the signal-to-noise ratio.
We only evaluated Equations (7.26) and (7.27) for positions j where the total number of sequence fragments covering both i and j was at least 50% of the maximum coverage.
For determining p-values, at least 300 re-estimates had to fulfil the quality criteria.

7.3 Results

7.3.1 Data Description

The entire 5′ UTR and the beginning of the Gag coding region fold in to a series of functional domains, regulating key steps of the HIV-replication cycle [51, 53, 57], including intracellular gRNA production and packaging into nascent virions. Therefore, the first $\sim$ 500 nucleotides of the viral HIV-1 gRNA were targeted for the functional analysis. The *in cell* MIME procedure was performed using six mutant libraries conducted in three independent experiments (two mutant libraries per experiment). The segregated RNAs in the cells and virions were physically separated and sequenced. Additionally, the DNA which was transfected into the cells was sequenced. The three types of samples were derived from both the wild type and the mutant libraries. The sequences of the non-mutated control experiment were used for the above explained correction of RT- and sequencing errors.



(A) Substitution rates for all experiments and libraries.

(B) Error corrected substitution rates of the mutation libraries.

FIGURE 7.4: **A**: Box and whisker plot of substitution frequencies ($\log_{10}$) for wild type library (red) and mutations libraries (green and blue) in DNA, cell and virus raw data of three different experimental replicates. The mutation rates for the mutation libraries are significantly higher than for the wild type controls ($P < 0.01$ respectively, Wilcoxon rank sum test). **B**: Box and whisker plot of the error corrected mutation frequencies ($\log_{10}$) for the mutation libraries in DNA, cell and virus raw data of three different experimental replicates (with two libraries per experiment). Figures adapted from [12], Supplementary Figures.

In total, we aligned 80 million sequences to the reference genome, finding 1.08×10^8

mutations from 2.15×10^{10} nucleotides. Raw substitution rates were found to be significantly higher in the mutant library compared to the wild type control experiments (Figure 7.4a), demonstrating that biologically relevant mutations could be clearly distinguished from the background errors (p-value < 0.01 with Wilcoxon rank sum test, respectively).

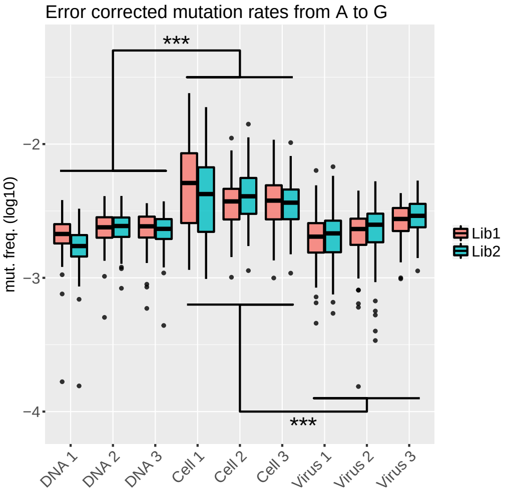


FIGURE 7.5: Box and whisker plot of the error corrected mutation frequencies ($\log_{10}$) of nucleotide A to G for the mutation libraries in DNA, cell and virus raw data of three different experimental replicates. The A to G mutation frequencies in cells are significantly higher than in DNA and virus ($P < 0.01$, Wilcoxon rank sum test). Figure taken from [12], Supplementary Figures.

We were able to use the substitution frequencies in the wild type control to obtain error-corrected mutation frequencies, thus eliminating any biases from errors introduced during library preparation and sequencing. As seen in Figure 7.4b, error corrected mutation rates were similar across all six independent libraries from the three independent experiments (Figure 7.4b) and were highly reproducible for all classes of mutations. They continuously decreased from DNA (median $= 4.8 \times 10^{-3}$), cellular gRNA (median $= 4.2 \times 10^{-3}$) to virion gRNA (median $= 3.3 \times 10^{-3}$) indicating purifying selection with the proceeding viral lifecycle.

Surprisingly, we observed a consistent elevation of A to G mutations within the cellular gRNA compared to the input DNA and the RNA extracted from the virus (p-value < 0.01), clearly recognisable in Figure 7.5. These cellular A to G substitutions were enriched at 5′ AA 3′ and 5′ UA 3′ dinucleotides and seemed to cluster at unpaired adenines in close proximity of double stranded RNA structure. Although we have no clear biological explanation for this phenomenon, it is reminiscent of editing characteristics of the dsRNA adenosine deaminases, ADAR1 or ADAR2 [159, 160]. Despite being an interesting conspicuity, it is most likely unrelated to the investigated processes of gRNA production and packaging. Therefore, we omitted this class of mutation in our analysis.

7.3.2 Regulation of intracellular gRNA Production

First, we focused on the detection of regulatory motifs affecting the viral gRNA production in the infected host cell. The mutation frequencies of the initially transfected DNA library was compared to the frequencies of the viral RNA extracted from the cells. As derived above, the effect of a mutation m at position i can be expressed by the relation of the mutation frequencies in the DNA versus the cell, denoted as $K_m^{\text{prod}}(i)$ (cf. Eq. (7.12)), and in addition, statistically ascertained. $K_m^{\text{prod}}(i) > 1$ means that the mutation decreases gRNA production and stability. Conversely, $K_m^{\text{prod}}(i) < 1$ identifies mutations that increase gRNA production and stability.

The analysis revealed three distinct regions strongly and significantly affecting the packaging process, both in a positive and negative way, mapped to TAR, PolyA and SL2, respectively. The regions are highlighted in Figure 7.6, showing the log effect on gRNA production $log_2 K^{\text{prod}}_m(i)$ for the whole sequence.

Unsurprisingly, gRNA production is positively regulated by the motif found in TAR, indicated by the strong depletion of TAR mutants in the cellular gRNA, clearly seen in Figure 7.7. This is concordant with its known allocation of binding sites for the Tat protein and the cellular factor P-TEFb [161–166]. The mapping of the signals for TAR on the structure in Figure 7.7 shows a substantially stronger effect of the apical part of the stemloop compared to the distal portion of the stem. Again, this agrees with different studies showing that the distal portion of TAR is less important for gene expression compared to the apical portion [167, 168].

The second signal affecting gRNA production was found in 5′ PolyA. Here, an accumulation of mutated gRNA was found inside the cell in relation to the transfected mutant frequency in the DNA library, unveiling a negatively regulating role of PolyA. The ^{73}AAUAAA^{78} hexamer within the apical loop was identified as the precise signal when mapped to the structure (see Figure 7.8). The effect on the pro-

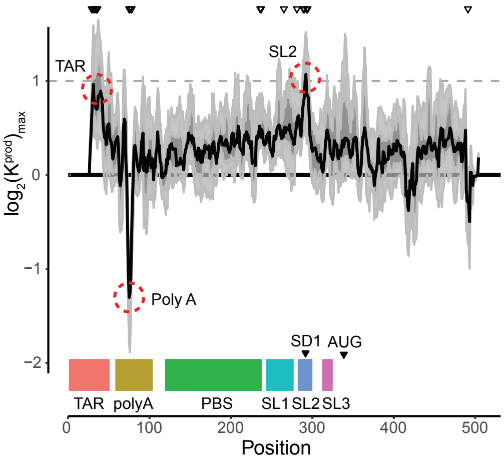


FIGURE 7.6: *In cell* MIME discovery of RNA motifs regulating HIV-1 gRNA production: $\log_2 K^{\text{prod}}$ of the mutation with the maximal effect on RNA production for the 5′ UTR and Gag coding region (smoothed with a linear, two-sided convolution filter of width 2). Functional domains are indicated with coloured boxes below the graph. Black triangles above indicate significant effects. Three regions with significant ($P < 0.05$) and strong ($\log_2 K^{\text{prod}} \geq 1$ or ≤ -1, gray dotted line) effects are highlighted with red circles. Figure extracted from [12].

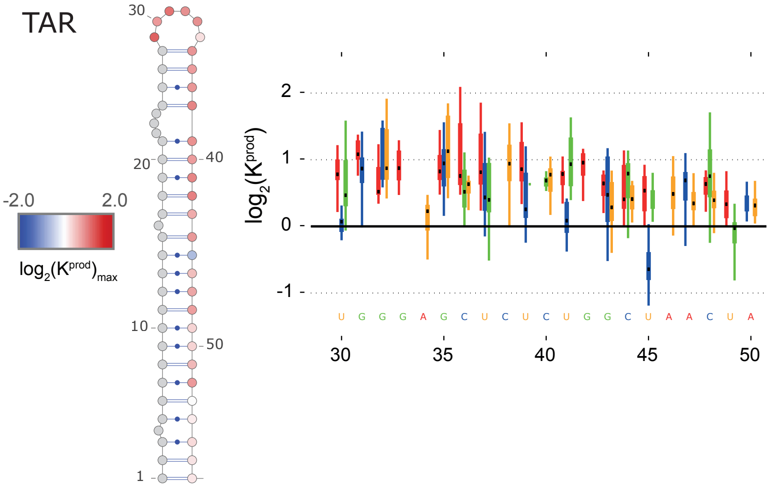


FIGURE 7.7: Regulatory motif for gRNA production in TAR. Mutations with maximal effect on $\log_2 K^{\text{prod}}$ mapped on TAR structure. Box and whisker plot of the effect of each class of mutation on $\log_2 K^{\text{prod}}$. Black dot shows median, box shows IQR, whiskers show extremes (1.5× IQR). Mutation classes are colour coded: red = A, green = C, blue = G, yellow = U. Figure extracted from [12].

duction includes all mutations within this signal, except for the mutant at position 74 from A to U, pointed out in the boxplot in Figure 7.8. As the motifs AAUAAA and AUUAAA are recognised as common cellular polyadenlyation signals [169], the results indicate a role of the cellular polyadelyation machinery in regulating intracellular gRNA levels.

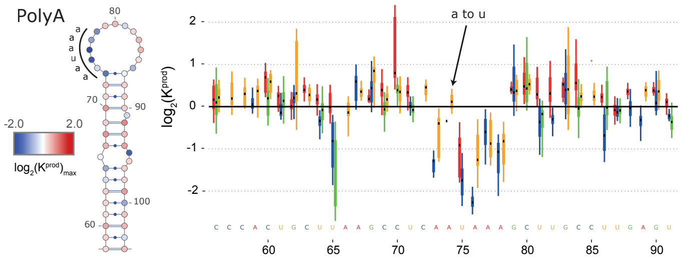


FIGURE 7.8: Regulatory motif for gRNA production in 5′ PolyA. Mutations with maximal effect on $\log_2 K^{\text{prod}}$ mapped on PolyA structure. For explanation of the plot see Fig. 7.7. Figure extracted from [12].

The third signal mapped to SL2, shown by a lower mutation frequency in this region within the cells than in the transfected DNA, thus implying a regulatory activity of SL2 in gRNA production. Interestingly, mutations disrupting gRNA production mapped precisely to the U1 snRNA binding site ^{289}GGUGAGU^{295} (Figure 7.9), and includes all types of mutations. This observation was surprising, as one might expect that disrupting the splice donor site, and with that eliminating the splicing of viral RNAs, would enrich unspliced gRNA in cells. Nevertheless, the opposite effect is observed here, and our data argue that an interaction between U1 snRNA and the splice donor site is required for gRNA production.

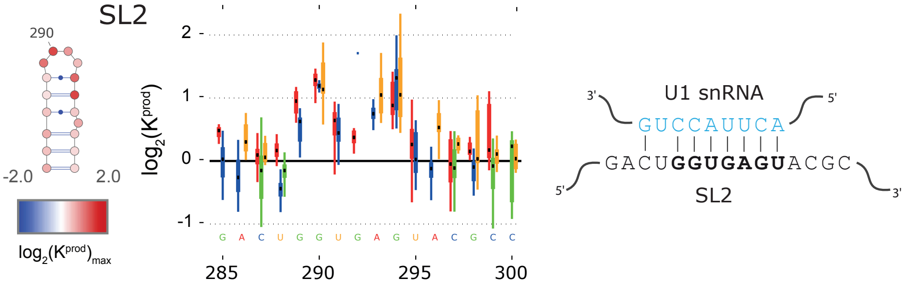


FIGURE 7.9: Regulatory motif for gRNA production in SL2. Mutations with maximal effect on $\log_2 K^{\text{prod}}$ mapped on SL2 structure. Mutations impairing gRNA production cluster to the U1 snRNA binding site. For explanation of the plot see Fig. 7.7. Figure extracted from [12].

7.3.3 Regulation of gRNA Packaging

Following, we determined regulatory motifs affecting the packaging of intracellular gRNA into nascent virions. This process comprises multiple molecular events, including the formation of a protein-RNA packaging complex, its transport to the cell surface, and the packaging and assembly of viral particles, hence presumably involving several regulatory components. The relation of the mutation frequencies in the RNA extracted from the cells and the virions can be used to detect the effect of mutation m at position i on packaging, given by $K_m^{\text{pack}}(i)$ (cf. Eq. (7.20)).

$K_m^{\text{pack}}(i) > 1$ means that the mutation decreases gRNA packaging, and $K_m^{\text{pack}}(i) < 1$ involves a mutation that increases packaging.

We were able to find two strong and significant signals mapped to 5′ PolyA and SL1-SL3, both important for the packaging process, as seen in Figure 7.10 showing the log packaging defect $log_2 K_m^{\text{pack}}(i)$ for the whole sequence.

Suprisingly, we observed that the packaging signal within 5′ PolyA mapped precisely to the same 5′ PolyA sequence ^{73}AAUAAA^{78} that we identified as a strong regulator of gRNA production (Figure 7.11). Like their effect on gRNA production, all mutations to this hexamer sequence impaired gRNA packaging into virions, except for a single mutation at position 74 from A to U, tagged in the boxplot in Figure 7.11. Again, the fact that the motifs AAUAAA and AUUAAA are well known polyadenylation signals suggests an important role for the cellular polyadenylation machinery for gRNA packaging in addition to its influence on gRNA production [169].

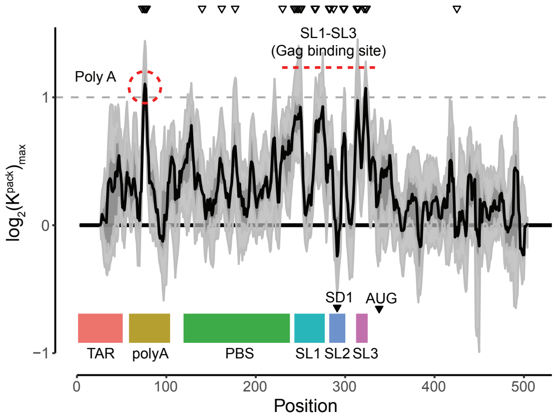


FIGURE 7.10: *In cell* MIME discovery of RNA motifs regulating HIV-1 gRNA packaging: $\log_2 K^{\text{pack}}$ of the mutation with the maximal effect on RNA packaging for the HIV-1 5′ UTR and Gag coding region (smoothed with a linear, two-sided convolution filter of width 2). Functional domains are indicated with coloured boxes. Black triangles above indicate significant effects. Two regions with significant ($P < 0.05$) and strong ($\log_2 K^{\text{pack}} \geq 1$, gray dotted line) effects are highlighted with dot red line/circle. Figure extracted from [12].

The second packaging signal spans the SL1, SL2, and SL3 structures in the psi domain, as highlighted in Figure 7.12. In the beginning of this chapter, we already pointed out that this region was previously identified with *in vitro* MIME experiments as the Pr55$^{\text{Gag}}$ binding site [8]. Consequently, our results here confirm the idea that Pr55$^{\text{Gag}}$ is a central player in the selection of gRNA. Yet, comparing the signal of the *in vitro* Pr55$^{\text{Gag}}$ binding versus the results *in cell* directing gRNA packaging reveals differences between the precise sequence motifs.

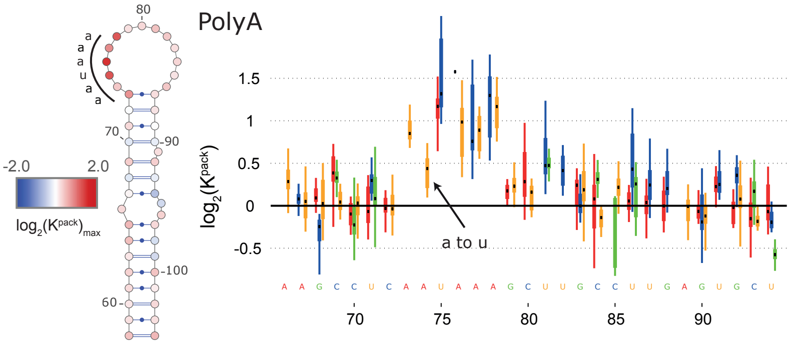


FIGURE 7.11: Regulatory motif for gRNA packaging in 5′ PolyA. Mutations with maximal effect on $\log_2 K^{\text{pack}}$ mapped on PolyA structure. Box and whisker plot of the effect of each class of mutation on $\log_2 K^{\text{pack}}$. Black dot shows median, box shows IQR, whiskers show extremes (1.5× IQR). Mutation classes are colour coded: red = A, green = C, blue = G, yellow = U. Figure extracted from [12].

We compared the precise motifs within the stems of SL1 and SL3, since both were considered important for binding to Pr55^{Gag} *in vitro*. Mapping the effects on the structures of SL1 and SL3 in Figure 7.13 illustrates the differences. Suprisingly, the ^{257}GCGCGC^{262} sequence within the SL1 apical loop, which was found to be crucial for Pr55^{Gag} binding *in vitro*, did not have an significant effect on gRNA packaging in cells.

