Bachelor Thesis

Viral mRNA Secondary Structures Affect the Thermodynamics of Frameshifting

prepared by

Annke de Maeyer

at the Max Planck Institute
for Multidisciplinary Sciences

Matriculation number:	21972492
Submission date:	18 October 2022
Supervisors:	Dr. Lars Bock and Sara Gabrielli
First Referee:	Prof. Dr. Helmut Grubmüller
Second Referee:	Prof. Dr. Stefan Klumpp

Acknowledgement

First, many thanks to Prof. Helmut Grubmüller for giving me the opportunity to become a part of his team, and for his support during my work on the bachelor thesis. I am also thankful to Prof. Stefan Klumpp for being my second referee.

Special thanks to my supervisors Dr. Lars Bock and Sara Gabrielli for their patient guidance, countless advice, and encouragement. There was never a bad timing for questions, and I am very grateful for the detailed answers and discussions that followed them.

Finally, I would like to acknowledge the whole department for the friendly and welcoming working environment and especially Lisa-Marie Heß for becoming a close friend throughout my work on this thesis.

Contents

1. Introduction

Production of proteins is essential for life on earth. This process is called protein biosynthesis and occurs inside cells in a biomolecular complex termed ribosome. The ribosome decodes the genetic information stored in messenger RNA (mRNA), a single-stranded molecule which contains a sequence of nucleotides. During protein synthesis, the nucleotides are read by the ribosome in groups of three (codons). Each codon in the mRNA is translated into one amino acid in the synthesized protein. Hence, mRNA works as a template and encodes the sequence of amino acids for a specific protein. The ribosome moves along the mRNA one codon at a time and adds the encoded amino acid to the growing protein. The movement follows the so-called downstream direction, which is opposite to the upstream direction [1, 2].

In general, the process of the protein synthesis is well understood, but many details are still not fully known, one of them being a mechanism called frameshifting. During frameshifting, the ribosome „slips“ on the mRNA, in such a way that the reading frame of the ribosome is shifted by one, two, or four nucleotides in comparison to the original frame, termed 0 frame [3]. The region of the mRNA where frameshifting typically takes place is called slippery sequence [4]. After frameshifting occurred, the codons are read in the shifted reading frame and are therefore generally different from the ones that are read when frameshifting does not occur. As a consequence, a different protein with a different sequence of amino acids is synthesized [2]. This work will focus on -1 frameshifting, during which the ribosome shifts into the -1 reading frame, such that the codons start one nucleotide upstream of the original position in the 0 frame. Usually, frameshifting only occurs rarely [5] and does not result in functional proteins. However, in many organisms sequences have evolved that lead to a high probability of frameshifting. In this case the decoding of both reading frames results in functional proteins. This phenomenon is called programmed ribosomal frameshifting (PRF) and has the advantage that one sequence of mRNA can decode multiple proteins [1, 6]. For several viruses this process is crucial: reducing the PRF in HIV causes the virus to be less infectious [7].

PRF is quantified by the frameshifting efficiency, which is the probability for the ribosome to shift [6]. In the case of *E. coli* and for a specific gene *dnaX*, it was shown that the frameshifting efficiency can be reproduced and predicted with a thermodynamic model [8]: the free-energy difference between the 0 frame and -1 frame at a fixed temperature gives a probability corresponding to the frameshifting efficiency.

PRF takes place when translation is slowed down, because this gives the ribosome enough time to overcome the free-energy barrier between the 0 frame and -1. The stalling of translation is often due to the presence of secondary structure elements in the mRNA (e.g. stem-loops or pseudoknots), which interact with the translating ribosome

and impede its downstream movement [6]. In addition, the interaction of the structured elements with the ribosome and their destabilization during the movement of the ribosome along the mRNA might also affect the frameshifting efficiency. In this work I want to test this hypothesis by analyzing high-throughput data from Mikl et al. [9], where measurements related to the frameshifting efficiencies of more than 12,000 mRNA sequences, based on viral, bacterial, and human PRF events, are provided. I will employ Bayesian statistics to estimate the free-energy difference between the 0 frame and -1 frame from the experimentally determined frameshifting efficiency. In particular, I aim at comparing the effect of the different downstream structures on the free energies.

2. Biological Background

In order to understand the mechanisms occurring during protein synthesis, I explain the basics of DNA, mRNA, tRNA, and of the ribosome in this section. The knowledge of how proteins are synthesized in cells is essential to study the process of ribosomal frameshifting. Since the aim of this work is to compare the effects of different downstream structures on the frameshifting efficiency, the subsequent section provides information about secondary structure elements in the mRNA associated with frameshifting. Afterwards, since the analysis will be based on the frameshifting measurements of mainly viral sequences, I give an introduction to viruses in general. Finally, I present each virus taken into account to investigate the thermodynamics of frameshifting.

2.1. DNA, mRNA, tRNA, and the Ribosome

Generally, DNA (deoxyribonucleic acid) is a double stranded molecule forming a double helix, each strand containing a sequence of nucleotides. A nucleotide consists of a base, a pentose-sugar-ring and one or more phosphates. The nucleotides in DNA exhibit four different bases: adenine (A), cytosine (C), guanine (G), and thymine (T). The carbon atoms in the sugar ring of the nucleotide are numbered and labeled using a prime ('). This convention makes it possible to define an end-to-end orientation of the strands: each strand lacks, at one end, one nucleotide at the 5' position and, at the other end, a nucleotide at the 3' position. That is why, the ends of a strand are called 5' end and 3' end. In DNA the two strands of the double helix are oriented in opposite directions (5' to 3' and 3' to 5') and hydrogen bonding between their bases stabilizes the double helix structure of the DNA. Strong hydrogen bonds can only form between the complementary bases G and C, and between the complementary bases A and T. A stable base pair of G and C and of A and T is called a Watson-Crick base pair [1, 2, 10].

During the process of transcription, the genetic code of one strand of the DNA is used as

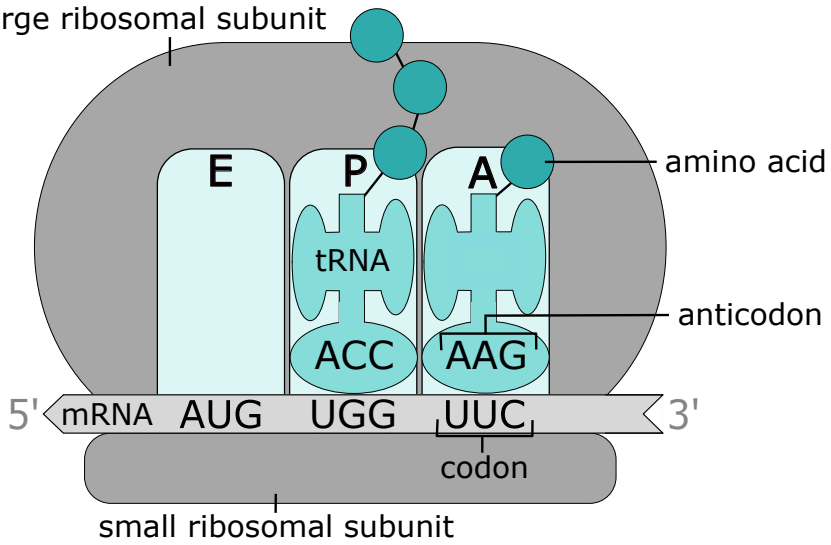

Figure 1: Schematic of the components involved in the translation process: the ribosome with its two subunits and three binding sites E, P, and A, the mRNA with three codons, and aminoacylated tRNAs with anticodons.

a template to produce single stranded mRNA (messenger ribonucleic acid). The mRNA is complementary to one strand of the DNA, but replaces T with the nucleobase uracil (U), which can form a Watson-Crick base pair with A. The mRNA carries genetic information to the ribosome, where protein synthesis takes place. The bases in the mRNA are read by the ribosome in codons from the 5' to the 3' end. Each codon consists of three bases and codes for one amino acid. This correct decoding is ensured by adapter molecules, termed transfer RNAs (tRNAs). Each tRNA contains a distinct anticodon consisting of three nucleotides which can base pair with the complementary codon nucleotides in the mRNA. The tRNAs entering the ribosome are aminoacylated, meaning that they are bound to an amino acid. The type of amino acid bound to the aminoacyl-tRNA depends on the anticodon. In this way, the sequence of codons in the mRNA is translated into the sequence of amino acids in the synthesized protein [1, 2, 10].

The ribosome catalyzes the translation process, whose components are indicated in Fig. 1. The ribosome consists of two subunits, which differ in size. The small ribosomal subunit is responsible for mRNA recruitment and decoding of mRNA codons. The large ribosomal subunit is in charge of peptide-bond formation and contains an exit tunnel through which the growing peptide emerges from the ribosome [6, 10, 11]. The ribosome has three binding sites for mRNA-tRNA base pairs: the aminoacyl (A) site, the peptidyl (P) site, and the exit (E) site [1].