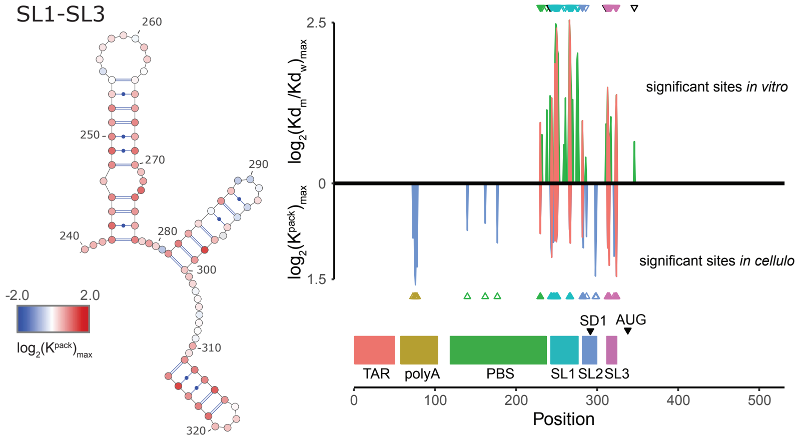


FIGURE 7.12: Regulatory motif for gRNA packaging in Psi domain. Mutations with maximal effect on $\log_2 K^{pack}$ mapped on S1, SL2 and SL3 structures. Qualitative comparison between the significant effects of mutations on Pr55^{Gag} binding determined by *in vitro* MIME (upper portion, green) and the effects on gRNA packaging by *in cell* MIME (lower portion, blue). Sites significantly affecting both are pictured red. Arrows above for *in cell* and below for *in vitro* in the colour of the affected functional domain. Filled arrows show significant effects in both. Figure extracted from [12].

Furthermore, mutations to the stem of SL1 and SL3 had similar effects on packaging in cells. However, mutations of SL1 *in vitro* showed much more damaging impact on protein binding that mutations to SL3.

Additionally, mutation of SL2 seems to be more critical for gRNA packaging in cells compared to Pr55^{Gag} binding *in vitro*, though, less important for packaging than SL1 and SL3 (cf. Figure 7.12).

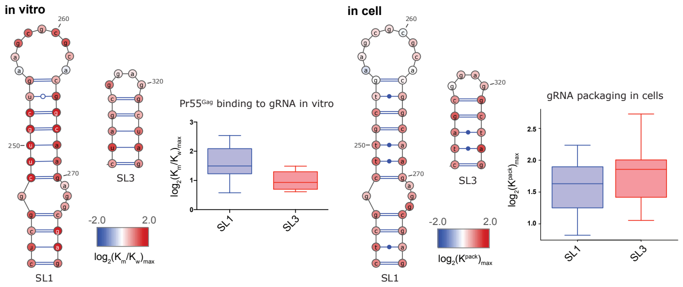


FIGURE 7.13: Comparison of SL1 and SL3 MIME data. Pr55^{Gag} binding to gRNA *in vitro* vs gRNA packaging in cells. Figure extracted from [12], Supplementary Figures.

7.3.4 Discussion

The HIV-1 life cycle is regulated in all its central steps by non-coding gRNA elements, e.g. by providing binding sites for viral and cellular enzymes, factors, and other proteins, or other nucleotide molecules. Many of the crucial processes remain unexplained for now. Yet, the discovery of unknown functional regions in the genome and their corresponding structures, as well as resolving regulatory interactions slowly takes shape and provides a promising opportunity for new therapeutic attempts for fighting the persistent pandemic of HIV-1.

RNA is functionally flexible, meaning one motif can be involved in diverse (also contradictive) processes. This complicates the determination of regulatory motifs with traditional truncation and deletion mutagenesis. The problem is strengthened by the fact that often, especially in viral RNA, regulatory and coding regions overlap, which impedes the identification of functional motifs. MIME [8, 11] provides a method to detect regulatory regions in an unbiased fashion, which has been shown to perform well *in vitro*. Here, we adapted the technique for *in cell* experiments, to shed light on the biological coherences in the HIV-1 life cycle. We were able to find interconnected regulatory motifs within the 5′ UTR important for the production and packaging of viral gRNA, summarised in Figure 7.14. To our suprise, a common signal within the 5′ PolyA apical loop negatively regulated the gRNA production, and in contrast positively regulated the packaging of the RNA into virions.

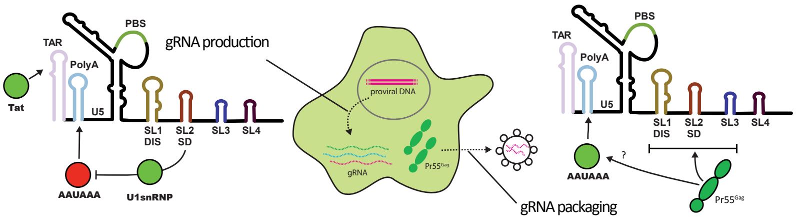


FIGURE 7.14: Five regulatory elements controlling HIV-1 replication. gRNA production is positively regulated by sequences within TAR and by the U1snRNP binding site within SL2. gRNA production is negatively regulated by the AAUAAA motif in 5′ polyA. The U1snRNP binding site is required for repression of 5′ polyadenylation. gRNA packaging into virions requires both the Pr55^{Gag} binding site (SL1-SL3), and the AAUAAA motif in 5′ PolyA. Regulatory elements are colorcoded: positive = green, negative = red. Figure taken from [12].

One of the identified motifs controlling gRNA production allowed us to confirm the appropriate performance of *in cell* MIME: Mutations within TAR, particularly the apical part of the stem, had a negative impact on the ability to produce gRNA. This is conistent with results of various studies localising binding sites in TAR for the Tat protein and transcription elongation factor P-TEFb [170–173].

Furthermore, we discovered a motif within SL2 with an equally high impact on the production process. Futher experiments in [12], which we omitted in this chapter, could approve the involvement of SL2 in the regulation of gRNA production.

Mutations of the third recovered signal within the apical loop of PolyA, on the contrary, led to an increase of gRNA production.

For the process of packaging viral gRNA into virions, we determined two distinct and equally strong signals. The central motif loacalised at the psi domain spanning SL1, SL2 and SL3, closely coincides with the Pr55^{Gag} binding site found *in vitro* [8, 52]. However, we found out that the major signal encountered *in vitro*, the sequence ^{257}GCGCGC^{262} in the apical loop of SL1, was less important for gRNA packaging in cells. After all, we reveal that the SL1 stem including the internal loop are crucial for gRNA packaging, which agrees with the consequence of SL1 deletion leading to severe packaging defects [64, 174].

The second motif with strong impact on gRNA packaging matches the same hexamer in $5'$ PolyA already determined as suppressor of full length gRNA production. Other studies came to analogous results, showing that mutations within PolyA decrease gRNA packaging [175, 176] and a complete deletion of PolyA has a similar effect as deleting both SL1 and SL3 [64, 176].

In summary, we could demonstrate that MIME can be applied to complex cellular environments to infer regulatory elements within functional RNA at single site resolution. We detected motifs in the $5'$ UTR of HIV-1 regulating gRNA production and packaging, one of which being involved in both processes. These results might be useful for further investigation to gain deeper insights into HIV-1 biology and for the development of potential RNA-based drug therapy.

Chapter 8

Combining MIME and DCA Techniques: Accuracy Benchmark

In MIME, presented in Chapter 5 and applied to *in cell* experiments in Chapter 7, **local** frequencies are used to infer functionally important sites and structural relationships within non-coding RNA. The true effects of a mutation on the particular function may be confounded by effects that are caused by mutations at other positions in the sequence. This is particularly the case, if the assessed position is involved in epistatic interactions.
The methods of DCA, explained in Chapter 4, derive a global maximum entropy model to disentangle interactions in order to derive structural elements within a functional sequence.

In this chapter, we want to incorporate the ideas of DCA into the MIME framework to derive a **global** model for describing the probability distribution for the functionally separated sequence pools of the MIME data. Instead of the position-wise frequencies, we apply MIME methods on the probabilities of the respective sequence. The aim is to improve the prediction of functional and structural constraints by combining MIME and DCA methods. After examining the problem of intruding effects of hidden mutations, we derive a global sequence model for the inference of mutational and epistatic effects. We benchmark two DCA methods based on mean field approximation in order to optimise our predictions.

8.1 Intruding Effects from Epistasis

In [177], we develop an optimised protocol for MIME experiments that considers varying protein concentrations in the system and the choice of an optimal mutation rate. The latter has to be large enough to be distinguishable from noise. However, concerns may arise that with a too high mutation rate, predictions for mutational effects at certain position may be confounded by the effects of mutations at other positions within the sequence, illustrated in Figure 8.1.
In the case of no epistasis in the whole assessed sequence, we can assume

$$\frac{Kd_{\mathcal{S}}}{Kd_{\mathcal{W}}} = \prod_{i=1}^{N} \frac{Kd_m}{Kd_w}(i), \tag{8.1}$$

with Kd denoting the respective dissociation constant, e.g. in a binding process, of wt = the wild type sequence without any mutation, $\mathcal{S}$ = any sequence containing either a wild type or a mutation at each position i.

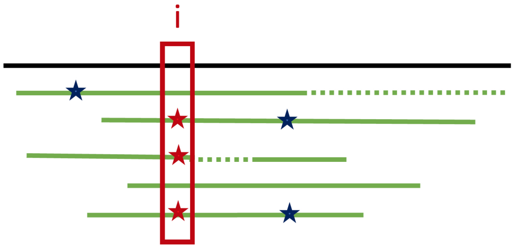


FIGURE 8.1: Confounding effects of mutations. The sequence reads (green), which are mapped to the reference (black) comprise several mutations (stars). If the mutations of a certain position i are evaluated locally (red stars), they might be affected by mutations at other positions within the sequence.

For the case of epistatic interactions, that do not involve the assessed position i, we have

$$\frac{Kd_{\mathcal{S}}}{Kd_{\mathcal{W}}} = \frac{Kd_m}{Kd_w}(i) \cdot \frac{Kd_J}{Kd_{\text{wt}}}, \tag{8.2}$$

where position i has no interaction with any position $j \in J$. Positions J may confer epistatic effects among each other.

As an illustrative example, we assume a sequence of length $L = 2$ and both positions are interacting with each other. Obviously, we have for the wild type

$$\frac{Kd_{w,w}}{Kd_{\text{wt}}} = 1. \tag{8.3}$$

In this example, position 1 has no effect on the function when mutated, i.e.

$$\frac{Kd_{m,w}}{Kd_{\text{wt}}} = \frac{Kd_m}{Kd_w}(1) \cdot \frac{Kd_w}{Kd_w}(2) = 1, \tag{8.4}$$

and a mutation at position 2 leads to a 2-fold increase, i.e. a relative effect of

$$\frac{Kd_{w,m}}{Kd_{\text{wt}}} = \frac{Kd_w}{Kd_w}(1) \cdot \frac{Kd_m}{Kd_w}(2) = 2. \tag{8.5}$$

In the case of no epistasis, the effect of the double mutant is simply given by the product of both mutational effects

$$\frac{Kd_{m,m}}{Kd_{\text{wt}}} = \frac{Kd_m}{Kd_w}(1) \cdot \frac{Kd_m}{Kd_w}(2) = 2. \tag{8.6}$$

When we compute relative Kd values based on the MIME output, i.e. position-wise mutation frequencies, we focus on one base after the other, basically ignoring the presence/absence of additional mutations. Hence, instead of looking at the effect of the whole sequence where only one position i is mutated, we take all sequences into account, which have a mutation at position i, regardless of the remaining nucleotide composition (cf. Figure 8.1). This means, for the effect of a mutation at position 1 in the example above, we look at the set of sequences with mutation m at position i, regardless of the rest of the sequence: $\mathcal{M}(1) = S_{m,*} = S_{m,w} + S_{m,m}$. The same applies for the apparent wild type sequence set, consisting of $\mathcal{W}(1) = S_{w,*} = S_{w,w} + S_{w,m}$. Both sets are confounded by the presence of mutants at position 2.

For the case of protein in excess, i.e. RNA $<<$ protein, it can be shown that the following equation holds, if we would assume that the frequencies of the sequences within set $\mathcal{M}$ are identical:

$$Kd_{\mathcal{M}}(i) = \frac{1}{|\mathcal{M}(i)|} \sum_{\mathcal{S} \in \mathcal{M}(i)} Kd_{\mathcal{S}}. \tag{8.7}$$

For our example we have for position $i = 1$

$$\begin{aligned} Kd_{\mathcal{M}}(1) &= \frac{1}{2} \cdot Kd_{\text{wt}} \cdot (1+2) \\ &= Kd_{\text{wt}} \cdot 1.5, \end{aligned} \tag{8.8}$$

i.e. the absolute Kd estimate for the set of sequences which contains a mutation at position i is false: $\frac{Kd_{\mathcal{M}}}{Kd_{\text{wt}}}(1) = 1.5 \neq 1 = \frac{Kd_{m,w}}{Kd_{\text{wt}}}$. Analogously, for the wild type set we have

$$\begin{aligned} Kd_{\mathcal{W}}(1) &= \frac{1}{2} \cdot Kd_{\text{wt}} \cdot (1+2) \\ &= Kd_{\text{wt}} \cdot 1.5. \end{aligned} \tag{8.9}$$

Again, we compute the absolute effect incorrectly, with $\frac{Kd_{\mathcal{W}}}{Kd_{\text{wt}}}(1) = 1.5 \neq 1 = \frac{Kd_{w,w}}{Kd_{\text{wt}}}$.
However, since we are inspecting the relative effect of each mutation in comparison to the wild type sequence, we have

$$\frac{Kd_{\mathcal{M}}}{Kd_{\mathcal{W}}}(1) = \frac{Kd_m}{Kd_w}(1) = 1, \tag{8.10}$$

meaning the relative effect can be determined correctly.
In summary, the following results were concluded for the case of no epistasis involving position i:

- When protein is used in excess (binding saturation), the true relative effect is determined.
$\Rightarrow \frac{Kd_{\mathcal{M}}}{Kd_{\mathcal{W}}}(i) = \frac{Kd_m}{Kd_w}(i)$
- When we are in the linear binding regimen (little protein), the true relative effect is determined.
$\Rightarrow \frac{Kd_{\mathcal{M}}}{Kd_{\mathcal{W}}}(i) = \frac{Kd_m}{Kd_w}(i)$
- In the worst case scenario (non-linear binding regimen, many confounding mutations $p \longrightarrow 1$, extremely strong impact of confounders $\frac{Kd_C}{Kd_w}(i) \longrightarrow \infty$) we get a false negative prediction.
$\Rightarrow \frac{Kd_{\mathcal{M}}}{Kd_{\mathcal{W}}}(i) \longrightarrow 1$

For the case of epistasis involving position i, we get a different result:
We extend the example from above with an epistatic interaction of the two positions, such that the co-occurrence of both mutations lead to a 3-fold amplification of the effect compared to no epistasis. Note: Instead of using the epistasis value E introduced in Chapter 5.1.5, we denote the magnitudal epistatic effect in the following as $\mathcal{E} = e^{-E}$, i.e. the amount of increase or decrease of the absolute effect (we will refer to that later in this chapter).

With an epistatic effect of $\mathcal{E}_{1,2} = 3$, we have a relative effect of the double mutant given by

$$\frac{Kd_{m,m}}{Kd_{\text{wt}}} = \frac{Kd_{m,w}}{Kd_{\text{wt}}} \cdot \frac{Kd_{w,m}}{Kd_{\text{wt}}} \cdot \mathcal{E}_{1,2} = 6. \tag{8.11}$$

If we again consider the set of sequences containing a mutation at position 1 we have

$$\begin{aligned} Kd_{\mathcal{M}}(1) &= \frac{1}{2} \cdot Kd_{\text{wt}} \cdot (1+6) \\ &= Kd_{\text{wt}} \cdot 3.5. \end{aligned} \tag{8.12}$$

The effect for the wild type is given by the average of the set

$$\begin{aligned} Kd_{\mathcal{W}}(1) &= \frac{1}{2} \cdot Kd_{\text{wt}} \cdot (1+2) \\ &= Kd_{\text{wt}} \cdot 1.5. \end{aligned} \tag{8.13}$$

Hence, the relative Kd for a mutation at position 1 is given by

$$\frac{Kd_{\mathcal{M}}}{Kd_{\mathcal{W}}}(1) = \frac{3.5}{1.5} = 2.3, \tag{8.14}$$

which is a vast overestimation of the true effect $\frac{Kd_{m,w}}{Kd_{\text{wt}}} = 1$.
When we investigate the effect for a mutation at position 2, we come to a similarly drastic deviation: The average absolute effect of a mutation at position 2 is given by

$$\begin{aligned} Kd_{\mathcal{M}}(2) &= \frac{1}{2} \cdot Kd_{\text{wt}} \cdot (2+6) \\ &= Kd_{\text{wt}} \cdot 4, \end{aligned} \tag{8.15}$$

and the for the wild type we derive

$$\begin{aligned} Kd_{\mathcal{W}}(2) &= \frac{1}{2} \cdot Kd_{\text{wt}} \cdot (1+1) \\ &= Kd_{\text{wt}} \cdot 1. \end{aligned} \tag{8.16}$$

The relative effect is thus given by

$$\frac{Kd_{\mathcal{M}}}{Kd_{\mathcal{W}}}(2) = \frac{4}{1} = 4, \tag{8.17}$$

a 2-fold deviation from the true effect for a single mutation at position 2 $\frac{Kd_{m,w}}{Kd_{\text{wt}}} = 2$.

Hence, we seek to find a possibility to predict true functional and epistatic effects with the help of DCA.

8.2 Potts Model applied to the MIME Framework

Recapitulating, MIME allows the quantification of relative effects that a certain mutation has on a particular function in comparison to the wild type. For this, the sequence numbers S of the selected fraction b and the non-selected fraction u are approximated by statistically assessment. Furthermore, the method facilitates the derivation of the structural commitment of each position. Simply put, the relative effect is calculated by the ratios of mutant counts m vs the wild type counts w of the selected and unselected fraction, cf.