2.2. Protein Synthesis

Once the synthesis of mRNA is completed in the transcription step, there are five more stages until a protein is fully synthesized: activation of amino acids, initiation, elongation, termination, and enzymatic processing and folding.

During the activation of amino acids (1.) the amino acids are bound to their corresponding tRNA: the tRNA gets aminoacylated.

During initiation (2.) the mRNA and the initial aminoacyl-tRNA Met-tRNA$^{\text{Met}}$ bind to the small subunit of the ribosome. In this complex, the anticodon nucleotides of Met-tRNA$^{\text{Met}}$ are paired with the starting codon nucleotides AUG, which signals the beginning of the polypeptide, in the P-site. The binding of the large subunit follows to form the initiation complex.

During elongation (3.) the codons are read and translated one by one along the 5' to 3' direction of the mRNA. This cycle is indicated in Fig. 2. The aminoacyl-tRNA containing the anticodon that is complementary to the next codon, binds to the A site of the initiation complex. Afterwards, the amino acid attached to the tRNA in the P site forms a peptide bond with the amino acid bound to the tRNA in the A site, such that a deacylated tRNA$^{\text{Met}}$ remains in the P site and a dipeptidyl-tRNA is formed in the A site. Next, during a step called translocation, the ribosome moves one codon towards the 3' end of the mRNA. As a consequence, the dipeptidyl-tRNA moves from the A site to the P site and shifts the deacylated tRNA from the P site to the E site. Afterwards, the deacylated tRNA in the E site dissociates from the ribosome. After translocation is completed, the A site is ready to be occupied by the next aminoacyl-tRNA, and the next elongation cycle begins.

Finally, a stop codon in the mRNA signals the termination (4.) of the polypeptide, which is then released from the ribosome. The ribosome is recycled and, afterwards, it is ready for the synthesis of the next protein.

The new polypeptide may undergo enzymatic processing and folds (5.) into a three-dimensional configuration [1].

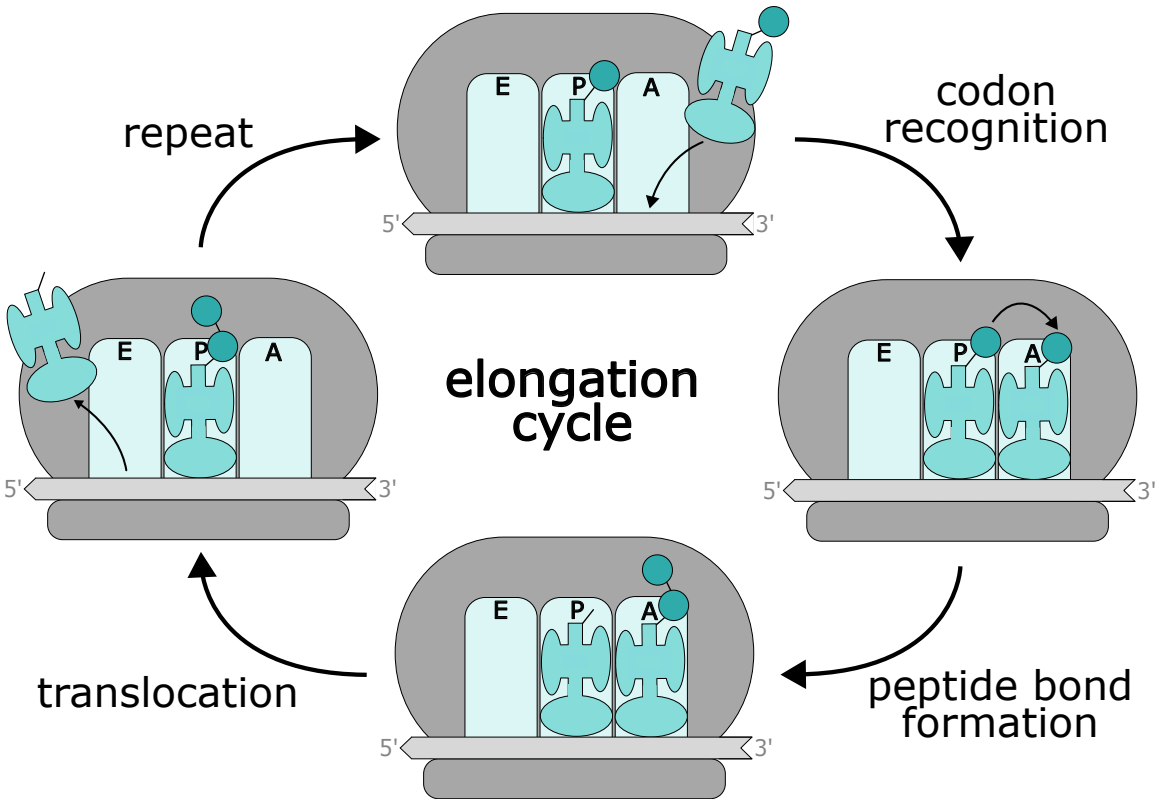

Figure 2: The elongation cycle consists of three steps: codon recognition, peptide bond formation, and translocation.

2.3. Ribosomal Frameshifting

During translocation of the tRNAs in the ribosome, a ribosomal frameshift event can occur. In this case the ribosome shifts by one, two, or four nucleotides on the mRNA. As a consequence, the reading frame of the ribosome is shifted by the same number of nucleotides in comparison to the original frame, termed 0 frame. Starting from the position in the mRNA where a frameshift took place, the codons that are read in the new frame typically differ from the codons in the 0 frame. As a result, the synthesized sequence of amino acids is different from the sequence of amino acids that is synthesized when frameshifting does not occur. The most common type of ribosomal frameshifting is -1 frameshifting, where the ribosome shifts by one nucleotide towards the 5' end of the mRNA. As a consequence, the codons read by the ribosome in the new -1 frame start one nucleotide upstream of the original position in the 0 frame [6, 12]. A schematic of -1 frameshifting is displayed in Fig. 3.

Usually, the sequence of amino acids that is synthesized when a frameshift occurs does not result in a useful protein. On average, this error occurs once in 10^{-4} to 10^{-5} codons [14]. However, certain mRNA sequences have evolved to undergo frameshifting at high efficiencies, a process called programmed ribosomal frameshifting (PRF). In this case, the

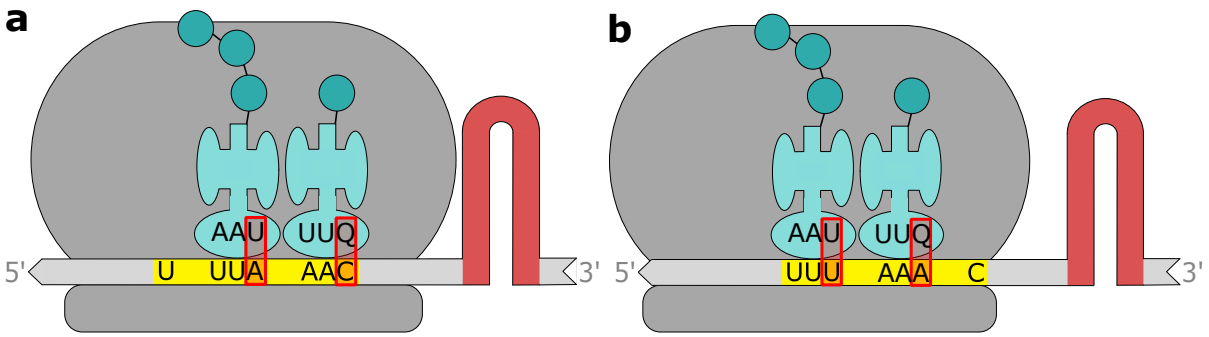

Figure 3: Schematic of -1 frameshifting in the presence of an mRNA stem loop. (**a**) Before frameshifting, the ribosome is in the 0 frame. (**b**) after frameshifting the ribosome shifted one nucleotide upstream to the -1 frame. The codons in the slippery sequence (yellow) in the new -1 frame differ from them in the 0 frame. The secondary structure (red, stem loop) is introduced in Section 2.4. Queuosine (Q) is a modified base [13].

resulting sequence of amino acids can form into a functional protein. In this way, PRF increases the information content of the genome and is used to regulate the expression of proteins [6, 15].