Eq. 5.6. This relation can also be expressed by the ratio of local probabilities of mutant and wild type at position i :

$$\begin{aligned} K_m(i) = \frac{Kd_m}{Kd_w}(i) &= \frac{S^u_m}{S^u_w}(i) \cdot \frac{S^b_w}{S^b_m}(i) \\ &= \frac{\frac{S^u_m}{S^u}}{\frac{S^u_w}{S^u}}(i) \cdot \frac{\frac{S^b_w}{S^b}}{\frac{S^b_m}{S^b}}(i) \\ &= \frac{P^u_m}{P^u_w}(i) \cdot \frac{P^b_w}{P^b_m}(i) \end{aligned} \tag{8.18}$$

(Note: This is the raw computation of the Kd, without error correction and quality filtering).

This way of relations may contain transitive information, as only local frequencies are taken into account. The frequencies are considered in isolation. To disentangle the direct from indirect correlations, we aim to apply methods of the direct coupling analysis, which adapts a global model for the probability distribution of sequence variants, as introduced in Chapter 4.3. Our aim is to incorporate these inferred direct interactions to derive quantitative measures for mutational effects.

Instead of using the local frequencies of nucleotides, we fit the maximum entropy model to deduce the probability distribution (cf. Eq. (4.17))

$$P(\sigma) = \frac{1}{Z} e^{\mathcal{H}(\sigma)} = \frac{1}{Z} \exp\left(\sum_i h_i(\sigma_i) + \sum_{i<j} g_{ij}(\sigma_i, \sigma_j) \right)$$

for the wild type sequence σ_{wt} containing no mutation and mutated sequences σ_{m_i} with only one mutation m at position i

$$P(\sigma_{\text{wt}}) = P(\sigma \mid \sigma_i = w \; \forall \, i) \text{ and} \tag{8.19}$$

$$P(\sigma_{m_i}) = P(\sigma \mid \sigma_i = m, \, \sigma_j = w \; \forall \, j \neq i). \tag{8.20}$$

These probabilities are plugged into Equation (8.18), substituting the local frequencies, to determine the relative effects of the single mutations:

$$\begin{aligned} K^P_m(i) = \frac{Kd_m}{Kd_w}(i) =& \frac{P^u(\sigma_{m_i})}{P^u(\sigma_{\text{wt}})} \cdot \frac{P^b(\sigma_{\text{wt}})}{P^b(\sigma_{m_i})} \\ =& \frac{\frac{1}{Z^u} e^{-\mathcal{H}^u(\sigma_{m_i})}}{\frac{1}{Z^u} e^{-\mathcal{H}^u(\sigma_{\text{wt}})}} \cdot \frac{\frac{1}{Z^b} e^{-\mathcal{H}^b(\sigma_{\text{wt}})}}{\frac{1}{Z^b} e^{-\mathcal{H}^b(\sigma_{m_i})}} \\ =& \frac{e^{-\mathcal{H}^u(\sigma_{m_i})}}{e^{-\mathcal{H}^u(\sigma_{\text{wt}})}} \cdot \frac{e^{-\mathcal{H}^b(\sigma_{\text{wt}})}}{e^{-\mathcal{H}^b(\sigma_{m_i})}} \\ =& e^{-\mathcal{H}^u(\sigma_{m_i}) + \mathcal{H}^u(\sigma_{\text{wt}}) + \mathcal{H}^b(\sigma_{m_i}) - \mathcal{H}^b(\sigma_{\text{wt}})}. \end{aligned} \tag{8.21}$$

Conveniently, for computing the relative effect with respect to the wild type sequence, the costly partition function Z cancels out and the Hamiltonians are sufficient.

8.2.1 Gauge Fixing

As we saw in Chapter 4.3.1, the number of independent parameters in the statistical model is given by $\binom{L}{2}q^2 + Lq$, with a sequence length of L and q possible symbols per site. In order to attain a unique solution, the number of parameters has to be fixed to

match the number of constraints, which is $\binom{L}{2}(q-1)^2 + L(q-1)$.
The gauge fixing can be achieved by setting all couplings and local fields for e.g. the last symbol q to zero: $h_i(q) = 0$ and $g_{ij}(q,*) = g_{ij}(*,q) = 0$. The remaining parameters are thus measured with respect to state q.
In the cases for the application to an MSA [9], the symbol "–" (insertion or deletion) is set to zero. In our case, we do not consider indels. When the NGS reads are mapped to the reference, indels are ignored and only position-wise nucleotide frequencies are of interest. Instead of setting the parameters for one of the nucleotides to zero, we reformat the count files: We reorder the counts for the nucleotides

```
A C G U
```

to

```
wt mut1 mut2 mut3
```

where the three possible mutations mut1, mut2 and mut3 are in the same order as A, C, G and U (e.g. wt = C, mut1 = A, mut2 = G, mut3 = U).
Now, always the wild type nucleotide for each position is set to zero to achieve gauge fixing:

$$
\begin{aligned}
h_i(\mathrm{wt}) &= 0 \\
g_{ij}(\mathrm{wt},*) = g_{ij}(*,\mathrm{wt}) &= 0.
\end{aligned}
$$

Since the Hamiltonian for the wild type sequence will be always zero (and thus $e^0 = 1$), the exponential Hamiltonian for a mutation sequence is already given in relation to the wild type ($\frac{e^{-\mathcal{H}(\sigma^m)}}{e^{-\mathcal{H}(\sigma^w)}} = e^{-\mathcal{H}(\sigma^m)}$).
As a result, the relative effect of a mutation is given by

$$
\begin{aligned}
K_m(i) &= \frac{e^{-\mathcal{H}^u(\sigma^{m_i})}}{e^{-\mathcal{H}^b(\sigma^{m_i})}} \\
&= e^{-\mathcal{H}^u(\sigma^{m_i}) + \mathcal{H}^b(\sigma^{m_i})}. \qquad (8.22)
\end{aligned}
$$

8.2.2 Regularisation

Pseudocounts

The determination of couplings with mean field approximation involves the inversion of the correlation matrix (cf. Eq. (4.41) in Chapter 4.3.2). In the case of undersampling, the matrix is not invertible, which would lead to infinite coupling values.
The most simple way for regularisation is to add pseudo-observations to the real data. If there is a variable which never occurs in the data set of M sequences, i.e. $S_i(A) = 0$, one solution would be to add an "extra" $(M+1)^{th}$ observation where this variable occurs.
For the analysis of protein and RNA sequence data (MSA), this method is very popular due to strong undersampling [9, 178]. Here, typically a prior distribution is chosen where each symbol of S_i is considered equally likely. The strength of the pseudocount is given by a weight θ and is usually chosen according to the sample size M, e.g. $\theta = \frac{1}{M}$ or similar. The amplitude θ is expected to vanish with $M \to \infty$.

In all published studies of DCA [9, 128, 129, 132, 141] applied to sequence families in MSAs, conserved sequences are considered. Only functional and prevailed sequences are analysed, which emerged and became evolutionary fixed in the population after many generations due to selection pressure. Hence, many intermediate sequences, either not

functional, or not as reproductive as other variants, may not be contained within the sample set, leading to a massive undersampling.
On the contrary, in MIME just one single step of evolution is conducted, and both the functionally selected and unselected sequence sets are taken into consideration. Although, the wild type RNA is mutated with a certain, small, mutation rate p_{mut}, millions of sequences are mutated in parallel. Since the mutagenesis procedure per site and symbol is assumed to be independent and (at least roughly) identically Gaussian distributed, and the sample size is very huge, we expect to have a complete sampling in terms of heterogeneity per site and occurrence of symbol per site.

L_1- and L_2 Regularisation

Another way to prevent inferred coupling terms from being infinite, is the L_2 regularisation term, which is added to the log-likelihood. Here, a prior probability distribution for the couplings is considered which penalises large coupling values [178].
Since we do not have infinite couplings due to perfect sampling, i.e. all symbols per site are observed, we do not need to take this kind of regularisation into consideration.

The L_1-norm regularisation of couplings, corresponding to a Laplacian prior distribution, produces a sparse interaction matrix [178]. Instead of small coupling values, this regularisation favours coupling terms that are zero and is again achieved by adding the penalty term to the log-likelihood. This regularisation is of particular interest, if one expects a sparse but strong interaction network. Especially for the mapping of pairwise contacts the large coupling values are of interest.
However, we seek to quantify effects as accurate as possible, which may also involve low interaction terms. Therefore, we refrain from regularisation in our analysis. Note, that the absence of a ground truth complicates the choice of an optimal weight parameter for the regularisation term.

8.2.3 Parameter Determination

In Chapter 4.3, different methods are introduced to infer the model parameters h and g. DCA is applied on multiple sequence alignments in order to retrieve information about direct interactions between pairs of residues. Pseudo-Likelihood-Maximisation (PLM) has been shown to yield excelling results in comparison to other methods such as the mean field approximation [128]. Nevertheless, PLM requires the knowledge about the distribution of the full-length sample sequences in order to derive the pseudo-likelihood. In MIME, however, the samples are sequenced using NGS, resulting in a massive amount of short reads, i.e. no information about the full-length sequence frequencies. These conditions do not suit the capability of PLM.
Besides, the adaption of DCA in MIME is fortunately applicable without the partition function Z, which enormously simplifies the parameter inference. Therefore, the mean field approximation is the most simple way to determine parameters h and g, conveniently using the standard MIME analysis input, i.e. the position– and pairwise nucleotide frequencies $f_i(A)$ and $f_{ij}(A, B)$.

In the naive mean field approximation, the coupling terms g^{MF} can be inferred by inverting the covariance matrix $C_{ij}(A, B) = f_{ij}(A, B) - f_i(A)f_j(B)$, as derived in Equation (4.41) in Chapter 4.3:

$$g_{ij}^{\mathrm{MF}}(A, B) = -(C^{-1})_{ij}(A, B). \tag{8.23}$$

Parameters h^{MF} can be subsequently ascertained, derived in Equation (4.40), with

$$h_i^{\mathrm{MF}}(A) = \ln\left(\frac{f_i(A)}{f_i(\mathrm{wt})}\right) - \sum_{j\neq i}\sum_{B\neq \mathrm{wt}} g_{ij}(A,B) f_i(B). \tag{8.24}$$

Adding the second order expansion of the Gibbs potential (Onsagar correction term) to the mean field equation yields the TAP equation, as shown in Chapter 4.3.2. The coupling terms g^{TAP} are derived according to Equation (4.46)

$$g_{ij}^{\mathrm{TAP}}(A,B) = \frac{\sqrt{1 - 8\cdot f_i(A) f_j(B)[(C^{-1})_{ij}(A,B)]} - 1}{4\cdot f_i(A)\cdot f_j(B)}, \tag{8.25}$$

as well as the parameters h^{TAP} in Equation (4.44):

$$\begin{aligned} h_i^{\mathrm{TAP}}(A) =& \ln\left(\frac{f_i(A)}{f_i(\mathrm{wt})}\right) - \sum_{j\neq i}\sum_{B\neq \mathrm{wt}} g_{ij}(A,B)\cdot f_i(B) \\ &+ \sum_{j\neq i}\sum_{B\neq \mathrm{wt}} g_{ij}(A,B)^2 \cdot f_i(A)\cdot \left(1 - f_j(B)^2\right). \end{aligned} \tag{8.26}$$

8.3 Set up the Test Data

In our benchmark, we aim to compare the accuracy regarding the prediction of single site effects, as well as the quantification of epistasis with the methods stated above. Therefore, we generate a test data set in order to examine the deviation from the ground truth and to be able to compare the precision of each method.

8.3.1 Sampling Sequence Variants

For the test data set, we simulate the whole process of mutating a certain number of sequences M of length L, comprising q symbols, with a mutation rate p_{mut}, since the number of available sequences, is much lower than the number of possible sequence variants $S = q^L$.
Although realistic RNA data would include four nucleotides ($q = 4$ symbols), the sampling of the data set with $q = 2$ (wild type and mutant) is sufficient for the test purposes here. The number of sampled, mutated sequences is chosen to be $M = 12\cdot 10^6$.

We simulate the mutation process sequence by sequence, starting from the original wild type reference. Sampling each position of each sequence with a Bernoulli experiment would be too much computational effort. Instead, we draw the number of mutations n^* for a sequence in the first step of the mutant sampling. The probability for n mutations is given by the binomial distribution

$$p_n = \binom{L}{n}\cdot p_{\mathrm{mut}}^n \cdot (1 - p_{\mathrm{mut}})^{L-n}.$$

Then, we uniformly draw a variant with n^* mutations. For this, we give IDs for all possible sequence variants as follows: The wild type sequence with 0 mutations has ID 1, the sequence with only one mutation at position 1 has ID 2, one mutation at position 2 has ID 3, etc. The sequence is drawn from the ID range comprising n^* mutations.

We restrict the maximal number of mutations n_{max}. The mutation probability is quite

small, resulting in a very low probability for large numbers of mutations per sequence, such that we can neglect these. The expected occurrence of a certain sequence variant σ with n_σ mutations is given by $\mathbb{E}\left(S_\sigma^{\text{tot}}\right) = M \cdot p_{\text{mut}}^{n_\sigma} \cdot (1 - p_{\text{mut}})^{L-n_\sigma}$. The expected absolute number of a sequence variant $\mathbb{E}\left(S_\sigma^{\text{tot}}\right)$ with n_σ mutations has to be large enough in order to segregate into the functionally selected and unselected pool.
As an example, for length $L = 50$ and a mutation rate of $p_{\text{mut}} = 0.01$, the expected emergence of a particular sequence variant with $n = 4$ mutations is $\mathbb{E}\left(S_\sigma^{\text{tot}}\right) = 12 \cdot 10^6 \cdot 0.01^4 \cdot (1 - 0.01)^{46} \approx 0.07$, so not even one is expected. The expected occurrence of a variant with $n = 3$ mutations is given by $\mathbb{E}\left(S_\sigma^{\text{tot}}\right) = 12 \cdot 10^6 \cdot 0.01^3 \cdot (1 - 0.01)^{47} \approx 7.48$. Here, we would consider $\mathbb{E}\left(S_\sigma^{\text{tot}}\right)$ large enough to be accounted for. So in this case, the choice for the maximal number of mutations would be $n_{\max} = 3$.

After sampling all M sequences, we count the appearance of each sequence S_σ^{tot} and the total frequency of a sequence $f_\sigma^{\text{tot}} = \frac{S_\sigma^{\text{tot}}}{M}$.

8.3.2 Sampling Kds and Epistasis

The absolute dissociation constant of the wild type is given by Kd_{wt}, which we chose to be 1 for convenience. A mutation m at position i has an effect on the function with probability p_{kd}. As default we set $Kd_m(i) = Kd_{\text{wt}}$ for all positions i that have no effect. If there is an effect, it is randomly sampled from the lognormal distribution with standard deviation of 1

$$Kd_m(i) = \exp(X), \text{ with } X \sim \mathcal{N}(0,1). \tag{8.27}$$

As mentioned in Chapter 5.1 in Equation 5.18, pairwise epistasis is given by

$$E(i,j) = \log\left(\frac{K_{m,w}(i,j) \cdot K_{w,m}(i,j)}{K_{m,m}(i,j)}\right),$$

with $-\infty < E < \infty$, denoting the decrease and increase of fitness. We are observing the magnitude of how much a double mutant is increasing or lowering the effect of the two single mutational effects. It could be denoted as the "defect epistasis" so to say, i.e. the inverse of the fitness epistasis, now in the range of $0 < \mathcal{E} < \infty$:

$$\mathcal{E}(i,j) = e^{-E(i,j)} = \frac{K_{m,m}(i,j)}{K_{m,w}(i,j) \cdot K_{w,m}(i,j)}. \tag{8.28}$$

In the following, "epistatic effects" refers to this measure of magnitude.
Two position pairs i and j have epistatic effects with probability $p_{\text{epistasis}}$ and the effect is also considered to be lognormal distributed, where we chose a standard deviation of 1

$$\mathcal{E}_{i,j} = \exp(X), \text{ with } X \sim \mathcal{N}(0,1). \tag{8.29}$$

Note: also mutations without an effect can be involved in epistatic effects, if they occur together with other mutations.
The Kd for a sequence variant is computed by multiplying the single Kds for each residue of the sequence and the epistasis values for pairwise mutations:

$$Kd_\sigma = \prod_{i=1}^{L}\left(Kd_{\sigma_i}(i) \prod_{j>i} \mathcal{E}_{i,j}\right). \tag{8.30}$$

8.3.3 Deriving Species Frequencies in Equilibrium

Model Specification: The mass action kinetics between the different sequence variants describing the binding competition experiment presented in Chapter 5.1.1 is given by

$$dS_\sigma^u = -k_{\text{on},\sigma} \cdot S_\sigma^u \cdot B + k_{\text{off},\sigma} \cdot S_\sigma^b \tag{8.31}$$

$$dS_\sigma^b = -k_{\text{off},\sigma} \cdot S_\sigma^b + k_{\text{on},\sigma} \cdot S_\sigma^u \cdot B \tag{8.32}$$

$$dB = \sum_\sigma k_{\text{off},\sigma} \cdot S_\sigma^b - k_{\text{on},\sigma} \cdot S_\sigma^u \cdot B \tag{8.33}$$

where S^u and S^b denote the unbound and bound concentration of sequence variant σ. B gives the concentration of free protein, and k_{on} and k_{off} the respective rates of dissociation and association. The ratio of the latter gives the dissociation constant $Kd = \frac{k_{\text{off}}}{k_{\text{on}}}$.
In a steady state condition, the rates of change (left hand side) become zero. Solving one of the first equations for the unbound fraction of sequence σ results in

$$S_\sigma^u = \frac{S_\sigma^b \cdot Kd_\sigma}{B}. \tag{8.34}$$

The total concentration of sequence σ is simply the sum of unbound and bound fraction

$$S_\sigma^{\text{tot}} = S_\sigma^u + S_\sigma^b. \tag{8.35}$$

If we plug in (8.34) in (8.35) and solve for the bound sequence, we get

$$S_\sigma^b = \frac{B \cdot S_\sigma^{\text{tot}}}{B + Kd_\sigma} \tag{8.36}$$

The total number of protein in the system is given by

$$B_{\text{tot}} = B + \sum_\sigma S_\sigma^b, \tag{8.37}$$

and plugging in (8.36) yields the equation

$$B_{\text{tot}} - \sum_\sigma \frac{B \cdot S_\sigma^{\text{tot}}}{B + Kd_\sigma} - B = 0. \tag{8.38}$$

The square of Equation (8.38) is the objective function to minimise in order to get the number of unbound protein B in the steady state condition. The input are the sampled sequence concentrations and Kds, and a fixed total concentration of protein in the system B_{tot}. After obtaining B, we can compute the amount of bound and unbound sequences S^b and S^u (Equations (8.36) and (8.35)), corresponding to read counts in the real data.
The determination of the bound fraction in Eq. (8.36) involves the Kd values, which are real numbers, with $0 < Kd < \infty$. To avoid biases due to rounding errors, and to simulate a result close to reality, we sample the allocation for bound or unbound fraction according to their relative frequencies. For example, if $S_\sigma^{\text{tot}} = 80$ and the calculated fraction of bound sequences is given by $S_\sigma^b = 28.712$, we draw the number of bound sequences according the binomial distribution $S_\sigma^{b,\text{sim}} \sim \mathcal{B}(80, \frac{28.712}{80})$.
In our simulations, we chose a protein concentration of $2 \times M$.

8.3.4 Adding Noise

Since read sequencing is error prone, we simulate these errors by adding random noise. The probability that a nucleotide at one site is incorrectly identified is given by p_{err}. For each read in the bound and unbound fraction of a sequence, we simulate the erroneous nucleotide detection by drawing the number of errors from a binomial distribution

$$n_{\text{err}} \sim \mathcal{B}(L, p_{\text{err}}),$$

and uniformly draw the positions i where the error occurs

$$i_{\text{err}} \sim \mathcal{U}(L).$$

The wrongly identified sequence ω is incremented ($S_{\omega}^{u/b} = S_{\omega}^{u/b} + 1$) and the number of the original sequence σ is decreased by 1 ($S_{\sigma}^{u/b} = S_{\sigma}^{u/b} - 1$). Since the vast majority of the nucleotides are in wild type configuration, this error affects in particular false positive mutant detections.

Note: because of reasons stated above, we allow only a total number of (also wrongly detected) mutations of $n_{\max}$.

We use a sequencing error probability per site smaller than the mutation rate, e.g. for a mutation probability of $p_{\text{mut}} \approx 10^{-2}$ we choose $p_{\text{err}} = 10^{-3}$.

8.3.5 Output

After sampling the error, the final sequence variant counts for the bound and unbound pools are used to count the position– and pairwise nucleotide occurrences, respectively. Additionally, a "control experiment" with no true mutations ($p_{\text{mut}} = 0$), but the same error rate is sampled, also resulting in a bound and a unbound fraction of the wild type sequences. For each single position and pair of positions the nucleotides are counted, including the (wrongly appearing) mutations. This wild type data is required for the error correction procedure in the following analysis.

The nucleotide counts of the four samples are written into count files (1d and 2d) in the given result directory, as described in the Software Chapter 6.1.2.

The files for the mutant and wild type samples are numbered with a consecutive id in the following way:

number	**sample type**
1	wild type bound
2	wild type unbound
3	mutation bound
4	mutation bunound

We also save the sampled Kds and epistasis values in the result directory, in order to compare the predictions of the different methods with the ground truth.

8.3.6 Implementation

The sampling routines are implemented in C++.

For minimising the objective function, we used a numerical L-BFGS-B solver (from `https://github.com/PatWie/CppNumericalSolvers`).

The program is called from the command line with

```
DCA_Benchmark <resultDir> <L> <pMut> <pError> <pKd> <pEpi>
```

and requires the parameters:

resultDir	(string)	path to directory where the output files are saved
L	(int)	sequence length
pMut	(float)	probability per site of a sampled read to be mutated
pErr	(float)	probability per site of a sampled read to be contain an error
pKd	(float)	probability per site to have an effect on the function when mutated
pEpi	(float)	probability per position pair to have epistatic effects when mutated

8.4 Evalutating different Methods

The different methods for the determination of the mutational and epistatic effects are also implemented in C++. For the inversion of the covariance matrix we used the Eigen package. For each method, the determined single site Kds, and the epistatic effects are respectively written into an outputfile. The script for the evaluation and for generating the plots is implemented in R.

We compare the precision of the standard MIME method (cf. Chapter 5.1), which we denote as "MIME" in the following, with different DCA methods based on the mean field approximation, including the naive mean field approach ("MF"), and the TAP equation ("TAP").
We focus on the raw Kd estimation and omit the quality filtering such as the signal-to-noise ratio. The input for all methods are the nucleotide counts written into output files, as described in the sampling routine above. The frequencies are error corrected according to Section 5.1.2:

$$S_m^{\mathrm{corr}}(i) = S_m(i) - \kappa_{w\rightarrow m}(i) \cdot S_w(i), \tag{8.39}$$

with $\kappa_{w\rightarrow m}$ denoting the probability of falsely detecting a mutant instead of a wild type at position i, which is derived from the additional sample set comprising no mutations, only errors.

For the MIME approach, the single site effects are determined with the resampling procedure as described in Chapter 5.1.3, using the error corrected pairwise nucleotide counts. For the effects of the double mutants, the respective point estimate is used, also using the 2d data. The epistasis is computed according to Equation (8.28).
The error corrected frequencies are also the input for both DCA methods. First, the respective parameters h and g are calculated as described above for the "unbound" and "bound" fractions, and used for the determination of the probabilities for the sequences comprising one single mutation and pairs of mutations. These values are used to compute the effects akin to the original approach, according to Equation (8.22).
The resulting Kds and epistasis values for each position/–pairs estimated by each method are written into individual output files.

For each test setting, we conducted $N = 20$ sampling runs and evaluated the pooled results of all runs. We compare the absolute deviation of the relative log Kd: $\log(K_m^{\mathrm{estimate}}) - \log(K_m^{\mathrm{true}}))$, to have the positive and negative effects on the same scale $(-\infty < \log(K_m) < \infty)$. Thus, we have the same weight of deviation increasing and decreasing the effect. The same is done for the epistatic effect $\mathcal{E}$, with $\log(\mathcal{E}^{\mathrm{estimate}}) - \log(\mathcal{E}^{\mathrm{true}})$.
Furthermore, we evaluate the tendencies of the method to deviate depending on the respective value, to get an idea whether the evaluated method under– or overestimates certain effects, i.e. true value vs the respective deviation. Plotting the true value against

the estimated value will additionally help to assess how well the predictions match the truth and when they cause problems. For this, we also compute the coefficient of determination R^2, which is the proportion of the variance in the dependent variables that is predictable from the independent variables. In other words, R^2 measures the goodness of fit, to assess how well the values correspond to the model. The closer R^2 is to 1, the better are the predictions.

8.5 Benchmark Results

First, we investigated whether the considered sequence length L biases the prediction accuracy for the different methods. For this, we ascertained the deviation of the estimates from the true value without any single site effect or interaction, i.e. $p_{\text{kd}} = 0$ and $p_{\text{epistasis}} = 0$. Consequently, $K_m^{\text{true}}(i) = \mathcal{E}(ij) = 1 \ \forall \ i, j$, and since we consider the deviation of the log value, we assess the respective prediction deviation from zero.
We examined the outcome for different sequence lengths $L \in \{50, 75, 100\}$, respectively. For all evaluations we chose a mutation rate of $p_{\text{mut}} = 0.03$, and an error probability of $p_{\text{err}} = 0.001$. In Figure 8.2 it can be seen, that the deviations for single site effects $\log(K_m)$ predicted with MIME are unaffected by the sequence length, while the deviation for MF enhance slightly with longer sequences. A reason for this might be a growing sample size $\sum_{n=0}^{n_{\max}} \binom{L}{n}$ including more outliers. Surprisingly, the deviations from the single site effects with TAP disperse more extremely with growing length. We therefore chose for the further analysis length $L = 50$, to avoid the bias due to the sequence length for the performance assessment for the respective scenarios.

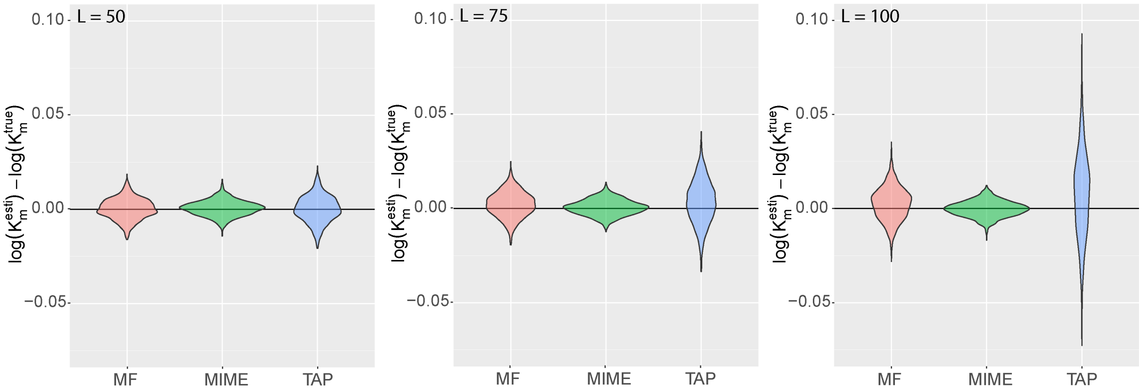


FIGURE 8.2: Violin plots showing the absolute deviation of the estimated single effects $\log(K_m)$ from the true value for methods naive MF, standard MIME and TAP. The other parameters for the data sampling are $p_{\text{kd}} = p_{\text{epistasis}} = 0$, $p_{\text{mut}} = 0.03$, and $p_{\text{err}} = 0.001$.

Notably, the deviation from effects of double mutants and zero epistasis are not considerably affected by the choice of L and are similar for all three methods. A representative example is shown in Figure 8.3.
We tested the methods with different parameter settings:
The test cases included the following combinations of p_{kd} and $p_{\text{epistasis}}$:

p_{kd}	0.2	0.8	0.0	0.0	0.5	0.8
$p_{\text{epistasis}}$	0.0	0.0	0.2	0.8	0.1	0.3

In summary, we have three different scenarios:

1. Only single site effects, no epistasis

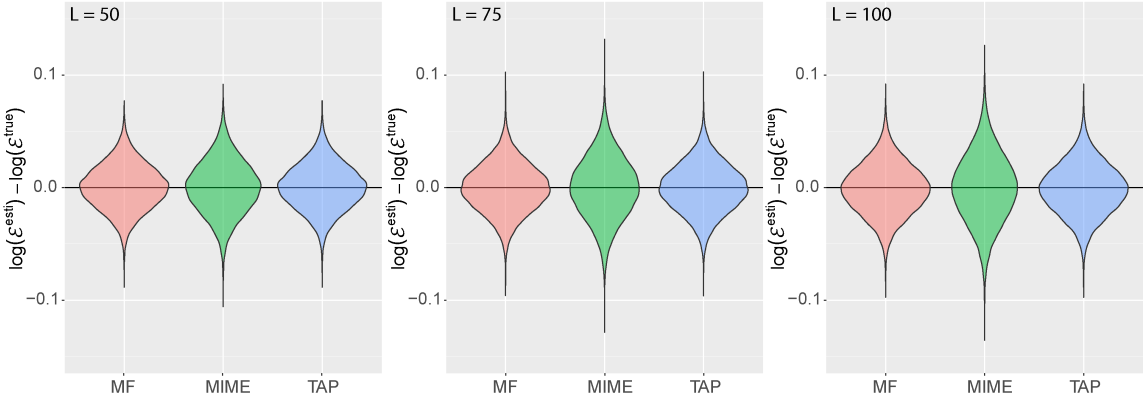


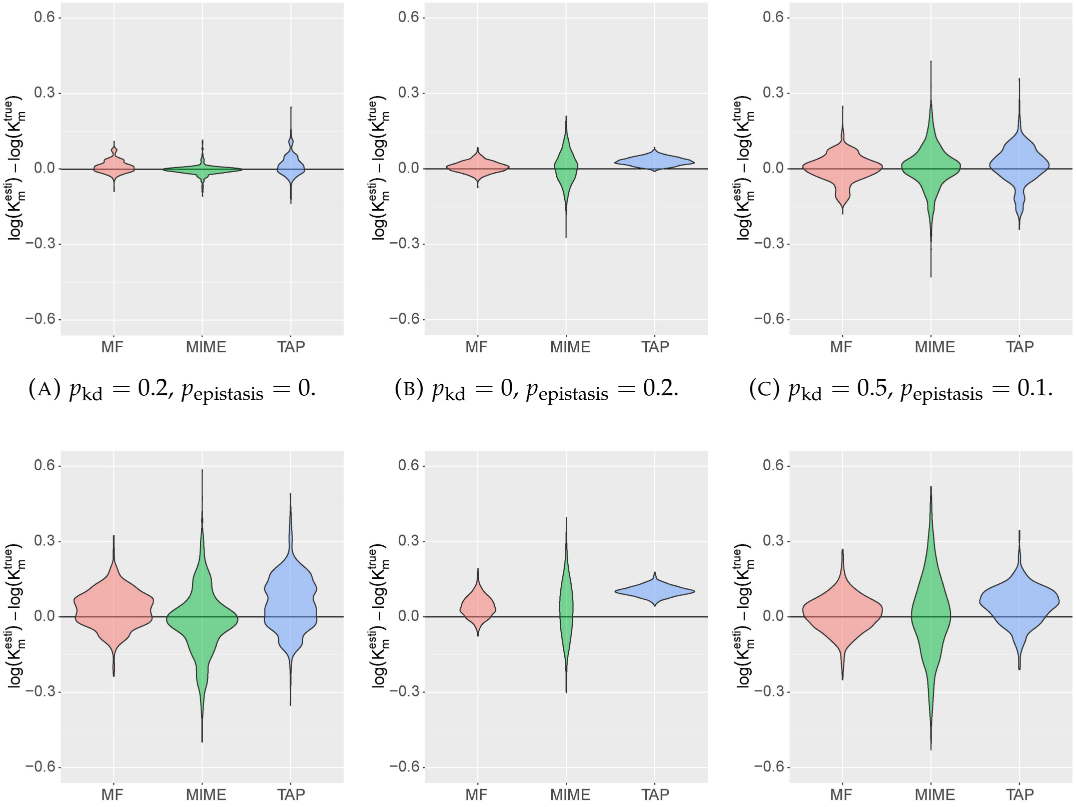
FIGURE 8.3: Violin plot showing the absolute deviation of the estimated log epistasis values from zero (the true value) for methods naive MF, standard MIME and TAP. The parameters for the data sampling are $p_{\text{kd}} = p_{\text{epistasis}} = 0$, $p_{\text{mut}} = 0.03$, and $p_{\text{err}} = 0.001$.

2. No single site effects, only epistasis
3. Single site effects and epistasis

In the first scenario, mutations may have an effect, however no interactions are present. We conducted the simulations with a low probability for each mutation to have an effect ($p_{\text{kd}} = 0.2$) and with a high probability ($p_{\text{kd}} = 0.8$).
In the second scenario, we constructed test cases where single mutations have no effect ($p_{\text{mut}} = 0$) unless they co-occur with certain other mutations ($p_{\text{epistasis}} = 0.2$ and 0.8).
These two cases are unlikely to be observed in reality. Their purpose is to analyse the respective prediction behaviour of each method in the presence of the particular effect, and facilitates the interpretation of the third scenario, where we have both single site effects and interactions, with $p_{\text{mut}} > p_{\text{epistasis}}$. We assume, that a large proportion of the sequence would affect the function of the RNA if mutated, including being directly involved in the respective process, and indirectly associated by providing the functional structure. The latter concerns sites in contact, i.e. direct interactions. Since the interactions constitute only a part of the functionally important sites, epistasis occurs less frequently. Again, we chose a scenario with a low probabilities for effects ($p_{\text{kd}} = 0.5$ and $p_{\text{epistasis}} = 0.1$) and with high probabilities ($p_{\text{kd}} = 0.8$ and $p_{\text{epistasis}} = 0.3$).

8.5.1 Estimation of Single Site Effects

In the following, we will analyse the accuracy of the methods to estimate relative effects of single mutations K_m.
In Figure 8.4, the plots show the absolute difference of the log estimate from the ground truth for each method. In general, we can observe that the qualitative characteristics of the distribution of differences are retained with varying effect probability (from top to bottom). However, the absolute prediction errors grow with the number of sites that confer an effect.
In all three scenarios, the error distribution for the standard MIME method centres around zero, but comprises a larger deviation from zero in comparison to the MF method when epistatic occur. On the contrary, the naive mean field approach shows a slight overestimation of the values when strong epistasis is involved, though the predictions deviate more when actual single site effects are present. The results for the TAP equation behaves similar to MF, yet, is prone to overestimation in all scenarios. Especially noticeable in the



(A) $p_{\text{kd}} = 0.2$, $p_{\text{epistasis}} = 0$.

(B) $p_{\text{kd}} = 0$, $p_{\text{epistasis}} = 0.2$.

(C) $p_{\text{kd}} = 0.5$, $p_{\text{epistasis}} = 0.1$.

(D) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0$.

(E) $p_{\text{kd}} = 0$, $p_{\text{epistasis}} = 0.8$.