During PRF, the frameshift takes place while a sequence of seven nucleotides called the slippery sequence resides in the ribosome. While this sequence is translated, a frameshift occurs with a certain probability, which is called the frameshifting efficiency (FS). From experimentally measured amounts of the produced peptide, the frameshifting efficiency is generally calculated as

$$\mathrm{FS} = \frac{x\text{-frame products}}{x\text{-frame products} + 0\text{-frame products}}\,, \tag{2.1}$$

where x represents the (positive or negative) number of shifted nucleotides [6]. The amount of products is measured with various techniques, such as western blotting [7, 16], monitoring of fluorescence [9], chromatography or with radioactive labels [8]. The slippery sequence is very sensitive to mutations and expected to be optimized for the programmed frameshifting event, such that mutations in the slippery sequence mostly lead to a reduced frameshifting efficiency [7]. The slippery sequence typically follows the pattern X XXY YYZ, where the spaces indicate the 0 frame codons before the frameshift and the X, Y, and Z denote different bases [17, 18]. The slippery sequences ensure that the codons in the 0 and -1 frame can form stable interactions with the same tRNAs, since the third codon position allows a so-called wobble base pair [6]. A wobble base pair is rather loose and can be formed between the third base of a codon and the corresponding base of its anticodon. In this case, the bases do not need to be complementary in order to pair with each other, as in Fig. 3. In general, a wobble base pair is believed to be beneficial, because it permits a rapid dissociation of tRNA and mRNA, which results in a higher

rate of protein synthesis [1, 2].

2.4. mRNA Downstream Structure

The region downstream (toward the 3' end) of the slippery sequence plays an important role as a frameshifting stimulatory element: secondary structure elements in the mRNA interact with the ribosome and impede its downstream movement. This results in a slowed down translocation, which is necessary for PRF (here -1 PRF), since it gives the ribosome enough time to overcome the free-energy barrier between the 0 frame and the -1 frame.

Two of the most common types of structures into which mRNA folds downstream of the slippery sequence are the stem loop and the pseudoknot. Folding into a stem loop occurs typically when two regions of the strand have complementary bases when reading them in opposite directions. As in Fig. 4a, b, the strand then folds into a structure looking like a hairpin. A pseudoknot as in Fig. 4c, d contains at least two loops and nucleotides in a loop pair with complementary bases outside the loop [12].

Both stem loop and pseudoknot can stimulate frameshifting by slowing down translocation and, thus, giving time to overcome the free-energy barrier between the reading frames. In the presence of a stem loop, experimental evidence from Bao et al. suggests that -1 PRF can occur through two pathways. Firstly, as illustrated in Fig. 5a, the stem loop is expected to be able to interact with the mRNA entry channel and act as a "roadblock", inhibiting the downstream movement of the ribosome, and, therefore, tRNA-mRNA translocation. Secondly, the stem loop is assumed to be able to dock into the A site of the ribosome. As a consequence, the binding of the next tRNA would be inhibited, such that translocation is stalled [21]. In this work I consider only the first mechanism of the "roadblock" effect, because firstly, there is no experimental evidence, that the stem loop stalls the ribosome by docking into the A site specifically for HIV-1 and SIVmac239 (included viruses with a stem loop in this work, see Section 2.5). Secondly, there is currently no evidence that frameshifting would occur when the A site is occupied by a stem loop.

As regards the pseudoknot, on the other hand, there is structural evidence [22] that, in the 0 frame, unfolding occurs as the pseudoknot approaches the entry channel. This is probably due to the higher complexity and larger size of the pseudoknot compared to the stem-loop. The unfolding is done by the ribosome itself, which acts as a helicase [23]. My hypothesis is that, since for the pseudoknot it is energetically favorable to stay folded, it resists the unfolding and, thus, generates a back-pull of the mRNA towards the -1 frame, as visualized in Fig. 5b. As both the "roadblock" effect and this mechanism inhibit translocation, I assume that, also in the case of the pseudoknot, equilibrium between 0 frame and -1 frame is reached. In addition, I hypothesize that the back-pull reduces the

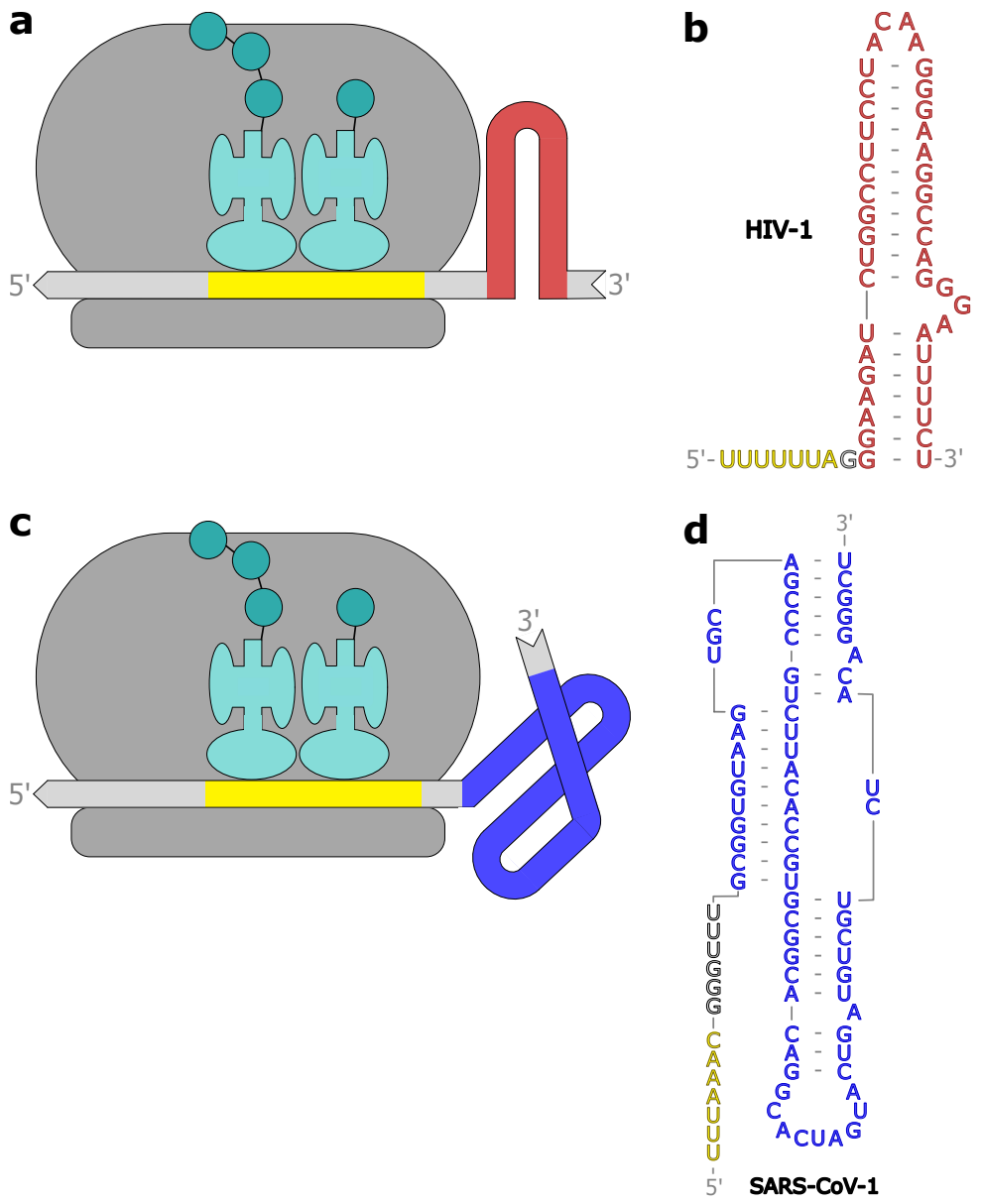

Figure 4: (**a**) Schematic of the translating ribosome and the mRNA with a stem loop (red). (**b**) Sequence of the stem loop of HIV-1 mRNA [7, 19]. The spacer region (grey) between slippery sequence (yellow) and stem loop (red) consists of one nucleotide. (**c**) Schematic of the translating ribosome and the mRNA with a pseudoknot (blue). (**d**) Sequence of the stem loop of SARS-CoV-1 mRNA [20]. The spacer region (grey) between slippery sequence (yellow) and pseudoknot (blue) consists of six nucleotides.