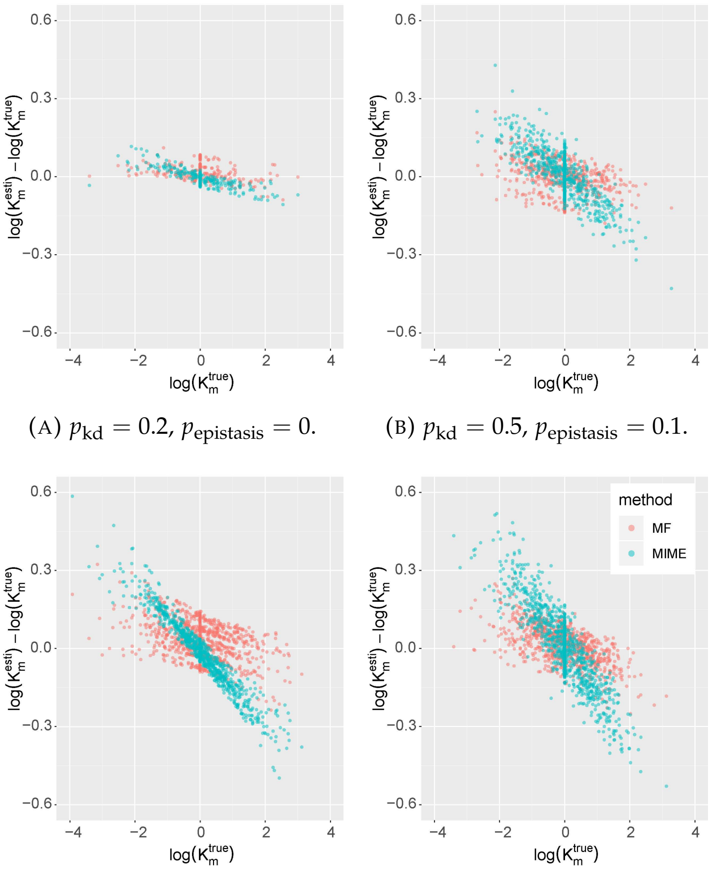
(F) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0.3$.

FIGURE 8.4: Violin plots showing the absolute difference of the prediction for single site effects ($\log(K_m^{\text{esti}})$) from the true values ($\log(K_m^{\text{true}})$) for MF, standard MIME and TAP.

scenario involving only epistasis, where all differences are above zero. This implies, that negative effects on function ($\log(K_m) > 0$) are overestimated, and at the same time positive effects on function($\log(K_m) < 0$) are underestimated. For the further analysis of the single site effect estimation, we therefore omit the results for TAP.

We investigate, if the absolute differences are connected to the strength of the true value. In Figure 8.5, we plot the deviations of the estimated single site effects $\log(K_m^{\textbf{esti}} - \log(K_m^{\textbf{true}})$ depending on the effect strength $\log(K_m^{\textbf{true}})$ for MF and standard MIME. Here, only scenarios with varying single site effects can be evaluated, i.e. the cases with $p_{\text{kd}} = 0$ are not included.
Both plots show increasing positive deviations with $log(K_m^{\text{true}}) \longrightarrow -\infty$ and growing negative deviations from zero with $log(K_m^{\text{true}}) \longrightarrow \infty$, i.e. an underestimation of the positive and negative effects on function. Interestingly, the plot for the results with the standard MIME approach exhibits a steeper slope in contrast to MF, i.e. more deviation with larger effects. The MF approach, on the contrary, shows a more widespread distribution of differences, regardless of the true effect, particularly for the case without epistasis in Figure 8.5 (C).

To evaluate, whether the deviations lead to under– or overestimation of effects, we examine the estimated value against the true value, plotted in Figure 8.6. In order to assess how well the estimates fit to the given model, i.e. the ground truth, we additionally compute the respective coefficient of determination R^2. The plots and the coefficients of determination (all coefficients R^2 are > 99%) already suggest that the predictions of K_m



(A) $p_{\text{kd}} = 0.2$, $p_{\text{epistasis}} = 0$.

(B) $p_{\text{kd}} = 0.5$, $p_{\text{epistasis}} = 0.1$.

(C) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0$.

(D) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0.3$.

FIGURE 8.5: Dotplot of the true single site effect values $\log(K_m^{\text{true}})$ versus the absolute deviation of the estimates from the truth $\log(K_m^{\text{esti}}) - \log(K_m^{\text{true}})$ predicted with MF and standard MIME.

are quite accurate for both methods. However, as we apprehended in the beginning, the predictions for standard MIME tend to underestimate positive and negative effects of mutations, when pairwise interactions occur in the data (Figure 8.6 (D)). For naive mean field, the predictions are highly precise, regardless of the involved effects.

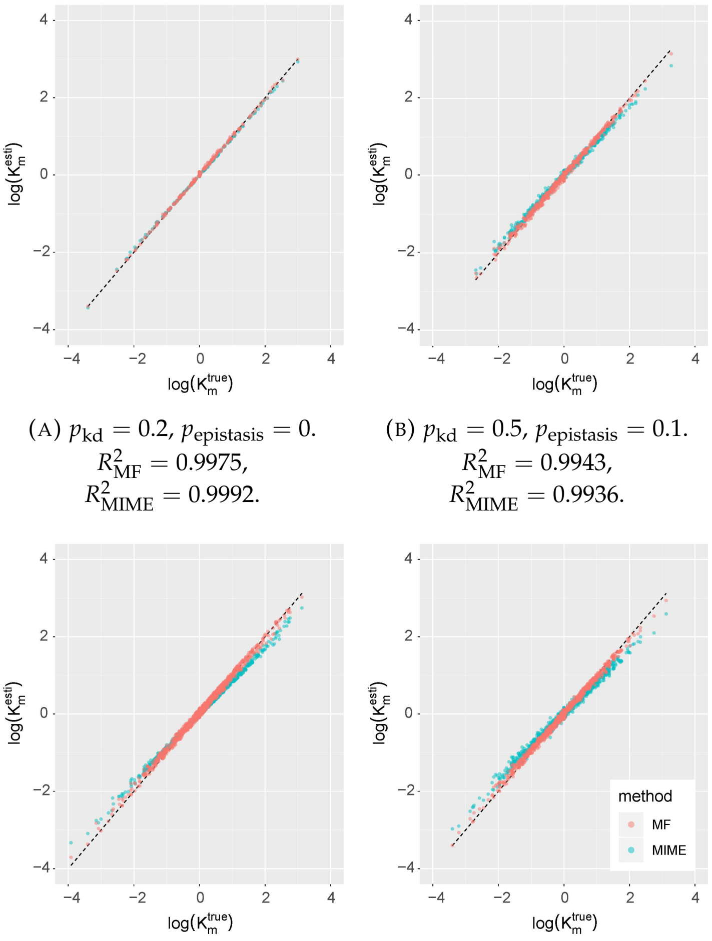
As shown in Equation (8.22), the Hamiltonians for the bound and unbound sample have to be constructed in order to compute the single site effects. For the sequence with one mutation m at position i, it is given by

$$K_m^{\text{MF}}(i) = \frac{e^{-\mathcal{H}^u(\sigma^{m_i})}}{e^{-\mathcal{H}^b(\sigma^{m_i})}} \tag{8.40}$$

$$= \frac{e^{h_i^u(m_i)}}{e^{h_i^b(m_i)}}. \tag{8.41}$$

The remaining parameters are set to zero, since the rest of the sequence is in wild type configuration, and epistasis is only given if at least two mutations occur. In the naive mean field approach, parameters h^{MF} are calculated according to Equation (8.24):

$$h_i^{\text{MF}}(A) = \ln\left(\frac{f_i(A)}{f_i(\text{wt})}\right) - \sum_{j \neq i} \sum_{B \neq \text{wt}} g_{ij}(A,B) f_i(B).$$



(A) $p_{\text{kd}} = 0.2$, $p_{\text{epistasis}} = 0$. $R^2_{\text{MF}} = 0.9975$, $R^2_{\text{MIME}} = 0.9992$.

(B) $p_{\text{kd}} = 0.5$, $p_{\text{epistasis}} = 0.1$. $R^2_{\text{MF}} = 0.9943$, $R^2_{\text{MIME}} = 0.9936$.

(C) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0$. $R^2_{\text{MF}} = 0.9948$, $R^2_{\text{MIME}} = 0.9982$.

(D) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0.3$. $R^2_{\text{MF}} = 0.9954$, $R^2_{\text{MIME}} = 0.9902$.

FIGURE 8.6: Dotplot of the true single site effect values $\log(K_m^{\text{true}})$ versus the estimates $\log(K_m^{\text{esti}})$ of MF and standard MIME. The black dashed diagonal shows the identity.

We noticed in the data, that for all scenarios the inferred couplings terms g were almost always between -1 and 0, whereas parameters h have values around -3. Furthermore, the frequency of the mutation $f_i(A)$ is comparatively small, due to low mutation rate. Hence, the second term of the equation, although summing up over all L-1 positions, is very small. The first part, with $f_i(\text{wt}) > f_i(A)$, dominates the equation. This leads to a rough approximation of

$$h_i(A) \approx \log\left(\frac{f_i(A)}{f_i(\text{wt})}\right), \tag{8.42}$$

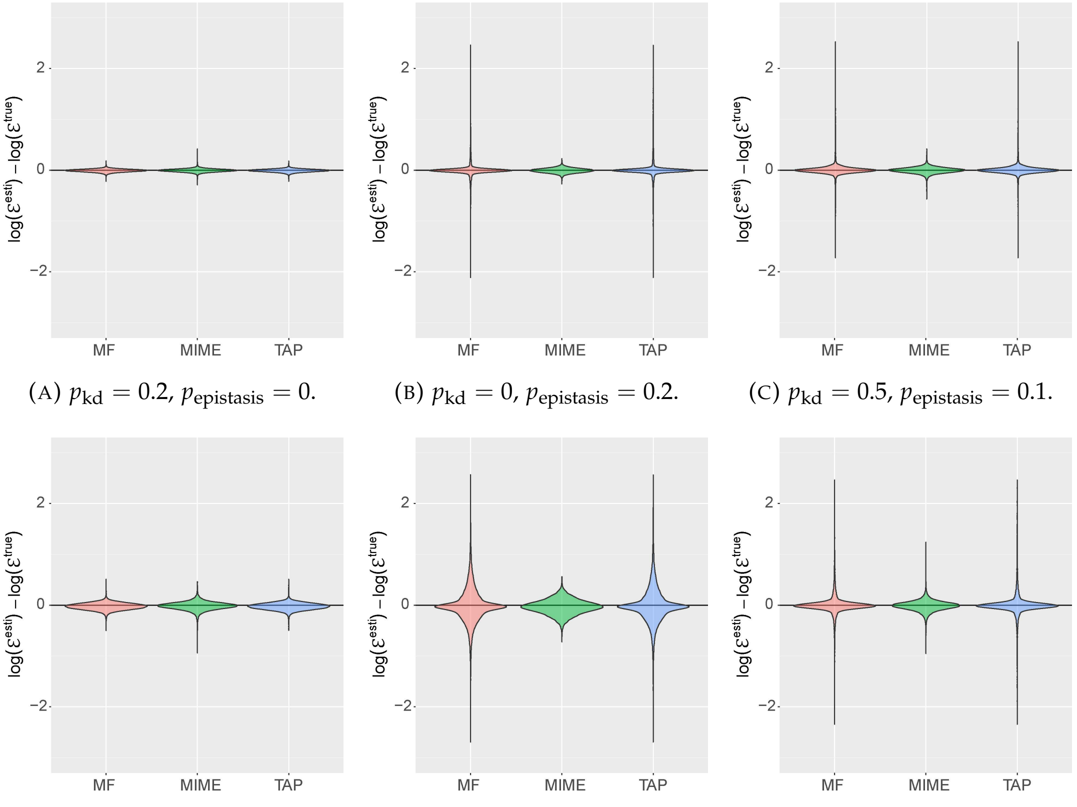
and hence

$$\begin{aligned} K_m^{\text{MF}}(i) &\approx \frac{e^{h_i^u(m_i)}}{e^{h_i^u(m_i)}} &(8.43)\\ &= \frac{f_i^u(m_i)}{f_i^u(\text{wt})} \cdot \frac{f_i^b(\text{wt})}{f_i^b(m_i)}, &(8.44)\end{aligned}$$

which corresponds to the Equation 8.18, i.e. the standard MIME evaluation for single site effects. Although the second term in Equation (8.24) for h is relatively small, it has an amending effect on the prediction.

8.5.2 Estimation of Epistatic Effects

We performed the analogous analysis as seen above for the epistasis estimates.
A first impression of the performance of the three methods MF, MIME and TAP is given in the violin plots of the absolute deviation of the estimates $\log(\mathcal{E}^{\text{esti}})$ from the true epistasis values $\log(\mathcal{E}^{\text{true}})$ in Figure 8.7.



(A) $p_{\text{kd}} = 0.2$, $p_{\text{epistasis}} = 0$.

(B) $p_{\text{kd}} = 0$, $p_{\text{epistasis}} = 0.2$.

(C) $p_{\text{kd}} = 0.5$, $p_{\text{epistasis}} = 0.1$.

(D) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0$.

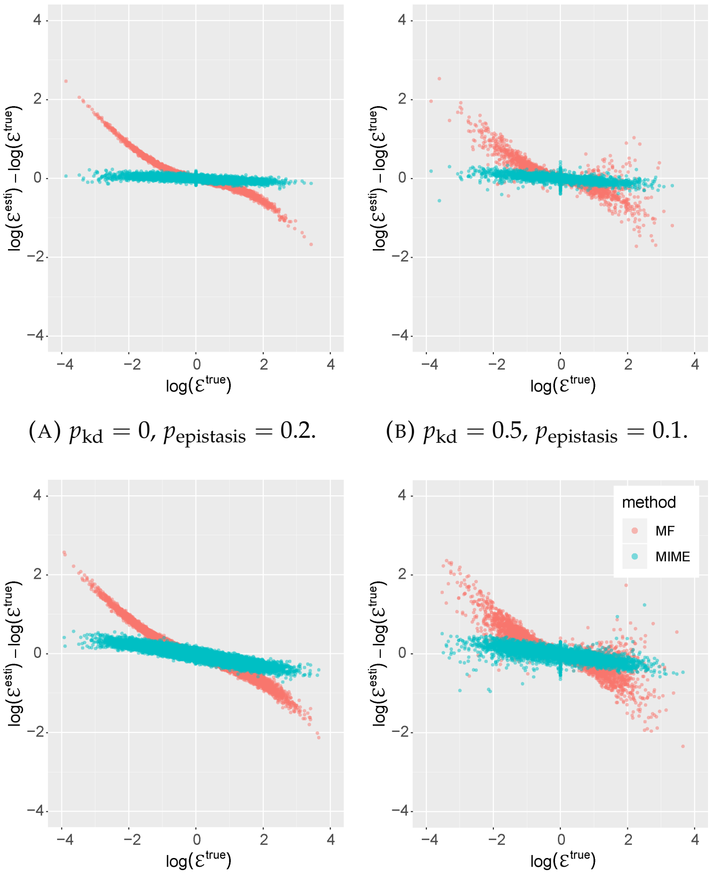
(E) $p_{\text{kd}} = 0$, $p_{\text{epistasis}} = 0.8$.

(F) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0.3$.

FIGURE 8.7: Violin plots showing the absolute difference of the predictions for epistasis ($\mathcal{E}_{ij}$) from the true values (both log) for MF, standard MIME and TAP.

We can notice that in all scenarios the deviations from the ground truth centre around zero. However, the mean field approaches are substantially prone to false predictions when epistasis is involved with a slight tendency of underestimation, whereas the standard MIME method shows elevated biases with more frequent occurrence of single mutation effects.
For the evaluations regarding epistasis values, all outcomes for MF and TAP looked highly similar. We therefore omit the results for TAP in the following plots for clearer comparison with the standard MIME method.

The relation of the estimate deviation to the true value are depicted in Figure 8.8. Here, we evaluate only the scenarios with $p_{\text{epistasis}} > 0$. Very unexpectedly, the predictions with MF are predominantly underestimating the interactions, resulting in massive false negatives, regardless of the amount of epistatic effects in the data. The plots with no single



(A) $p_{\mathrm{kd}} = 0$, $p_{\mathrm{epistasis}} = 0.2$.

(B) $p_{\mathrm{kd}} = 0.5$, $p_{\mathrm{epistasis}} = 0.1$.

(C) $p_{\mathrm{kd}} = 0$, $p_{\mathrm{epistasis}} = 0.8$.

(D) $p_{\mathrm{kd}} = 0.8$, $p_{\mathrm{epistasis}} = 0.3$.

FIGURE 8.8: Dotplot of the true single site effect values $\log(\mathcal{E}^{\mathrm{true}})$ versus the absolute deviation of the estimates from the truth $\log(\mathcal{E}^{\mathrm{esti}}) - \log(\mathcal{E}^{\mathrm{true}})$ of MF and standard MIME.

site effects show a clean curve, suggesting a wrong data representation for the evaluation of the epistasis is used here. With presence of single site effects, however, the deviations blur out.

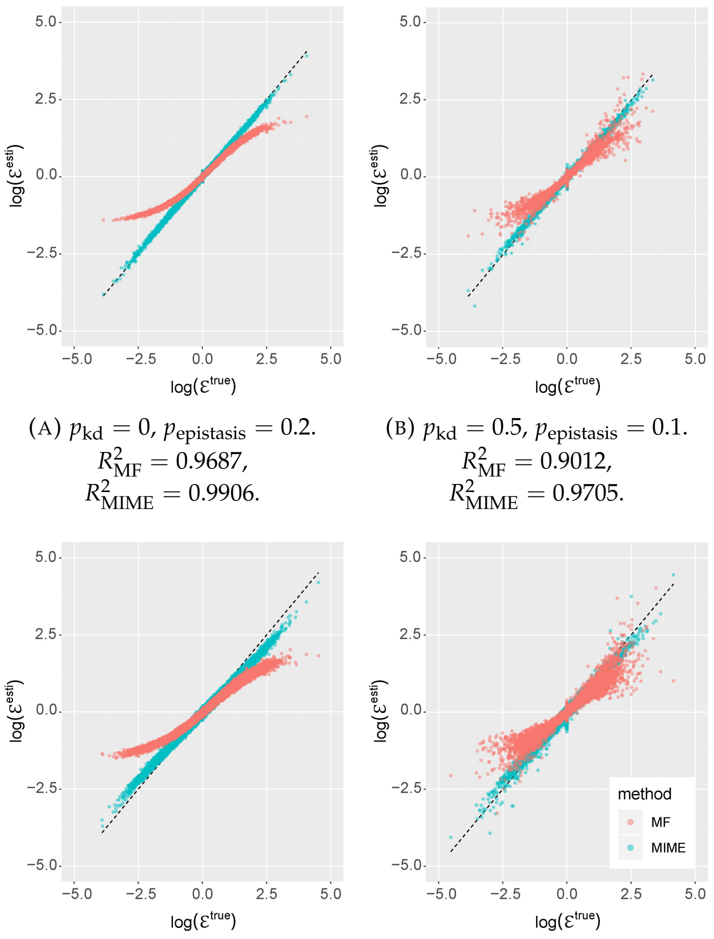
For the standard MIME method, we can notice a quite precise determination of epistatic effects, if p_{kd} is zero . With occurring single site effects, the accuracy slightly drops with large values.