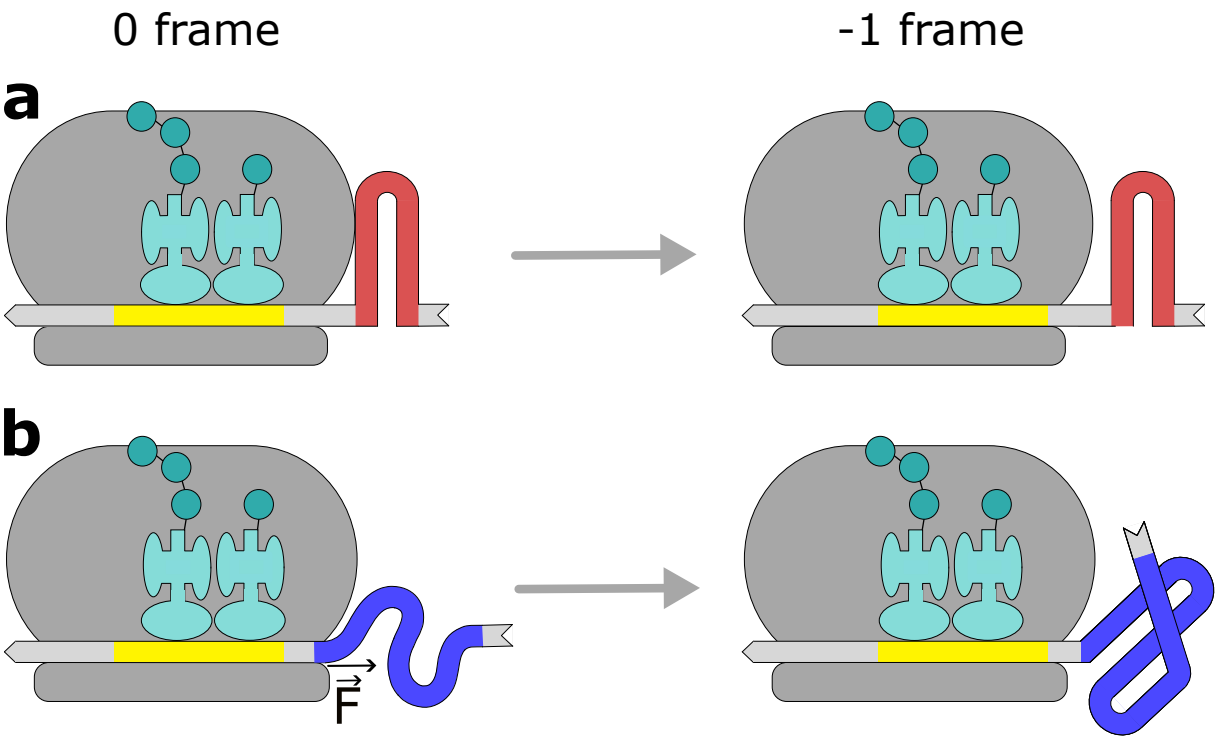

Figure 5: (**a**) A stem loop stalls the ribosome during frameshifting by interacting with the mRNA entry channel and inhibiting translocation [21]. (**b**) On the left: schematic view of a pseudoknot which is partially unfolded in the 0 frame [22]. The pseudoknot might resist the unfolding and generate a back-pull towards the -1 frame (on the right).

free-energy difference between the 0 and the -1 frame. The free energy is introduced in Section 3.1.

I only consider a thermodynamic effect on the frameshifting efficiency and not a kinetic one for the following reasons: Bock et al. support the notion that (with a stem loop) translocation is sufficiently slower than a tRNA slippage, such that there is enough time to overcome the free-energy barrier [8]. However, if translocation was not stalled enough by a pseudoknot, it would result in a lower frameshifting efficiency. Since my results show mostly an increased frameshifting efficiency for sequences with a pseudoknot (see Section 7.5), this effect would be negligible.

2.5. Viruses

A virus is an intracellular pathogen, whose size ranges from five to a few hundred nanometers. As viruses do not have a metabolism, they can only multiply via host cells. Outside of an infected cell, viruses exist as particles called virions. A virion in its simplest case consists of DNA or RNA (never both) and a capsid (coat built of proteins). During an infection, the virion attaches to the host cell and the capsid penetrates into the host cell's cytoplasm. If the virus contains RNA and is a so-called retrovirus, it is able to reverse the

normal flow of genetic information from DNA to RNA by using the enzyme reverse transcriptase, which synthesizes DNA from RNA. The DNA is then integrated into the host cell's nuclear genome, such that the virus takes advantage of the host cell's mechanisms of protein synthesis to produce viral proteins. These proteins then form a new virus core structure, which can be released as a new virion [24–26].

During protein biosynthesis in the host cell, many viruses employ PRF in order to compress genomic information into a smaller amount of space. Additionally, PRF allows for regulation of the relative amounts of proteins produced in the the 0 and shifted frames [15, 25]. For HIV-1 for example, PRF is crucial: Dulude et al. investigated the frameshifting efficiency of eight mutations of the coding sequence of the HIV wild type. Each of the eight mutations had a lower frameshifting efficiency compared to the wild type, which resulted in a decreased synthesis of a certain protein and consequently in a reduced incorporation of viral enzymes into the virions. Therefore, all mutants were attenuated in long-term virus replication [7].

In my work I will take 9 viruses into account to analyze the effect of the mRNA secondary structures downstream of the slippery sequence on the frameshifting free-energy differences. Here, I will give a short introduction to each of these viruses, which I will refer to as HIV, SIV, HERV, HTLV, PLRV, RSV, SARS, SRV, and WNV. HIV and SIV have a stem loop downstream of their slippery sequence (Fig. 6). The other seven viruses exhibit a pseudoknot as their downstream secondary structure (Fig. 6).

The human immunodeficiency virus (HIV) targets the immune system of infected people by destroying and impairing their immune cells, such that their defense against many infections and some types of cancer is weakened. This retrovirus is responsible for a high mortality worldwide as the virus has claimed around 40.1 million lives so far. The acquired immunodeficiency syndrome (AIDS) is the most advanced stage of HIV infection [19, 27].

Simian immunodeficiency viruses (SIV) are closely related to HIV and represent a large group of viruses, found naturally in an extensive number of African primate species. Other than HIV and SRV, SIVs do not lead to an AIDS-like terminal stage of infection. It was proposed that the viruses have been associated and co-evolved with their hosts over a long period of time, such that the SIV-positive primates do not show clinical symptoms [28]. The type of SIV considered in this work is SIVmac239.

HERV stands for human endogenous retrovirus. HERVs are naturally integrated in the human DNA and passed on from one host cell generation to the next one in the host cell's genome (provirus). They can originate from ancient retroviral infections. Their presence in the genome is not expected to have any major effect. In this work, I will include the type HERV-K10 [29, 30].

The human T-lymphotropic virus (HTLV) is the first human oncogenic retrovirus that was discovered. It causes, for instance, adult T-cell leukemia, a type of cancer. The

retrovirus has many endemic areas, such as Southern Japan, Central and South America, and the Caribbean. In my work, I consider the coding sequence of HTLV-1 [31, 32].

PLRV stands for potato leaf roll virus. It mostly infects potato plants and causes leafrolling and stunting [33].

The Rous Sarcoma Virus (RSV) was discovered by Peyton Rous in 1911. It is a retrovirus causing sarcoma in fowls [34].

The severe acute respiratory syndrome (SARS) is caused by the SARS-associated coronavirus (SARS-CoV). In my work I will include SARS-CoV-1, the first type of this disease, which was first reported in China in late 2002. Its symptoms are typically fever, followed by a dry nonproductive cough and shortness of breath. SARS-CoV-1 is not to be confused with SARS-CoV-2, which caused the COVID-19 pandemic [35, 36]. SARS-CoV-1 and SARS-CoV-2, however, have very similar sequences and the frameshift site is almost completely conserved. The the frameshift stimulating pseudoknot differs only by one nucleotide [37].

The simian retrovirus type 1 (SRV-1) is a retrovirus causing simian acquired immune deficiency syndrome (SAIDS) in rhesus macaques. Although SRV-1 is genetically unrelated to HIV, an infection leads to a pathology resembling that of terminal AIDS in humans, as it causes the depletion of certain immune cells [38, 39].

The West Nile Virus (WNV) is a neuropathogen, primarily transmitted by mosquitoes. It is indigenous in Africa, Asia, Europe, and Australia. Birds are its natural hosts, such that the virus maintains naturally in a mosquito-bird-mosquito transmission cycle. Human WNV infections are often subclinical. However, clinical infections can, for instance, lead to a severe meningitis [40].

Originally, I additionally considered the human protein CCR5 (C-C chemokine receptor type 5), a receptor on the surface of white blood cells that plays a role in inflammatory responses [41]. In 2014, Belew et al. described a -1 PRF signal in the mRNA encoding CCR5 with a pseudoknot as a secondary structure [42]. However, during my work on this thesis, experimental evidence by Khan et al. was published stating that PRF does not take place during CCR5 decoding [43]. That is why, I excluded CCR5 from my investigation. Interestingly, Khan et al. deduced with their results that, apart from mutant states of genes and retroelement-derived genes, there are currently no known human genes functionally utilizing efficient -1 PRF [43]. Retroelement genes are transposable (jumping) genes, which are transcribed into RNA, reverse-transcribed into DNA and afterwards, introduced into a new site of the genome [44].