In Figure 8.9 we compare the true versus the estimated *log* value and calculate the corresponding coefficient of determination R^2. As we saw already in the plots before, the standard MIME method predicts the epistatic effects accurately with certain biases with increasing amount of single mutation effects.

In this data representation, for the case without single site effects, the estimated value predicted with MF reminds of a root- or log-like function. We plotted two different functions on the dotplot in Figure 8.10(A), which could presumably reflect the function behind the log MF estimates :

- $f(x) = sign(x)\sqrt{|x|}$
- $f(x) = sign(x)\log(|x| + 1)$

The logarithmic function shows the closest similarity to our data, especially for negative and small positive values. Contrarily, the root function resembles the larger positive values quite well, however strongly deviates for the rest. We transformed the MF estimates



(A) $p_{\text{kd}} = 0$, $p_{\text{epistasis}} = 0.2$. $R^2_{\text{MF}} = 0.9687$, $R^2_{\text{MIME}} = 0.9906$.

(B) $p_{\text{kd}} = 0.5$, $p_{\text{epistasis}} = 0.1$. $R^2_{\text{MF}} = 0.9012$, $R^2_{\text{MIME}} = 0.9705$.

(C) $p_{\text{kd}} = 0$, $p_{\text{epistasis}} = 0.8$. $R^2_{\text{MF}} = 0.9801$, $R^2_{\text{MIME}} = 0.9922$.

(D) $p_{\text{kd}} = 0.8$, $p_{\text{epistasis}} = 0.3$. $R^2_{\text{MF}} = 0.913$, $R^2_{\text{MIME}} = 0.9718$

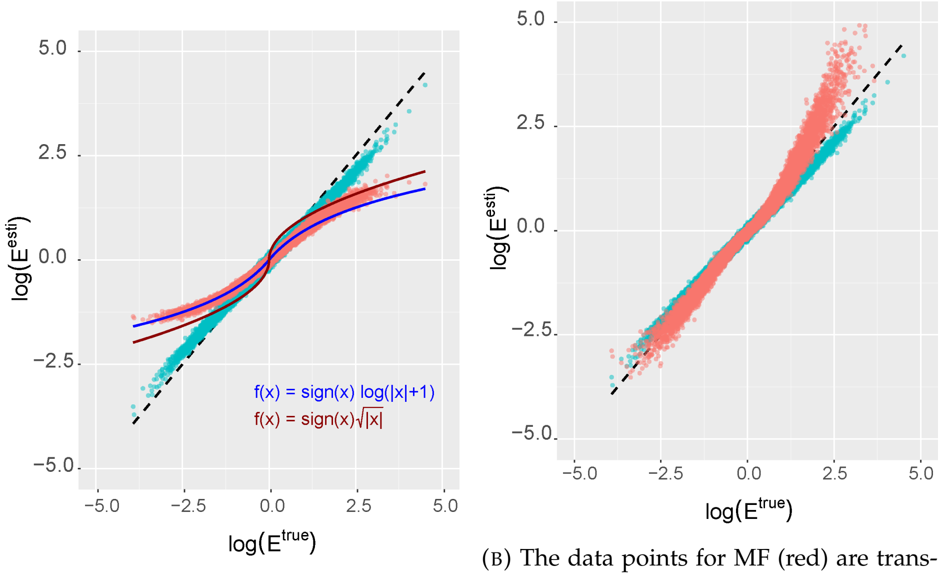
FIGURE 8.9: Dotplot of the true epistasis values $\log(\mathcal{E}^{\text{true}})$ versus the estimates $\log(\mathcal{E}^{\text{esti}})$ of MF and standard MIME. The black dashed diagonal shows the identity.

with the inverse function of the log function, resulting in a remarkably amended accordance with the true values, seen in Figure 8.10(B). As expected, the largest proportion matches quite well, with increasing deviation towards bigger positive values.
Hence, the chosen scale as derived in the beginning is not the appropriate choice for the assessment of epistasis. If we look again at the determination of epistasic effects $\mathcal{E}$ with MF, evaluated in the MIME based approach, we have

$$\mathcal{E}^{\text{MF}}(i,j) = e^{-E^{\text{MF}}(i,j)} = \frac{K_{m,m}(i,j)}{K_{m,w}(i,j) \cdot K_{w,m}(i,j)} \tag{8.45}$$

$$= \frac{e^{-\mathcal{H}^u(\sigma^{m_i m_j})}}{e^{-\mathcal{H}^b(\sigma^{m_i m_j})}} \cdot \frac{e^{-\mathcal{H}^b(\sigma^{m_i})}e^{-\mathcal{H}^b(\sigma^{m_j})}}{e^{-\mathcal{H}^u(\sigma^{m_i})}e^{-\mathcal{H}^u(\sigma^{m_j})}} \tag{8.46}$$

$$= \frac{e^{h_i^u(m_i)+h_j^u(m_j)+g_{ij}^u(m_i,m_j)}}{e^{h_i^b(m_i)+h_j^b(m_j)+g_{ij}^b(m_i,m_j)}} \cdot \frac{e^{h_i^b(m_i)}e^{h_j^b(m_j}}{e^{h_i^u(m_i)}e^{h_j^u(m_j})} \tag{8.47}$$



(A) Plotted curves for a logarithmic function (blue) and a square root function (red).

(B) The data points for MF (red) are transformed with the inverse of the logarithmic function.

FIGURE 8.10: True epistasis values $\log(\mathcal{E}^{\text{true}})$ versus the estimates $\log(\mathcal{E}^{\text{esti}})$ of MF and standard MIME for the scenario with $p_{\text{kd}} = 0$ and $p_{\text{epistasis}} = 0.8$.

where the parameter h are cancelled out and the epistatic effect is derived by

$$\mathcal{E}^{\text{MF}}(i,j) = \frac{e^{g^{u}_{ij}(m_i,m_j)}}{e^{g^{b}_{ij}(m_i,m_j)}}. \tag{8.48}$$

In this calculation, the epistasis is given by the relation of the exponential direct coupling terms of the unbound vs the bound samples. Nevertheless, we have seen, that this measure does not display the true epistatic effect correctly. The vast majority of the inferred couplings for all scenarios lies between 0 and -1, regardless of the sampled effect. This leads to an underestimation of the epistasis.

Whether the chosen methodology for the derivation of epistatic effects is unsuitable, and an appropriate score can be found, or the coupling inference (e.g. gauge fixing, regularisation) fails in the first place, needs to be explored in further analysis.

8.6 Discussion

The classical MIME approach is able to detect functionally important sites within RNA, by quantifying effects of mutations on that function relative to the wild type at single site resolution. Since these estimates may be confounded by effects of epistatic interactions within the sequence, our aim was to improve the predictions by combining the MIME approach with DCA related methods.

In our benchmark, we generated data sets with different parameters, i.e. the probabilities for the sequence to comprise single site effects and epistasis. The resulting pools of "bound" and "unbound" sequences served as input for the mentioned methods to estimate the sampled effects. We compared the deviation from the ground truth of standard MIME with the results of the combined approach using sequence probabilities acquired with the naive mean field approach and the TAP equation.

TAP yielded either similar or more inaccurate results compared to the mean field method.

For this reason, we excluded TAP for the more detailed analysis.
The estimation of effects of single mutations could indeed be improved with MF. Although standard MIME already obtains very good results, MF is able to reduce the biases for large effects.
For the inference of epistatic effects, however, both DCA methods yielded unsatisfying results at first sight. We found a logarithmic relationship of the estimated values $\mathcal{E}^{\text{esti}}$ to the initially sampled true value $\mathcal{E}^{\text{true}}$. However, these results are of suggestive nature and require further investigation.

The generated test cases were sufficient for test purposes, to confirm that the integration of a global sequence model can improve MIME predictions. In fact, we were able to ascertain an increase in precision, which motivates us to further pursue this approach. In subsequent studies, we are going to simulate the binding competition of longer sequences creating an enlarged sequence space. More importantly, we are going to refine the simulation of the sequencing procedure by including the random fragmentation. Here, we sampled and evaluated the full-length sequence, in contrast to next generation sequencing generating millions of short reads.
As soon as the approach is optimised, the aim is to undertake predictions with real sequencing data from MIME experiments.

Chapter 9

Discussion

In this work, we addressed the determination of evolutionary constraints with focus on the functional and structural constraints in RNA, and particular application to HIV genomic RNA.

We started with an introduction of the biology of HIV detailing all steps of the viral life cycle. Despite the multitude of complex, interlinked processes to produce new viable virions, no cure has been found yet that eliminates a viral infection completely. Nevertheless, effective treatment regimen exist, which disrupt essential processes of the replication cycle by blocking important enzymes and receptors, or by impeding reverse transcription with nucleoside analogues [5]. They inhibit the viral productivity and permit a nearly unimpaired quality of living for infected people. Still, in few cases, the virus is able to escape the treatment through (multi-)drug resistance mutations. Therefore, new treatment strategies are required.

Non-coding RNAs regulate virtually all cellular processes [179], and we have seen that the HIV gRNA plays a pivotal role in multiple mechanisms of the HIV life cycle. This circumstance suggests promising opportunities to expand the existing repertoire of antiviral therapeutics [180–183]. Already in early studies, the hairpin structure of the HIV transactivation response (TAR), facilitating the transcription of the viral DNA, was identified as a promising target for drug-like compounds [184–186], and could be confirmed and further developed in subsequent work [165, 187, 188].

One challenge for drug development is the determination of RNA motifs and structures which are related to essential functions. Moreover, molecules have to be developed which are eligible to bind to that functional element and simultaneously exhibit sufficient drug-likeliness [7]. This means, the considered molecule has to meet certain criteria to be approved as a potential drug. This involves several physicochemical properties, including lack of toxicity, cell– and tissue permeability, high affinity to bind the target, and high selectivity and specificity [7]. The latter imply a detailed knowledge about motifs in granular resolution, in order to specify and assess the structural sophistication of the potential target.

However, the regulatory landscape in the genomic RNA of HIV, together with the mapping of corresponding structures, requires substantial exploration as coherent mechanisms and function-structure relations are still incompletely understood [155].

Regions comprising functionally important elements constrain the evolution of the gRNA. Patterns and motifs prevail in the genetic pool of the population due to high fitness, since only a functional genome permits a viable organism.

We presented various methods to determine these evolutionary constraints. The quite recent direct coupling analysis (DCA) framework borrows approaches from statistical physics, implementing a maximum-entropy model to describe the information in biological sequence data within an MSA. The model comprises parameters, which need to be fit to derive a probability distribution of the sequence space. In most applications,

the coupling terms were used to infer direct pairwise interactions [9, 128, 129, 132, 141]. These direct couplings can be used to qualitatively nominate residues in close proximity and to conclude secondary– and tertiary structures. However, the data is only composed of viable organisms. Many generations of functional selection led to the accumulation of particularly strong and predominant variants, whereas intermediate variants are missing. Additionally, the MSA represents the variants with the highest total fitness corresponding to selective constraints, not necessarily in relation to a certain trait or function: It may happen that a genotype which would be more advantageous for a modular function is not or less observed, because the same genotype might be detrimental for another essential function or trait, and thus for the total fitness. Potentially valuable insights about structure-function relations might be masked.
Furthermore, these inferred structural information may represent a conglomerate of multiple different structures: Functional RNAs are compact information carriers exhibiting a high density of information. Often, gene segments encoding certain proteins overlap with regulatory regions, which already makes it difficult to detect functionally important sites. Moreover, RNAs can be multifunctional, i.e. the same regulatory region may form diverse structure conformations for different tasks [58, 63].

In MIME [8, 11], on the contrary, a single evolutionary step is executed to define mutational effects on a particular function. Copies of a reference sequence are randomly mutated and segregate into functionally selected and unselected sequence pools. Mutations which disrupt the function are expected to be less present in the selected pool relative to the allocation ratio of the reference sequence. Hence, the method allows to quantify relative effects that each mutation has on the function. Beyond the single site effects, MIME data can be used to estimate epistasis, including compensatory mutations: e.g. two individual mutations disrupt the function, but the co-occurrence of the mutations restores it, for instance by building a basepair involved in the structure.
MIME has been conducted in *in vitro* experiments to identify motifs within the $5'$ UTR of the HIV genome explicitly crucial for Pr55$^{\mathrm{Gag}}$ binding [8]. We adapted the MIME framework for *in cell* experiments [12], analysing two phases of the HIV replication cycle. We were able to find three motifs involved in gRNA production and two signals which are important for gRNA packaging. One of the motifs, mapped to the $5'$ PolyA structure, even had contradicting influence on both processes, which displays once more the complex and opaque correlations of the entire HIV replication process.

We were able to show that MIME can be used for functional structure mapping not only for fine-grained functions, but also for more sophisticated procedures in complex cell environments. Yet, we have seen, that the quantitative estimations might deviate when epsitatic interactions are involved. For the inference of mutational effects at a certain site, local nucleotide frequencies are used, which means effects of coupled mutation sites might intrude the predicted outcome. We attempted to facilitate ideas from DCA, the incorporation of a global model for the inference of sequence probabilities instead of frequencies, in order to improve the predictions. For the inference of the model parameters, we implemented the mean field approximation and its extension, the TAP equation. We generated test data for various scenarios comprising different portions of positions and pairs of positions which have an effect on the function when mutated. We compared the deviation from the ground truth of the MIME predictions and the results of the DCA methods. Indeed, we could achieve slight improvements for the determination of single site effects, especially for large values. However, the estimation for epistasis with DCA methods revealed unexpected behaviour. As we already discussed in the corresponding chapter, the derived scoring for the epistasis is not adequate, at least in combination with the low order of magnitude of the inferred coupling terms. In further investigations we would like to ascertain, if a more suitable measure exists, in order to harness the relationship of couplings of the selected and unselected fractions to correctly quantify epistasis. Another issue might be the choice of gauge fixing. We learnt, that the same number of independent parameters and constraints of the model can be achieved with diverse methods of gauge fixing. In our case, we set all parameters for one of the symbols to zero, which is convenient for the inference of sequence probabilities relative to the wild type sequence. However, some scoring schemes using inferred couplings in existing applications are not necessarily gauge invariant [132, 189]. We would like to inspect in more detail, if parameter inference with zero-sum gauge fixing leads to similar results, or if we are able to improve the prediction.

For future work, we consider *in silico* evolution experiments, akin to the directed enzyme evolution *in vitro* by Frances H. Arnold et al. [100, 190], which was honoured with a Nobel Prize in Chemistry in 2018. The idea here is to set up a cycle of directed selection: sequences are diversified (error-prone PCR, recombination, mutagenesis), selected by fitness, where the unselected sequences are discarded, and the selected pool is again diversified and the cycle starts over again, until a certain functional optimum is reached.
For our benchmark, we set up a similar approach, for the purpose of simulating a MIME experiment: We randomly mutate a reference sequence, sample their effect and epistatic effects on function and simulate the functional process. This functional process can be modelled in arbitrary complexity or modularity. We could accomplish several generations of the competitive experiment, always applied to the selected sequence fraction of the current generation, until we reach a somewhat stable population optimised for the simulated function. It would be interesting to evaluate and possibly optimise the performance of various methods for the determination of evolutionary constraints. The generated data underlies a genuine ground truth and we would be able to ascertain the behaviour of various algorithms under certain conditions.

An additional prospective aim is further detailed mapping of various RNA structures on specific functions in HIV or other organisms. Another approach to infer hints about structural information, which was not introduced in detail in this work, is the so called chemical Probing or Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE) experiment [191, 192]. The method allows to determine residues in folded RNA sequences, which are accessible to chemical probing, and as a consequence presumably unpaired. On the contrary, the knowledge of which positions are interacting is lacking here, and complicates a subsequent structure inference. As we have discussed earlier, RNAs often exhibit rich landscapes with multiple structures. It still remains a challenge to disentangle the true structure ensembles and associate the correct biological function. Combining the structural constraints of the chemical probing with the functional mapping and epistasis information of the MIME approach may lead to accurate assignment of structure-function relations. The results may provide constraints for 3D structure prediction and RNA targeting drug design.

Zusammenfassung

Seit ihrem Beginn vor etwa 30 Jahren, konnte die weltweite Epidemie von HIV noch immer nicht aufgehalten werden und kostete seither mehrere Millionen Menschenleben. Der Virus greift das Immunsystem seines Wirtes an. Bleibt eine HIV Infektion unbehandelt, führt sie zu schweren opportunistischen Erkrankungen, bis hin zu AIDS. In dieser Phase ist das Immunsystem bereits schwer geschädigt und führt rasch zum Tode der infizierten Person. Glücklicherweise ist die Zahl der jährlichen Neuinfektionen aufgrund umfassender Aufklärungskampagnen und effektiver Behandlungsmöglichkeiten rückläufig. Zwar wurde bis heute keine wirksame Impfung oder vollkommene Heilung entdeckt, jedoch können infizierte Personen mit Hilfe von antiretroviralen Therapien eine nahezu durchschnittliche Lebenserwartung erreichen. Allerdings ist HIV, bedingt durch hohe Mutationsrate und schnellem Replikationszyklus, ein besonders anpassungsfähiger Virus, was ihm ermöglicht, Resistenzen gegen Medikamente auszubilden. Die Situation wird besonders kritisch, wenn sich Multiresistenzen entwickeln. Die Entdeckung neuer Behandlungsstrategien ist daher von dringender Notwendigkeit.

RNAs enthalten neben Genregionen, in denen essentielle Proteine kodiert sind, eine Vielzahl regulatorischer Elemente und funktionaler Strukturen, die die meisten biologischen Prozesse beeinflussen. Eine Veränderung dieser Regionen durch Mutation, würde unter Umständen zur Folge haben, dass der Organismus nicht überlebt, da lebenswichtige Funktionen nicht ausgeführt werden können. Funtionsrelevante Motive und Strukturen bilden sogenannte Evolutionary Constraints, also funktions- und strukturabhängige Hemmung von Evolution. Das macht sie zu vielversprecheneden Angriffspunkten für neue Wirkstoffe. Dies setzt detailliertes Wissen über die regulatorischen Mechanismen und notwendigen Strukturen vorraus, die an lebenswichtigen Prozessen des HIV Zyklus beteiligt sind. Jedoch sind die genauen Abläufe und Zusammenhänge dieser Mechanismen nicht ausgiebig erforscht.

In der vorliegenden Arbeit behandeln wir Techniken zur Bestimmung von Evolutionary Constraints, im Kontext von genomischer RNA in HIV. Wir erläutern unterschiedliche Herangehensweisen, um qualitative und quantitative Rückschlüsse auf Evolutionary Constraints zu ziehen. Hauptaugenmerk legen wir dabei auf die beiden Methoden Direct Coupling Analysis (DCA) und Mutational Interference Mapping Experiment (MIME). Basierend auf Letzterer, präsentieren wir Software zur Vorhersage von funktionalen Elementen in RNA. Des Weiteren haben wir das MIME Framework für *in cellulo* Experimente adaptiert. Dabei konnten wir regulatorische Motive im $5'$ untranslatierten Bereich des HIV-1 Genoms detektieren, welche sowohl wichtig für die Produktion viraler RNA in Zellen, als auch für die Integration des viralen Genoms in neu entstehende Viren sind. Zum Schluss haben wir das Ziel verfolgt, die Vorhersage funktionaler Regionen in MIME zu verbessern, in dem wir an DCA angelehnte Methoden einbeziehen. Dazu führen wir einen Benchmark mit verschiedenen Szenarien durch. Wir generieren Daten mit unterschiedlich vielen Mutationseffekten, welche die Funktion der RNA einschränken, sowie paarweiser Epistasis (Evolutionary Constraints). Tatächlich konnten wir für einige Fälle Verbesserungen feststellen. Dies sind jedoch vorerst vorläufige Ergebnisse, die uns aber bestärken, diesen Ansatz weiter zu verfolgen.

Bibliography

[1] UNAIDS. *Epidemic transition metrics [Accessed: 28 December 2018]*. `https://AIDSinfo.unaids.org`. 2018.

[2] Kaveh Pouran Yousef et al. "Inferring HIV-1 transmission dynamics in Germany from recently transmitted viruses". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 73.3 (2016), pp. 356–363.

[3] Sulav Duwal et al. "Optimal treatment strategies in the context of 'treatment for prevention'against HIV-1 in resource-poor settings". In: *PLoS computational biology* 11.4 (2015), e1004200.

[4] Xiaowei Jiang et al. "Characterizing the diverse mutational pathways associated with R5-tropic maraviroc resistance: HIV-1 that uses the drug-bound CCR5 coreceptor". In: *Journal of virology* 89.22 (2015), pp. 11457–11472.

[5] AIDSinfo. *HIV Treatment: The Basics [Accessed: 12 January 2018]*. `https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/21/51/hiv-treatment--the-basics`. Nov. 2018.

[6] Harry LA Janssen et al. "Treatment of HCV infection by targeting microRNA". In: *New England Journal of Medicine* 368.18 (2013), pp. 1685–1694.

[7] Katherine Deigan Warner, Christine E Hajdin, and Kevin M Weeks. "Principles for targeting RNA with drug-like small molecules". In: *Nature Reviews Drug Discovery* 17.8 (2018), p. 547.

[8] Redmond P Smyth et al. "Mutational interference mapping experiment (MIME) for studying RNA structure and function". In: *Nature methods* 12.9 (2015), p. 866.

[9] Faruck Morcos et al. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". In: *Proceedings of the National Academy of Sciences* 108.49 (2011), E1293–E1301.

[10] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. "Inverse statistical problems: from the inverse Ising problem to data science". In: *Advances in Physics* 66.3 (2017), pp. 197–261.

[11] Maureen R Smith et al. "MIMEAnTo: profiling functional RNA in mutational interference mapping experiments". In: *Bioinformatics* 32.21 (2016), pp. 3369–3370.

[12] Redmond P Smyth et al. "In cell mutational interference mapping experiment (in cell MIME) identifies the 5′ polyadenylation signal as a dual regulator of HIV-1 genomic RNA production and packaging". In: *Nucleic acids research* 46.9 (2018), e57–e57.

[13] Paul M Sharp and Beatrice H Hahn. "Origins of HIV and the AIDS pandemic". In: *Cold Spring Harbor perspectives in medicine* 1.1 (2011), a006841.

[14] Feng Gao et al. "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes". In: *Nature* 397.6718 (1999), p. 436.

[15] Elizabeth Bailes et al. "Hybrid origin of SIV in chimpanzees". In: *Science* 300.5626 (2003), pp. 1713–1713.

[16] Avert. *History of HIV & AIDS overview [Accessed: 27 December 2018]*. `https://www.avert.org/professionals/history-hiv-aids/overview`. Jan. 2017.

[17] Nuno R Faria et al. "The early spread and epidemic ignition of HIV-1 in human populations". In: *science* 346.6205 (2014), pp. 56–61.

[18] Centers for Disease Control (CDC et al. "A cluster of Kaposi's sarcoma and Pneumocystis carinii pneumonia among homosexual male residents of Los Angeles and Orange Counties, California." In: *MMWR. Morbidity and mortality weekly report* 31.23 (1982), p. 305.

[19] Henry Masur et al. "An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction". In: *New England Journal of Medicine* 305.24 (1981), pp. 1431–1438.

[20] Centers for Disease Control (CDC et al. "Update on acquired immune deficiency syndrome (AIDS)–United States." In: *MMWR. Morbidity and mortality weekly report* 31.37 (1982), p. 507.

[21] Richard E Chaisson et al. "Cocaine use and HIV infection in intravenous drug users in San Francisco". In: *Jama* 261.4 (1989), pp. 561–565.

[22] Ctr's for Disease Control, Prevention (CDC), and United States of America. "Update: human immunodeficiency virus infections in health-care workers exposed to blood of infected patients". In: *Morbidity and Mortality Weekly Report* 36.19 (1987), pp. 285–289.

[23] JamesJ Goedert et al. "Mother-to-infant transmission of human immunodeficiency virus type 1: association with prematurity or low anti-gp120". In: *The Lancet* 334.8676 (1989), pp. 1351–1354.

[24] Françoise Barré-Sinoussi et al. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)". In: *Science* 220.4599 (1983), pp. 868–871.

[25] Robert C Gallo et al. "Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS". In: *science* 224.4648 (1984), pp. 500–503.

[26] Jean L Marx. "Strong new candidate for AIDS agent; a newly discovered member of the human T-cell leukemia virus family is very closely linked to the immunodeficiency disease". In: *Science* 224 (1984), pp. 475–478.

[27] Kathleen Case. "Nomenclature: human immunodeficiency virus". In: *Annals of internal medicine* 105.1 (1986), pp. 133–133.

[28] François Clavel et al. "Isolation of a new human retrovirus from West African patients with AIDS". In: *Science* 233.4761 (1986), pp. 343–346.

[29] John Coffin et al. "Human immunodeficiency viruses". In: *Science* 232.4751 (1986), p. 697.

[30] Ashley T Haase. "Pathogenesis of lentivirus infections". In: *Nature* 322.6075 (1986), p. 130.

[31] Harold Varmus. "Retroviruses". In: *Science* 240.4858 (1988), pp. 1427–1435.

[32] Robin A Weiss. "How does HIV cause AIDS?" In: *Science* 260.5112 (1993), pp. 1273–1279.

[33] Howard E Gendelrnan et al. "The macrophage in the persistence and pathogenesis of HIV infection". In: *Aids* 3.8 (1989), pp. 475–496.

[34] Li Wu and Vineet N KewalRamani. "Dendritic-cell interactions with HIV: infection and viral dissemination". In: *Nature reviews immunology* 6.11 (2006), p. 859.

[35] NAM Aidsmap. *HIV-1 and HIV-2 [Accessed: 27 December 2018]*. `http://www.aidsmap.com/HIV-1-and-HIV-2/page/1322970`. June 2012.

[36] Avert. *HIV strains and types [Accessed: 27 December 2018]*. `https://www.avert.org/professionals/hiv-science/types-strains`. July 2017.

[37] Anna Maria Geretti et al. "Effect of HIV-1 subtype on virologic and immunologic response to starting highly active antiretroviral therapy". In: *Clinical infectious diseases* 48.9 (2009), pp. 1296–1305.

[38] Massimo Gentile et al. "Determination of the size of HIV using adenovirus type 2 as an internal length marker". In: *Journal of virological methods* 48.1 (1994), pp. 43–52.

[39] David McDonald et al. "Visualization of the intracellular behavior of HIV in living cells". In: *The Journal of cell biology* 159.3 (2002), pp. 441–452.

[40] VM Vogt. "Retroviral virions and genomes". In: (1997).

[41] John AG Briggs et al. "Structural organization of authentic, mature HIV-1 virions and cores". In: *The EMBO journal* 22.7 (2003), pp. 1707–1715.

[42] Thomas Splettstoesse (www.scistyle.com). *Graphic: Structure of the RNA genome of HIV-1 [Accessed: 30 December 2018]*. `https://commons.wikimedia.org/wiki/File:HIV-genome.png`. License: https://creativecommons.org/licenses/by-sa/3.0/legalcode.

[43] Thomas Splettstoesse (www.scistyle.com). *Graphic: HIV capsid p24 [Accessed: 08 January 2018]*. `https://commons.wikimedia.org/wiki/File:P24_HIV-capsid.png`. License: `https://creativecommons.org/licenses/by-sa/4.0/legalcode`.

[44] Wesley I Sundquist and Hans-Georg Kräusslich. "HIV-1 assembly, budding, and maturation". In: *Cold Spring Harbor perspectives in medicine* (2012), a006924.

[45] David C Chan et al. "Core structure of gp41 from the HIV envelope glycoprotein". In: *Cell* 89.2 (1997), pp. 263–273.

[46] Simona Fiorentini et al. "Functions of the HIV-1 matrix protein p17". In: *Microbiologica-Quarterly Journal of Microbiological Sciences* 29.1 (2006), pp. 1–10.

[47] Owen Pornillos, Barbie K Ganser-Pornillos, and Mark Yeager. "Atomic-level modelling of the HIV capsid". In: *Nature* 469.7330 (2011), p. 424.

[48] Alan D Frankel and John AT Young. "HIV-1: fifteen proteins and an RNA". In: (1998).

[49] HIV Sequence Database. *HIV Gene Map [Accessed: 30 December 2018]*. `https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html`. 2018.

[50] Fred C Krebs et al. "Lentiviral LTR-directed expression, sequence variation, and disease pathogenesis". In: *HIV sequence compendium* (2001), pp. 29–70.

[51] Kevin A Wilkinson et al. "High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states". In: *PLoS biology* 6.4 (2008), e96.

[52] Ekram W Abd El-Wahab et al. "Specific recognition of the HIV-1 genomic RNA by the Gag precursor". In: *Nature communications* 5 (2014), p. 4304.

[53] Joseph M Watts et al. "Architecture and secondary structure of an entire HIV-1 RNA genome". In: *Nature* 460.7256 (2009), p. 711.

[54] Florence Baudin et al. "Functional sites in the 5′ region of human immunodeficiency virus type 1 RNA form defined structural domains". In: *Journal of molecular biology* 229.2 (1993), pp. 382–397.

[55] Rodney S Russell, Chen Liang, and Mark A Wainberg. "Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably?" In: *Retrovirology* 1.1 (2004), p. 23.

[56] Alexandra Valsamakis, Nancy Schek, and JAMES C Alwine. "Elements upstream of the AAUAAA within the human immunodeficiency virus polyadenylation signal are required for efficient polyadenylation in vitro." In: *Molecular and cellular biology* 12.9 (1992), pp. 3699–3705.

[57] Jean-Christophe Paillart et al. "First snapshots of the HIV-1 RNA structure in infected cells and in virions". In: *Journal of Biological Chemistry* 279.46 (2004), pp. 48397–48403.

[58] Ben Berkhout and Jeroen LB van Wamel. "The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure". In: *Rna* 6.2 (2000), pp. 282–295.

[59] Eugene Skripkin et al. "Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro". In: *Proceedings of the National Academy of Sciences* 91.11 (1994), pp. 4945–4949.

[60] Gaya K Amarasinghe et al. "NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the Ψ-RNA packaging signal. implications for genome recognition 1". In: *Journal of molecular biology* 301.2 (2000), pp. 491–511.

[61] ANDREW Lever et al. "Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions." In: *Journal of virology* 63.9 (1989), pp. 4085–4087.

[62] Amanda Zeffman et al. "The major HIV-1 packaging signal is an extended bulged stem loop whose structure is altered on interaction with the gag polyprotein1". In: *Journal of molecular biology* 297.4 (2000), pp. 877–893.

[63] Victoria D'Souza and Michael F Summers. "How retroviruses select their genomes". In: *Nature Reviews Microbiology* 3.8 (2005), p. 643.

[64] Laurent Houzet et al. "HIV controls the selective packaging of genomic, spliced viral and cellular RNAs into virions through different mechanisms". In: *Nucleic acids research* 35.8 (2007), pp. 2695–2704.

[65] Gaya K Amarasinghe et al. "Stem-loop SL4 of the HIV-1 Ψ RNA packaging signal exhibits weak affinity for the nucleocapsid protein. structural studies and implications for genome recognition1". In: *Journal of molecular biology* 314.5 (2001), pp. 961–970.

[66] Victoria W Pollard and Michael H Malim. "The HIV-1 rev protein". In: *Annual Reviews of Microbiology* 52.1 (1998), pp. 491–532.

[67] Mayte Coiras et al. "Understanding HIV-1 latency provides clues for the eradication of long-term reservoirs". In: *Nature Reviews Microbiology* 7.11 (2009), p. 798.

[68] Craig B Wilen, John C Tilton, and Robert W Doms. "HIV: cell binding and entry". In: *Cold Spring Harbor perspectives in medicine* (2012), a006866.

[69] David C Chan and Peter S Kim. "HIV entry and its inhibition". In: *Cell* 93.5 (1998), pp. 681–684.

[70] Alissa Bukrinskaya et al. "Establishment of a functional human immunodeficiency virus type 1 (HIV-1) reverse transcription complex involves the cytoskeleton". In: *Journal of Experimental Medicine* 188.11 (1998), pp. 2113–2125.

[71] Eric O Freed. “HIV-1 replication”. In: *Somatic cell and molecular genetics* 26.1-6 (2001), pp. 13–33.

[72] Sergey N Iordanskiy and Michael I Bukrinsky. “Reverse transcription complex: the key player of the early phase of HIV replication”. In: (2007).

[73] Serguei Popov et al. “Viral protein R regulates nuclear import of the HIV-1 pre-integration complex”. In: *The EMBO journal* 17.4 (1998), pp. 909–917.

[74] Philippe Gallay et al. “HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway”. In: *Proceedings of the National Academy of Sciences* 94.18 (1997), pp. 9825–9830.

[75] Véronique Zennou et al. “HIV-1 genome nuclear import is mediated by a central DNA flap”. In: *Cell* 101.2 (2000), pp. 173–185.

[76] John Hiscott, Hakju Kwon, and Pierre Génin. “Hostile takeovers: viral appropriation of the NF-kB pathway”. In: *The Journal of clinical investigation* 107.2 (2001), pp. 143–151.

[77] Eric O Freed. “HIV-1 assembly, release and maturation”. In: *Nature Reviews Microbiology* 13.8 (2015), p. 484.

[78] Noé Dubois et al. “Retroviral RNA Dimerization: From Structure to Functions”. In: *Frontiers in microbiology* 9 (2018), p. 527.

[79] Mauricio Comas-Garcia et al. “Dissection of specific binding of HIV-1 Gag to the "packaging signal" in viral RNA”. In: *Elife* 6 (2017), e27055.

[80] Yantao Yang et al. “Roles of Gag-RNA interactions in HIV-1 virus assembly deciphered by single-molecule localization microscopy”. In: *Proceedings of the National Academy of Sciences* 115.26 (2018), pp. 6721–6726.

[81] Sebla B Kutluay et al. “Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis”. In: *Cell* 159.5 (2014), pp. 1096–1109.

[82] Ian B Hogue et al. “Gag induces the coalescence of clustered lipid rafts and tetraspanin-enriched microdomains at HIV-1 assembly sites on the plasma membrane”. In: *Journal of virology* (2011), JVI–00743.

[83] Dzung H Nguyen and James EK Hildreth. “Evidence for budding of human immunodeficiency virus type 1 selectively from glycolipid-enriched membrane lipid rafts”. In: *Journal of virology* 74.7 (2000), pp. 3264–3272.

[84] Akira Ono and Eric O Freed. “Plasma membrane rafts play a critical role in HIV-1 assembly and release”. In: *Proceedings of the National Academy of Sciences* 98.24 (2001), pp. 13925–13930.

[85] Tsutomu Murakami and Eric O Freed. “Genetic evidence for an interaction between human immunodeficiency virus type 1 matrix and α-helix 2 of the gp41 cytoplasmic tail”. In: *Journal of virology* 74.8 (2000), pp. 3548–3554.