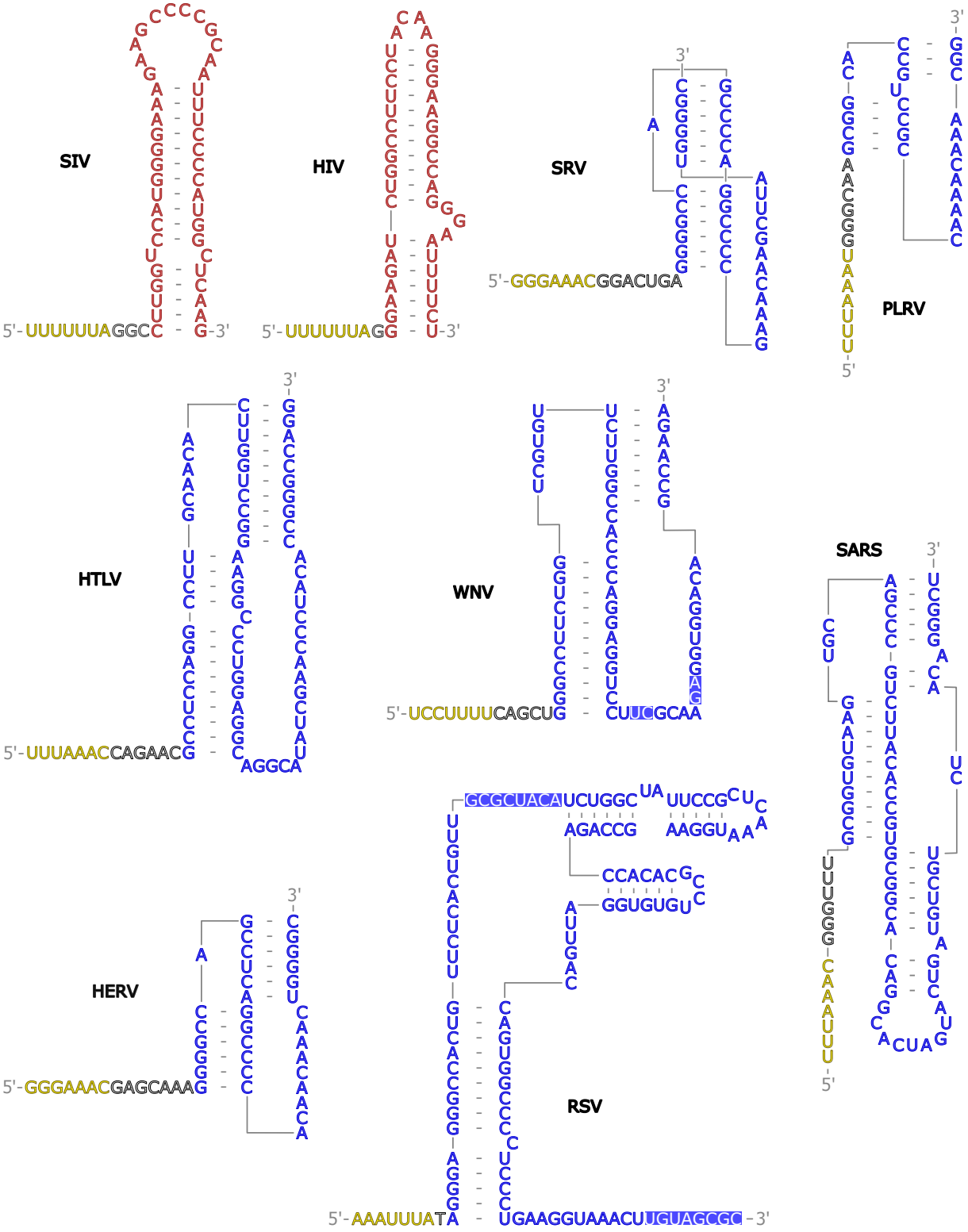

Figure 6: Secondary structures of the frameshift inducing mRNA sequences included in this work: Stem loop (red) of SIV [45] and HIV [7, 19], pseudoknots (blue) of SRV [39, 46, 47], PLRV [48–50], HTLV [51], WNV [52], SARS [20], HERV [53], and RSV [54]. For RSV and WNV, the bases with a blue background also base-pair in the pseudoknot. Additionally, the slippery sequences (yellow) and spacer regions (grey) are depicted.

3. Physical Background

In this section the physical background is introduced. The concept of free energy is explained and, subsequently, used in the thermodynamic model, which quantifies the frameshifting efficiency.

3.1. Free Energy

A state of a system in classical mechanics is described by an energy. The description of microscopic systems containing an ensemble of states, additionally requires the entropy. The analogue of the classical energy is the free energy in statistical physics. There are different free energies that describe different ensembles. The so-called Gibbs free energy G describes an isobaric-isothermal system (fixed pressure P and a fixed temperature T). It is defined as

$$G = U + PV - TS \tag{3.1}$$

with inner energy U, entropy S and volume V [55]. The change of the free energy is defined with the change of inner energy ΔU, volume ΔV, and entropy ΔS as

$$\Delta G = \Delta U + P\Delta V - T\Delta S\,. \tag{3.2}$$

A negative ΔG characterizes a process, that releases energy. In this case spontaneous processes occur to take the system to equilibrium, since the system attempts to minimize its energy. Hence, ΔG measures the favorability of a given reaction quantitatively, here, the favorability of the occurrence of a frameshift [1].

3.2. Thermodynamic Model Based on Free Energy

In the 0 frame and the -1 frame the tRNA anticodons typically pair with different codons in the mRNA slippery sequence (Fig. 3). That is why the base pairs in the codon-anticodon binding exhibit different free energies in the two frames. It can be assumed that these differences in free energy are additive and contribute to the total free-energy difference ΔG [8].

For *E. coli* and a specific gene *dnaX* there is experimental evidence provided by Bock et al. [8] that the frameshifting efficiency can be reproduced and predicted with a thermodynamic model: by assuming a thermodynamic equilibrium between 0 and -1 frame, the distribution of the two states follows a Boltzmann distribution (probability $\propto \mathrm{e}^{-G/k_BT}$).

The frameshifting efficiency (FS) can then be estimated quantitatively:

$$\mathrm{FS} = f_{\mathrm{model}}(\Delta G) = \frac{\mathrm{e}^{-\frac{G(-1\text{ frame})}{k_B T}}}{\mathrm{e}^{-\frac{G(-1\text{ frame})}{k_B T}} + \mathrm{e}^{-\frac{G(0\text{ frame})}{k_B T}}} = \frac{\mathrm{e}^{-\frac{\Delta G}{k_B T}}}{1 + \mathrm{e}^{-\frac{\Delta G}{k_B T}}} \tag{3.3}$$

with Boltzmann Factor k_B and temperature T. ΔG is the difference between the free energies in the -1 and 0 frame:

$$\Delta G = G(-1\text{ frame}) - G(0\text{ frame})\,. \tag{3.4}$$

I hypothesize that a back-pull in the presence of a pseudoknot downstream of the slippery sequence translates into an additional free-energy term that reduces the free-energy difference ΔG. Therefore, when considering the same slippery sequence, I call $\Delta\Delta G$ the difference between the ΔG in the presence of a downstream stem loop and the ΔG in the presence of a downstream pseudoknot. If my hypothesis that a pseudoknot reduces ΔG is true (see section Section 2.4), $\Delta\Delta G$ is expected to be larger than 0 kJ/mol. Eq. (3.3) has then to be extended:

$$\mathrm{FS}_{\mathrm{pseudoknot}} = \frac{\mathrm{e}^{-\frac{\Delta G - \Delta\Delta G}{k_B T}}}{1 + \mathrm{e}^{-\frac{\Delta G - \Delta\Delta G}{k_B T}}}\,. \tag{3.5}$$

4. Bayes' Theorem and Metropolis Algorithm

I employ Bayes' theorem in this work to use information from observed data in order to update available knowledge about parameters in a statistical model. For two events A and B, Bayes' theorem can be derived with a simple multiplicative rule of probability [56]:

$$P(A \cap B) = P(A)P(B|A)\,, \tag{4.1}$$

$$P(A \cap B) = P(B)P(A|B)\,. \tag{4.2}$$

Here, $P(A \cap B)$ is the probability of both, events A and B, being true. The marginal probabilities $P(A)$ and $P(B)$ are the probability for event A and B, respectively, to be true. The conditional probability $P(B|A)$ is the probability for event B to be true, under the condition that event A is true, for $P(A|B)$, this is reversed [56]. Rearranging Eq. (4.1) and Eq. (4.2) yields Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}\,, \tag{4.3}$$

if $P(B) \neq 0$. Instead of the events A and B, one can consider a data set „Data“ and model parameters „Parameters“. Then Bayes' theorem reads

$$P(\text{Parameters}|\text{Data}) = \frac{P(\text{Parameters})P(\text{Data}|\text{Parameters})}{P(\text{Data})}\,. \tag{4.4}$$

Since the marginal probability $P(\text{Data})$ is a normalizing factor and does not depend on the „Parameters“ it is often left out:

$$P(\text{Parameters}|\text{Data}) \propto P(\text{Parameters})P(\text{Data}|\text{Parameters})\,. \tag{4.5}$$

Associated with Bayes' theorem, the contributing terms of this proportionality have certain names, such that Eq. (4.5) is in words: given the observed data, the posterior probability $P(\text{Parameters}|\text{Data})$ is proportional to the prior probability $P(\text{Parameters})$ times the likelihood $P(\text{Data}|\text{Parameters})$ [56–58]. The prior probability $P(\text{Parameters})$ contains the prior knowledge of one or more given parameters of a statistical model. The likelihood function $P(\text{Data}|\text{Parameters})$ is the conditional probability of obtaining the input data, given the statistical model. The posterior probability density $P(\text{Parameters}|\text{Data})$ is the conditional probability of the parameters, given the data. It can be interpreted as „updated knowledge“, since it includes the information from both the prior and the obtained data. The posterior probability can be used to make predictions, but its functional form can be very complicated and high-dimensional [56–58]. Eq. (4.5) allows to compute the posterior as the product of prior and likelihood up to a proportionality constant. Moreover, the high number of parameters considered in this work translates into very complicated and high-dimensional posterior and likelihood functions. That is why, I employ the Metropolis algorithm which estimates the posterior densities of the parameters. It works as follows [57, 59, 60]:

Let f be a function proportional to the posterior probability from which random samples are to be obtained by the algorithm. Additionally, let x be the parameter (or vector of parameters) for which a probability density is to be obtained. First, an arbitrary initial value x_{old} is set and the probability of the initial value $p_{\text{old}} = f(x_{\text{old}})$ is determined. Afterwards, a large number of iteration steps are executed, each undergoing the following steps:

1. A random next candidate x_{new} is sampled from a proposal density (e.g. a Gaussian distribution with mean value x_{old}).

2. The probability $p_{\text{new}} = f(x_{\text{new}})$ is determined.

3. An acceptance ratio $\alpha = \frac{p_{\text{new}}}{p_{\text{old}}}$ is calculated.

4. A uniform random number $u \in [0, 1]$ is generated:
    a) If $u \leq \alpha$: the new value x_{new} is accepted and $p_{\text{old}} = p_{\text{new}}$ and $x_{\text{old}} = x_{\text{new}}$ are set.
    b) If $u > \alpha$: the candidate is rejected. Thus, p_{old} and x_{old} stay unchanged.

In other words, if the probability p_{new} of the candidate x_{new} is larger than the old one p_{old}, the candidate ΔG_{new} will always be accepted. If p_{new} is smaller than p_{old}, the candidate x_{new} will sometimes be accepted, depending on the α and u. After a large number of iteration steps x will converge and the accepted parameters yield a probability density of x or, if x contains more than one parameter, a probability density for each parameter.

5. Experimental Background

The data I use in the thesis is provided by Mikl et al. [9]. The authors assessed the frameshifting potential of more than 12,000 synthetic oligonucleotide involved in PRF, i.e. short polymers of nucleotides. The tested oligonucleotides included viral, bacterial, and human wild-type sequences and variants obtained by systematically introducing mutations into the wild-type sequences. The mutations were located either in the slippery sequence or in its proximity. In this thesis, I considered only sequences that varied in the slippery sequence and otherwise corresponded to the wild-type sequence. The following subsections describe the measurement method of frameshifting potential used by Mikl et al. [9]. In particular, the first subsection illustrates the frameshifting reporter construct. The second subsection describes how the method of fluoresce activated cell sorting (FACS) was used to assess the frameshifting potential. Finally, the method that was used to measure the background noise in the fluorescence signal is reported.

5.1. Frameshifting Reporter Construct

To investigate a variants' ability to induce a frameshift, a frameshifting reporter construct was developed. This construct was then introduced in human cells and processed by the human translation machinery. As illustrated in Fig. 7, the construct contains the sequence encoding the red fluorescent protein mCherry, followed by one of the tested oligonucleotides, and by the sequence encoding the green fluorescent protein (GFP). To investigate the -1 frameshifting, the GFP sequence is positioned $+1$ nucleotide downstream to the oligonucleotide sequence that is to be tested [9].

Figure 7: Frameshifting reporter construct: the coding sequence of mCherry is followed by the tested sequence, one nucleotide (N), and the coding sequence of GFP [9, modified version].

Only if -1 frameshifting occurs in the oligonucleotide sequence, the GFP coding sequence will be in frame and GFP will be synthesized. If frameshifting does not occur, the GFP coding sequence will not be in frame and GFP will not be synthesized. Thus, the -1 frameshifting potential can be quantified by the number of synthesized GFPs, which can be measured through the intensity of green fluorescence, emitted by GFP when exposed to blue or ultraviolet light [61]. The protein mCherry emits red fluorescence, which is monitored to check whether the construct was translated or not. Since the mCherry coding sequence is placed before the slippery sequence, it is synthesized before the frameshifting and, therefore, independent of the occurrence of a frameshift [9].

5.2. Fluorescence Activated Cell Sorting and Obtained Percent GFP Fluorescence

To make use of the human translation process the frameshifting reporter constructs were inserted into human cells, such that every cell contained constructs with one oligonucleotide variant and every variant had the same genomic environment. Experiments were carried out at a temperature of 310 K. The mCherry-positive cells were sorted by Fluorescence Activated Cell Sorting (FACS), which is a technique derived from flow cytometry. During flow cytometry, a large number of cells pass one after the other through one or more laser beams. Detectors measure scattered light from different angles and fluorescence emissions [62]. In addition to flow cytometry, during FACS the cells are also physically sorted according to their measured fluorescence. Specifically, in the experiments from Mikl et al., the intensity of green fluorescence of the mCherry positive cells was first measured in a setup similar to flow cytometry. After that, the cells were charged according to their green fluorescence intensity and sorted into 16 bins by an electric field. The DNA of the cells in each bin was then sequenced to determine the distribution of each variant across bins. Subsequently, the median of the log2 fluorescence for all cells in a bin of a certain variant was calculated and used as the fluorescence value associated with this bin and this variant. The resulting distribution for each variant across all bins was smoothed. For the variants with one or more peak in their distribution the weighted average was determined. The obtained value is called GFP fluorescence in the following as it was done in the paper by Mikl et al. [9]. To assign a percentage of GFP expression to every variant, the authors set the lowest obtained GFP fluorescence value to 0 % and the highest ob-

tained GFP fluorescence value to 100 % [9]. In the following, I will refer to these values as percent GFP fluorescence and I will call the set of all measured percent GFP fluorescence values M_{all}. These percentages must not be confused with the frameshifting efficiency. The connection of these two observables is discussed in Section 7.3.

5.3. Background Fluorescence

The measured values in M_{all} contain a background fluorescence caused by autofluorescence of the cells, which is a natural emission of light by components of the cell when they are excited by light with a suitable wavelength [63]. To measure this background noise, Mikl et al. adjusted the frameshifting reporter construct as depicted in Fig. 8.

Figure 8: Frameshift reporter construct to measure background noise. A stop codon was added directly after the coding sequence of mCherry [9, modified version].

With the added stop codon, the tested sequence, as well as the sequence of GFP, cannot be translated. Thus, GFP fluorescence coming from a ribosomal frameshift is excluded [9]. The measurements of the background fluorescence are also provided in the data set, such that I can treat the background fluorescence and the percent GFP fluorescence, resulting from a ribosomal frameshift, separately in the following.

6. Methods

In this section I present my methods to estimate the effect of different downstream structures on the free-energy difference between the 0 and the -1 frame. I start with my method to select the input measurements, where I explain how I pre-process the data from Mikl et al. [9]. Afterwards, I demonstrate, step by step, my methods to determine probability densities for the free-energy differences of selected sequences from the data set. Finally, I introduce my method to calculate differences of free-energy differences, which will be the foundation of discussing how different secondary structures affect the free energies.

6.1. Selection of the Input Measurements

In the case that a specific variant (sequence) was tested several times, there are multiple measured percent GFP fluorescence values for this variant. In the following, for each variant I will refer to the set of values obtained for the variant as M_{V}. For many variants, M_{V} contains only one value. However, for wild-type variants, M_{V} usually consists of a lot

more, sometimes around 30 values. For several cases, these sets M_{V} contain individual values, that differ dramatically from most of the other values in the set. To systematically identify and exclude these outliers, I use the interquartile range.

The interquartile range is the difference of two values called the quartiles Q_1 and Q_3. The lower quartile Q_1 is defined as the value below which 25 % of the measured values in M_{V} are. Above the upper quartile Q_3 are 25 % of the measured values in M_{V}. The interquartile range is given by [64]

$$Q = Q_3 - Q_1 \,. \tag{6.1}$$

Here, I define outliers as values below $Q_1 - 1.5Q$ or above $Q_3 + 1.5Q$. These values are then excluded in the further use of M_{V}.

Additionally, I exclude measurements that have more than one peak in their log2 fluorescence distribution across bins (see Section 5.2), since I do not expect to have two different GFP expression levels for the same sequence and, thus, I expect these multimodal measurements to be associated to measurement errors.