[86] Mary Ann Checkley, Benjamin G Luttge, and Eric O Freed. “HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation”. In: *Journal of molecular biology* 410.4 (2011), pp. 582–608.

[87] Tsutomu Murakami et al. “Regulation of human immunodeficiency virus type 1 Env-mediated membrane fusion by viral protease activity”. In: *Journal of virology* 78.2 (2004), pp. 1026–1031.

[88] Donald J Wyma et al. “Coupling of human immunodeficiency virus type 1 fusion to virion maturation: a novel role of the gp41 cytoplasmic tail”. In: *Journal of virology* 78.7 (2004), pp. 3429–3435.

[89] Margaret A Fischl et al. "The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex". In: *New England Journal of Medicine* 317.4 (1987), pp. 185–191.

[90] AIDSinfo. *What to Start: Choosing an HIV Regimen [Accessed: 12 January 2018]*. `https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/21/53/what-to-start--choosing-an-hiv-regimen`. Nov. 2018.

[91] Richard D Moore and Richard E Chaisson. "Natural history of HIV infection in the_era of combination antiretroviral therapy". In: *Aids* 13.14 (1999), pp. 1933–1942.

[92] Avert. *Starting antiretroviral treatment for HIV [Accessed: 12 January 2018]*. `https://www.avert.org/living-with-hiv/starting-treatment`. Dec. 2017.

[93] Didier K Ekouevi et al. "Antiretroviral therapy response among HIV-2 infected patients: a systematic review". In: *BMC infectious diseases* 14.1 (2014), p. 461.

[94] Theodosius Dobzhansky. "A review of some fundamental concepts and problems of population genetics". In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 20. Citeseer. 1955, pp. 1–15.

[95] Julius Van der Werf et al. *Adaptation and fitness in animal populations*. Springer, 2009.

[96] H Allen Orr. "Fitness and its role in evolutionary genetics". In: *Nature Reviews Genetics* 10.8 (2009), p. 531.

[97] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*. Vol. 1. Proc. 6th Int. Cong. Genet, 1932.

[98] Andrew L Ferguson et al. "Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design". In: *Immunity* 38.3 (2013), pp. 606–617.

[99] Frank J Poelwijk et al. "Empirical fitness landscapes reveal accessible evolutionary paths". In: *Nature* 445.7126 (2007), p. 383.

[100] Philip A Romero and Frances H Arnold. "Exploring protein fitness landscapes by directed evolution". In: *Nature reviews Molecular cell biology* 10.12 (2009), p. 866.

[101] John H Gillespie. "A simple stochastic gene substitution model". In: *Theoretical population biology* 23.2 (1983), pp. 202–215.

[102] Simon P Blomberg and Theodore Garland. "Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods". In: *Journal of Evolutionary Biology* 15.6 (2002), pp. 899–910.

[103] Janis Antonovics and Peter H van Tienderen. "Ontoecogenophyloconstraints? The chaos of constraint terminology". In: *Trends in Ecology & Evolution* 6.5 (1991), pp. 166–168.

[104] Thomas F Hansen. "Evolutionary constraints". In: *Oxford Bibliographies in Evolutionary Biology* (2014).

[105] D Brent Burt. "Evolutionary stasis, constraint and other terminology describing evolutionary patterns". In: *Biological Journal of the Linnean Society* 72.4 (2001), pp. 509–517.

[106] Günter P Wagner. "The influence of variation and of developmental constraints on the rate of multivariate phenotypic evolution". In: *Journal of Evolutionary Biology* 1.1 (1988), pp. 45–66.

[107] Kurt Schwenk and Gunter Wagner. "The relativism of constraints on phenotypic evolution". In: Jan. 2004, pp. 390–408.

[108] William Bateson and Gregor Mendel. *Mendel's principles of heredity*. "Part II: 1. Biographical notice of Mendel. 2. Translation of the paper on hybridisation. 3. Translation of the paper on Hieracium": p. [307]-368. Cambridge: University Press, 1909, p. 450. URL: `https://www.biodiversitylibrary.org/item/96877`.

[109] Heather J Cordell. "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". In: *Human molecular genetics* 11.20 (2002), pp. 2463–2468.

[110] Ilona Miko. "Epistasis: Gene Interactions and Phenotypic Effects". In: (2008).

[111] Ronald A Fisher. "XV.—The correlation between relatives on the supposition of Mendelian inheritance." In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52.2 (1919), pp. 399–433.

[112] Ramamurthy Mani et al. "Defining genetic interaction". In: *Proceedings of the National Academy of Sciences* 105.9 (2008), pp. 3461–3466.

[113] Mato Lagator et al. "On the mechanistic nature of epistasis in a canonical cis-regulatory element". In: *eLife* 6 (2017), e25192.

[114] Patrick C Phillips. "Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems". In: *Nature Reviews Genetics* 9.11 (2008), p. 855.

[115] Daniel M Weinreich, Richard A Watson, and Lin Chao. "Perspective: sign epistasis and genetic costraint on evolutionary trajectories". In: *Evolution* 59.6 (2005), pp. 1165–1174.

[116] David De Juan, Florencio Pazos, and Alfonso Valencia. "Emerging methods in protein co-evolution". In: *Nature Reviews Genetics* 14.4 (2013), p. 249.

[117] BT Korber et al. "Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis". In: *Proceedings of the National Academy of Sciences* 90.15 (1993), pp. 7176–7180.

[118] Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction". In: *Bioinformatics* 24.3 (2007), pp. 333–340.

[119] Elisabeth RM Tillier and Thomas WH Lui. "Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments". In: *Bioinformatics* 19.6 (2003), pp. 750–755.

[120] LC Martin et al. "Using information theory to search for co-evolving residues in proteins". In: *Bioinformatics* 21.22 (2005), pp. 4116–4124.

[121] Ulrike Göbel et al. "Correlated mutations and residue contacts in proteins". In: *Proteins: Structure, Function, and Bioinformatics* 18.4 (1994), pp. 309–317.

[122] William R Taylor and Kerr Hatrick. "Compensating changes in protein multiple sequence alignments". In: *Protein Engineering, Design and Selection* 7.3 (1994), pp. 341–348.

[123] Erwin Neher. "How frequent are correlated changes in families of protein sequences?" In: *Proceedings of the National Academy of Sciences* 91.1 (1994), pp. 98–102.

[124] Chen-Hsiang Yeang and David Haussler. "Detecting coevolution in and among protein domains". In: *PLoS computational biology* 3.11 (2007), e211.

[125] Julien Dutheil et al. "A model-based approach for detecting coevolving positions in a molecule". In: *Molecular biology and evolution* 22.9 (2005), pp. 1919–1928.

[126] Alan S Lapedes et al. “Correlated mutations in models of protein sequences: phylogenetic and structural effects”. In: *Lecture Notes-Monograph Series* (1999), pp. 236–256.

[127] Martin Weigt et al. “Identification of direct residue contacts in protein–protein interaction by message passing”. In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

[128] Magnus Ekeberg et al. “Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models”. In: *Physical Review E* 87.1 (2013), p. 012707.

[129] David T Jones et al. “PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments”. In: *Bioinformatics* 28.2 (2011), pp. 184–190.

[130] Lukas Burger and Erik Van Nimwegen. “Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method”. In: *Molecular systems biology* 4.1 (2008), p. 165.

[131] Claude Elwood Shannon. “A mathematical theory of communication”. In: *Bell system technical journal* 27.3 (1948), pp. 379–423.

[132] Richard R Stein, Debora S Marks, and Chris Sander. “Inferring pairwise interactions from biological data using maximum-entropy probability models”. In: *PLoS computational biology* 11.7 (2015), e1004182.

[133] Jan Krumsiek et al. “Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data”. In: *BMC systems biology* 5.1 (2011), p. 21.

[134] Edwin T Jaynes. “Information theory and statistical mechanics”. In: *Physical review* 106.4 (1957), p. 620.

[135] Harald Cramér. *Mathematical methods of statistics (PMS-9)*. Vol. 9. Princeton university press, 2016.

[136] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[137] T Plefka. “Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model”. In: *Journal of Physics A: Mathematical and general* 15.6 (1982), p. 1971.

[138] Antoine Georges and Jonathan S Yedidia. “How to expand around mean-field theory using high-temperature expansions”. In: *Journal of Physics A: Mathematical and General* 24.9 (1991), p. 2173.

[139] David J Thouless, Philip W Anderson, and Robert G Palmer. “Solution of solvable model of a spin glass”. In: *Philosophical Magazine* 35.3 (1977), pp. 593–601.

[140] Eleonora De Leonardis et al. “Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction”. In: *Nucleic acids research* 43.21 (2015), pp. 10444–10455.

[141] Caleb Weinreb et al. “3D RNA and functional interactions from evolutionary couplings”. In: *Cell* 165.4 (2016), pp. 963–975.

[142] Magdalena A Jonikas et al. “Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters”. In: *Rna* 15.2 (2009), pp. 189–199.

[143] Ulf-Peter Guenther et al. “Hidden specificity in an apparently nonspecific RNA-binding protein”. In: *Nature* 502.7471 (2013), p. 385.

[144] Jacob M Tome et al. "Comprehensive analysis of RNA-protein interactions by high-throughput sequencing–RNA affinity profiling". In: *Nature methods* 11.6 (2014), p. 683.

[145] Nicole Lambert et al. "RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins". In: *Molecular cell* 54.5 (2014), pp. 887–900.

[146] Jason D Buenrostro et al. "Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes". In: *Nature biotechnology* 32.6 (2014), p. 562.

[147] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[148] Illumina. *Illumina Sequencing Technology [Accessed: 05. March 2019]*. `https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`. 2010.

[149] Illumina. *An introduction to Next-Generation Sequencing Technology [Accessed: 05. March 2019]*. `https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf`. 2017.

[150] CeGaT GmbH. *Next-Generation Sequencing [Accessed: 05. March 2019]*. `https://www.cegat.de/services/next-generation-sequencing/`.

[151] Brent Ewing et al. "Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment". In: *Genome research* 8.3 (1998), pp. 175–185.

[152] Illumina. *Quality Scores for Next-Generation Sequencing [Accessed: 07. March 2019]*. `https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf`. 2011.

[153] Brent Ewing and Phil Green. "Base-calling of automated sequencer traces using phred. II. Error probabilities". In: *Genome research* 8.3 (1998), pp. 186–194.

[154] *FASTQ format - Encoding [Accessed: 24. March 2019]*. `https://en.wikipedia.org/wiki/FASTQ_format`.

[155] Stuart FJ Le Grice. "Targeting the HIV RNA genome: high-hanging fruit only needs a longer ladder". In: *The Future of HIV-1 Therapeutics*. Springer, 2015, pp. 147–169.

[156] Delphine Muriaux et al. "RNA is a structural element in retrovirus particles". In: *Proceedings of the National Academy of Sciences* 98.9 (2001), pp. 5246–5251.

[157] Mauricio Comas-Garcia, Sean Davis, and Alan Rein. "On the selective packaging of genomic RNA by HIV-1". In: *Viruses* 8.9 (2016), p. 246.

[158] John AG Briggs et al. "The stoichiometry of Gag protein in HIV-1". In: *Nature Structural and Molecular Biology* 11.7 (2004), p. 672.

[159] Katrina A Lehmann and Brenda L Bass. "Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities". In: *Biochemistry* 39.42 (2000), pp. 12875–12884.

[160] Rodolphe Suspène et al. "Double-stranded RNA adenosine deaminase ADAR-1-induced hypermutated genomes among inactivated seasonal influenza and live attenuated measles virus vaccines". In: *Journal of virology* 85.5 (2011), pp. 2458–2462.

[161] Joseph Sodroski et al. "Trans-acting transcriptional regulation of human T-cell leukemia virus type III long terminal repeat". In: *Science* 227.4683 (1985), pp. 171–173.

[162] Joseph Sodroski et al. “Location of the trans-activating region on the genome of human T-cell lymphotropic virus type III”. In: *Science* 229.4708 (1985), pp. 74–77.

[163] Christine H Herrmann and Andrew P Rice. “Lentivirus Tat proteins specifically associate with a cellular protein kinase, TAK, that hyperphosphorylates the carboxyl-terminal domain of the large subunit of RNA polymerase II: candidate for a Tat cofactor.” In: *Journal of Virology* 69.3 (1995), pp. 1612–1620.

[164] Christine H Herrmann, Moses O Gold, and Andrew P Rice. “Viral transactivators specifically target distinct cellular protein kinases that phosphorylate the RNA polymerase II C-terminal domain”. In: *Nucleic acids research* 24.3 (1996), pp. 501–508.

[165] Michael F Bardaro Jr et al. “How binding of small molecule and peptide ligands to HIV-1 TAR alters the RNA motional landscape”. In: *Nucleic acids research* 37.5 (2009), pp. 1529–1540.

[166] Sara Richter, Hong Cao, and Tariq M Rana. “Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1- Tat- TAR ternary complex formation”. In: *Biochemistry* 41.20 (2002), pp. 6391–6397.

[167] Bep Klaver and Ben Berkhout. “Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus.” In: *The EMBO journal* 13.11 (1994), pp. 2650–2659.

[168] David Harrich et al. “Differential growth kinetics are exhibited by human immunodeficiency virus type 1 TAR mutants.” In: *Journal of virology* 68.9 (1994), pp. 5899–5910.

[169] Bin Tian et al. “A large-scale analysis of mRNA polyadenylation of human and mouse genes”. In: *Nucleic acids research* 33.1 (2005), pp. 201–212.

[170] A Jakobovits et al. “A discrete element 3′ of human immunodeficiency virus 1 (HIV-1) and HIV-2 mRNA initiation sites mediates transcriptional activation by an HIV trans activator.” In: *Molecular and cellular biology* 8.6 (1988), pp. 2555–2561.

[171] Ben Berkhout, Robert H Silverman, and Kuan-Teh Jeang. “Tat trans-activates the human immunodeficiency virus through a nascent RNA target”. In: *Cell* 59.2 (1989), pp. 273–282.

[172] Sandy Feng and Eric C Holland. “HIV-1 tat trans-activation requires the loop sequence within tar”. In: *Nature* 334.6178 (1988), p. 165.

[173] Ping Wei et al. “A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA”. In: *Cell* 92.4 (1998), pp. 451–462.

[174] Jared L Clever and Tristram G Parslow. “Mutant human immunodeficiency virus type 1 genomes with defects in RNA dimerization or encapsidation.” In: *Journal of Virology* 71.5 (1997), pp. 3407–3414.

[175] Atze T Das et al. “A conserved hairpin motif in the R-U5 region of the human immunodeficiency virus type 1 RNA genome is essential for replication.” In: *Journal of Virology* 71.3 (1997), pp. 2346–2356.

[176] Jared L Clever, Daniel A Eckstein, and Tristram G Parslow. “Genetic dissociation of the encapsidation and reverse transcription functions in the 5′ R region of human immunodeficiency virus type 1”. In: *Journal of virology* 73.1 (1999), pp. 101–109.

[177] Max von Kleist and Maureen Smith. “Experimental design: Optimization of mutation rate for library preparation in MIME”. In: *Technical Report (Publication in Progress)* (2019).

[178] John P Barton et al. "Large pseudocounts and l 2-norm penalties are necessary for the mean-field inference of Ising and Potts models". In: *Physical Review E* 90.1 (2014), p. 012132.

[179] Kevin V Morris and John S Mattick. "The rise of regulatory RNA". In: *Nature Reviews Genetics* 15.6 (2014), p. 423.

[180] Thomas Hermann. "Strategies for the design of drugs targeting RNA and RNA–protein complexes". In: *Angewandte Chemie International Edition* 39.11 (2000), pp. 1890–1904.

[181] Jason R Thomas and Paul J Hergenrother. "Targeting RNA with small molecules". In: *Chemical reviews* 108.4 (2008), pp. 1171–1224.

[182] Thomas Hermann. "Small molecules targeting viral RNA". In: *Wiley Interdisciplinary Reviews: RNA* 7.6 (2016), pp. 726–743.

[183] Colleen M Connelly, Michelle H Moon, and John S Schneekloth Jr. "The emerging role of RNA as a therapeutic target for small molecules". In: *Cell chemical biology* 23.9 (2016), pp. 1077–1090.

[184] Houng-Yau Mei et al. "Inhibition of an HIV-1 Tat-derived peptide binding to TAR RNA by aminoglycoside antibiotics". In: *Bioorganic & Medicinal Chemistry Letters* 5.22 (1995), pp. 2755–2760.

[185] Christian Bailly et al. "The binding mode of drugs to the TAR RNA of HIV-1 studied by electric linear dichroism". In: *Nucleic acids research* 24.8 (1996), pp. 1460–1464.

[186] Houng-Yau Mei et al. "Discovery of selective, small-molecule inhibitors of RNA complexes—1. The tat protein/TAR RNA complexes required for HIV-1 transcription". In: *Bioorganic & medicinal chemistry* 5.6 (1997), pp. 1173–1184.

[187] Sunil Kumar and Dev P Arya. "Recognition of HIV TAR RNA by triazole linked neomycin dimers". In: *Bioorganic & medicinal chemistry letters* 21.16 (2011), pp. 4788–4792.

[188] Joanna Sztuba-Solinska et al. "Identification of biologically active, HIV TAR RNA-binding small molecules using small molecule microarrays". In: *Journal of the American Chemical Society* 136.23 (2014), pp. 8402–8410.

[189] Simona Cocco et al. "Inverse statistical physics of protein sequences: a key issues review". In: *Reports on Progress in Physics* 81.3 (2018), p. 032601.

[190] Keqin Chen and Frances H Arnold. "Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide". In: *Proceedings of the National Academy of Sciences* 90.12 (1993), pp. 5618–5622.

[191] Edward J Merino et al. "RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE)". In: *Journal of the American Chemical Society* 127.12 (2005), pp. 4223–4231.

[192] Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. "Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution". In: *Nature protocols* 1.3 (2006), p. 1610.