6.2. Determination of Free-Energy Differences

This section introduces my methods to determine the probability density of free-energy differences ΔG for the tested sequence variants. Here, $P(\Delta G|M_{\mathrm{V}})$ is the conditional probability density for a ΔG given a data set M_{V}. Maximizing this function or a function proportional to $P(\Delta G|M_{\mathrm{V}})$ yields the most probable value for ΔG given a data set M_{V}. At the same time, I aim at obtaining a probability density of ΔG, which entails the uncertainty. As explained in section 4, the functional form of $P(\Delta G|M_{\mathrm{V}})$ is not trivial. Therefore, I apply Bayes' theorem and multiply the likelihood $P(M_{\mathrm{V}}|\Delta G)$ and the prior $P(\Delta G)$ to obtain a function, that I call $P_{M_{\mathrm{V}}}$, which is proportional to the posterior $P(\Delta G|M_{\mathrm{V}})$. I will here report how I derive likelihood and prior to compute $P_{M_{\mathrm{V}}}$. I will then describe how I obtained the probability densities of the free-energy differences by applying the Metropolis algorithm which is introduced in section 4.

6.2.1. Bayes' Theorem Applied on Free-Energy Differences

My aim is to derive a function $P_{M_{\mathrm{V}}}$, which is proportional to the posterior probability $P(\Delta G|M_{\mathrm{V}})$. For this purpose, I use Bayes' theorem (see Section 4):

$$P(\Delta G|M_{\mathrm{V}}) \propto P(\Delta G)P(M_{\mathrm{V}}|\Delta G) = P_{M_{\mathrm{V}}}(\Delta G) \,. \tag{6.2}$$

For the prior $P(\Delta G)$, I make use of previously obtained knowledge to set a boundary

for ΔG. In the paper by Bock et al. the highest ΔG determined while changing two A-site and P-site base pairs simultaneously is around 10 kJ/mol [8]. Since more than three base pairs cannot be changed simultaneously in the A-site and P-site, I choose the highest possible ΔG of $3 \cdot 10$ kJ/mol = 30 kJ/mol as an upper boundary and -30 kJ/mol as a lower boundary. Hence, I define the interval $I = [-30 \text{ kJ/mol}, 30 \text{ kJ/mol}]$, such that the prior function is the indicator function

$$P(\Delta G) = \chi_I(\Delta G) = \begin{cases} 1, & \text{if } \Delta G \in I\,, \\ 0, & \text{if } \Delta G \notin I\,. \end{cases} \tag{6.3}$$

The likelihood function $P(M_{\mathrm{V}}|\Delta G)$ represents the conditional probability density of a data set M_{V} given a ΔG. This function is not as simple as the prior function and is derived in the subsequent Section 6.2.2.

6.2.2. Derivation of the Likelihood Function

I intend to obtain a conditional probability density function (likelihood function) of a data set M_{V} given a free-energy difference ΔG. Therefore, I assume that the percent GFP fluorescence values given in M_{all} consist of the random variable B representing the background fluorescence and a random variable S representing the percent GFP fluorescence resulting from a frameshift signal. Hence, the measured value given in M_{all} is a random variable $M = B + S$. These values I assume to be random, because firstly, the autofluorescence in B fluctuates randomly (due to different cell components and the fluctuating amount of light the components absorb) and secondly, B and S fluctuate, due to the limited accuracy of the measurement process of percent GFP fluorescence values. On the random variables B and S, I apply the central limit theorem (CLT). This is possible, since in B a large number of independent molecules contribute to the autofluorescence and I assume that every molecule exhibits the same probability density to exhibit fluorescence. For S (and additionally for B) this is possible, because of the large number of frameshifting events in each cell contributing to the fluorescence measurement. The CLT states that the probability densities of B and S are well approximated by a Gaussian distribution [65]:

$$f_B(b) = \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left(-\frac{(b-\mu_B)^2}{2\sigma_B^2}\right), \tag{6.4}$$

$$f_S(s) = \frac{1}{\sqrt{2\pi\sigma_S^2}} \exp\left(-\frac{(s-\mu_S)^2}{2\sigma_S^2}\right), \tag{6.5}$$

with values, i.e. observed measurements, b and s of the random variables B and S, mean values μ_B and μ_S, and standard deviations σ_B and σ_S.

Assuming that the background fluorescence B and the fluorescence resulting only from a frameshift signal S are independent, the probability density function of M is

$$f_M = f_B * f_S \,, \tag{6.6}$$

where $*$ denotes the convolution operator [65]. With the Convolution Theorem [66]

$$\mathcal{F}\{f_B * f_S\}(z) = \mathcal{F}\{f_B\}(k) \cdot \mathcal{F}\{f_S\}(k) \,, \tag{6.7}$$

Eq. (6.6) can simply be calculated by multiplying two Fourier transforms point-wise and applying the inverse Fourier transform. The calculation can be found in Appendix A. The result is again a Gaussian distribution, where m is the value of the random number M, i.e. an observed measurement from M_{all}:

$$f_M(m) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + \sigma_S^2}} \exp\left(-\frac{(m - \mu_B - \mu_S)^2}{2(\sigma_B^2 + \sigma_S^2)} \right) . \tag{6.8}$$

The mean value μ_B and the standard deviation σ_B can be calculated with the given background fluorescence measurements from Mikl et al. (see Section 5.3) [9]. The method is described in Section 6.2.3. The mean value of percent GFP fluorescence resulting from a frameshift signal μ_S is related to the expected frameshifting efficiency, which can be described by the thermodynamic model $f_{\text{model}}(\Delta G)$ from Eq. (3.3). Thus, $\mu_S = \mu_S(\Delta G)$ and consequently $f_M = f_M(m, \Delta G)$. Because the provided measurements by Mikl et al. [9] are in the dimension „percent GFP fluorescence“, $f_{\text{model}}(\Delta G)$ has to be converted from the dimension „frameshifting efficiency“ into the dimension „percent GFP fluorescence“ to equal μ_S. In Section 6.2.5, I explain the methods to derive a term for μ_S, in detail. Section 6.2.4 demonstrates the methods to determine a term for σ_S.

Eq. (6.8) is a probability density function for one m in M_{all}. The probability density of the ΔG values associated to a specific sequence is computed for each sequence separately. To that aim, I compute the product of all the $f_M(m, \Delta G)$ obtained from the M_{V} set corresponding to the considered sequence. Each one of the $f_M(m, \Delta G)$ is computed from one m measurement in the M_{V} set:

$$P(M_{\text{V}}|\Delta G) = \Pi_{m \in M_{\text{V}}} f_M(m, \Delta G) \,, \tag{6.9}$$

which is the conditional probability density of a data set M_{V} given a ΔG and therefore the required likelihood function.

Inserting the likelihood function Eq. (6.9) and the prior function Eq. (6.3) into Bayes'

theorem Eq. (6.2) yields

$$P(\Delta G|M_{\mathrm{V}}) \propto \chi_I(\Delta G) \cdot \Pi_{m \in M_{\mathrm{V}}} f_M(m, \Delta G) = P_{M_{\mathrm{V}}}(\Delta G)\,. \tag{6.10}$$

6.2.3. Determination of the Mean Background Fluorescence μ_B and its Standard Deviation σ_B

Since Mikl et al. provided the obtained measurements for the background fluorescence, the mean background fluorescence and its standard deviation can simply be calculated from the given data [9]. To that aim, I fit the Gaussian distribution from Eq. (6.4) (was derived from the CLT in Section 6.2.2) on a histogram of the measurements. However, these measurements contain values that are significantly larger than most of the other measurements. That is why, it might be reasonable to select the input measurements by employing the interquartile range (see Section 6.1) before determining μ_B and σ_B. Both ways, selecting input measurements and taking all measurements into account, I apply in Section 7.1.

6.2.4. Determination of the Relation between Variance of GFP Fluorescence and Mean GFP Fluorescence

As defined in Section 6.2.2, S is the percent GFP fluorescence resulting only from a frameshifting signal. In this section I describe a method to determine σ_S, which is not as trivial to calculate as μ_B or σ_B because the percent GFP fluorescence values provided by Mikl et al. contain the background fluorescence B (see Section 6.2.2). Thus, there is no data available that exclusively yields the percent GFP fluorescence resulting only from a frameshifting signal S [9].

A simple possibility would be to sample probability densities for both σ_S and ΔG, for each variant separately, with the Metropolis algorithm. This, however, leads to convergence problems when considering M_{V} sets that contain only one value. In this case, it is like trying to solve a system of equations with more parameters than equations: one cannot determine every single parameter, but only their combination, as the parameters are dependent from one another. Hence, there is no chance of convergence for two variables when M_{V} contains only one value. Thus, in order to reduce the number of variables, I test if, in the measurements, there is a relation between σ_S and μ_S from which σ_S can be determined. To that aim, I plot σ_S^2 against μ_S in Section 7.2. I chose to plot the variance σ_S^2 and not the standard deviation σ_S, since in the further course I will obtain higher probabilities when considering a linear relation between σ_S^2 and μ_S instead of a linear relation between σ_S and μ_S.

For the plot, I need probability densities for μ_S and σ_S^2 to consider their uncertainties.

Thus, I use Bayes' theorem and the Metropolis algorithm varying μ_S and σ_S as parameters. Therefore, μ_S and σ_S now correspond to the x from Section 4. In order for μ_S and σ_S to have a narrow density, i.e. low uncertainty, I run the algorithm only over the measurement sets M_V that contain many values. The largest numbers of measurements were obtained for the wild-type variants, thus I use the 12 sets M_V corresponding to the sequences of the following wild types: RSV, CCR5, HTLV, PLRV, PEG10 (retrotransposon-derived (see Section 2.5) human protein [67]), HIV, SIV, WNV, SRV, HERV, SARS, and PRRSV (Porcine reproductive and respiratory syndrome virus).

The probability density f from Section 4 corresponds now to P_{M_V}:

$$P_{M_\mathrm{V}}(\mu_S, \sigma_S^2) = \chi_2(\mu_S, \sigma_S^2) \cdot \Pi_{m \in M_\mathrm{V}} f_M(m, \mu_S, \sigma_S^2) \tag{6.11}$$

with

$$f_M(m, \mu_S, \sigma_S^2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + \sigma_S^2}} \exp\left(-\frac{(m - \mu_B - \mu_S)^2}{2(\sigma_B^2 + \sigma_S^2)} \right) \tag{6.12}$$

from Eq. (6.8) and μ_B and σ_B inserted from Section 7.1. The prior probability is modeled by the indicator function χ_2:

$$\chi_2(\mu_S, \sigma_S^2) = \begin{cases} 1, & \text{if } \mu_S \in [0\,\%, 100\,\%] \wedge \sigma_S^2 \geq 0\,, \\ 0, & \text{else}\,. \end{cases} \tag{6.13}$$

The indicator function has this form, because the mean value μ_S, as a percentage (percent GFP fluorescence), has to be between 0 % and 100 % and σ_S^2, as a variance, has to be greater than or equal to zero, per definition. When applying the Metropolis algorithm, I use two proposal functions for μ_S and σ_S, namely, two different Gaussian distributions with mean values μ_{S_old} and $\sigma_{S_\mathrm{old}}^2$ and standard deviations σ_μ and σ_σ. I choose σ_μ and σ_σ in order to get an acceptance rate between 20 % and 80 % as this usually results in a good convergence. Based on the results that are shown in Section 7.2, I assume a linear relation between μ_S and σ_S^2 with slope m_σ:

$$\sigma_S^2 = m_\sigma \cdot \mu_S\,, \tag{6.14}$$

which I use in the Metropolis algorithm to determine free-energy differences (see Section 6.2.6).

6.2.5. Determination of the Relation between Mean GFP Fluorescence and Frameshifting Efficiency

As explained in Section 5.2 and Section 6.2.2, the mean value of percent GFP fluorescence μ_S depends on the frameshifting efficiency $f_{\text{model}}(\Delta G)$, but is not assumed to be identical. To get a relation between the two units, I will plot μ_S in „percent GFP fluorescence“ obtained from the Mikl et al. [9] data for several variants against the frameshifting efficiencies of the corresponding variants independently reported in the literature. To that aim, I take frameshifting efficiencies from several papers into account: Biswas et al. [17], Dulude et al. [7], and Léger et al. [68] determined frameshifting efficiencies for coding sequences, for which also Mikl et al. [9] reported percent GFP fluorescence values. Since some of these data sets, which will be the input measurements, contain only one value, I again I cannot use both μ_S and σ_S^2 as parameters, but I can employ the gained relation between σ_S^2 and μ_S (Eq. (6.14)) and only use μ_S as a free parameter. Therefore, I once more use P_{M_V} as f from Section 4 and the Metropolis algorithm:

$$P_{M_\text{V}}(\mu_S) = \chi_3(\mu_S) \cdot \Pi_{m \in M_\text{V}} f_M(m, \mu_S)\,, \tag{6.15}$$

with

$$f_M(m, \mu_S) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + m_\sigma \cdot \mu_S}} \exp\left(-\frac{(m - \mu_B - \mu_S)^2}{2(\sigma_B^2 + m_\sigma \cdot \mu_S)}\right) \tag{6.16}$$

from Eq. (6.8) and μ_B and σ_B inserted from Section 7.1. The prior probability is modeled by the indicator function χ_3:

$$\chi_3(\mu_S) = \begin{cases} 1, & \text{if } \mu_S \in [0\,\%, 100\,\%]\,, \\ 0, & \text{else}\,. \end{cases} \tag{6.17}$$

The indicator function again results from μ_S being a percentage. Running the algorithm successively with all data sets available (eleven in total) gives a probability density of μ_S for each variant, which I then compare with the corresponding frameshifting efficiency. Based on the results that are shown in Section 7.3, I assume a linear relation between mean value μ_S and frameshifting efficiency $f_{\text{model}}(\Delta G)$ with slope m_μ:

$$\mu_S = m_\mu \cdot f_{\text{model}}(\Delta G)\,. \tag{6.18}$$

I employ the obtained dependency in the Metropolis algorithm to determine free-energy differences (see Section 6.2.6).

6.2.6. Determination of Frameshifting Free-Energy Differences with the Metropolis Algorithm

To sample a probability density for the free-energy difference of a certain variant $\tilde{V}$ (tested sequence), I employ the Metropolis algorithm introduced in Section 4. This is then repeated in order to obtain ΔG densities for each variant. For the function f in Section 4, I could just insert Eq. (6.14) and Eq. (6.18) with the obtained values for m_σ and m_μ from Sections 7.2 and 7.3 into P_{M_V} from Eq. (6.10). Then, the parameter that was called x in Section 4 would be only ΔG. I would use a Gaussian distribution with mean value ΔG_old and a standard deviation σ, which I would choose in order to get an acceptance rate between 20 % and 80 % as this usually results in a good convergence.

However, instead of using only the ΔG values as free parameters, I extend the algorithm to also include m_σ and m_μ as free parameters. This has the advantage that the algorithm always optimizes the probability densities of m_σ and m_μ given all measurements of the used wild-type sequences and all measurements of the tested sequence. Hence, I adjust the algorithm, in such a way that in each step it does not only generate a new candidate for ΔG, but a new set of candidates for 15 parameters. The set contains ΔG for the tested variant $\tilde{V}$, m_σ, m_μ, and the ΔG_i values for the 12 wild-type sequences. New candidates of these 15 parameters are calculated from 15 different Gaussian distributions: their mean value is the old candidate and their standard deviation σ of each Gaussian is set before running the algorithm and can be adjusted to get an acceptance rate between 20 % and 80 %. Furthermore the likelihood is extended, so that it consists of three parts:

1. The first part of the algorithm takes the percent GFP values of the data set $M_{\tilde{V}}$ and its purpose is to yield a probability density for ΔG. Its likelihood function is

$$L_{1,M_{\tilde{V}}}(\Delta G, m_\sigma, m_\mu) = \Pi_{m \in M_{\tilde{V}}} f_M(m, \Delta G, m_\sigma, m_\mu) \tag{6.19}$$

   with

$$f_M(m, \Delta G, m_\sigma, m_\mu) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + f_\mathrm{model}(\Delta G) \cdot m_\mu \cdot m_\sigma}} \exp\left(-\frac{(m - \mu_B - f_\mathrm{model}(\Delta G) \cdot m_\mu)^2}{2(\sigma_B^2 + f_\mathrm{model}(\Delta G) \cdot m_\mu \cdot m_\sigma)} \right). \tag{6.20}$$

2. The purpose of the second part of the likelihood is to obtain a probability density for m_σ. To this aim, I used the same data sets used in Section 6.2.4 and containing the percent GFP values for 12 wild-type sequences. These data sets are here called $M_{\mathrm{WT},i}$ with $i = 1, 2, ..., 12$. This part of the likelihood takes all data sets $M_{\mathrm{WT},i}$ into account and depends on the ΔG_i for each wild-type sequence i:

$$L_{2,M_\mathrm{WT}}(m_\sigma, m_\mu, \Delta G_1, ..., \Delta G_{12}) = \Pi_{i=1}^{12} \Pi_{m \in M_{\mathrm{WT},i}} f_M(m, \Delta G_i, m_\sigma, m_\mu) \tag{6.21}$$

with

$$f_M(m, \Delta G_i, m_\sigma, m_\mu) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + f_{\text{model}}(\Delta G_i) \cdot m_\mu \cdot m_\sigma}} \exp\left(-\frac{(m - \mu_B - f_{\text{model}}(\Delta G_i) \cdot m_\mu)^2}{2(\sigma_B^2 + f_{\text{model}}(\Delta G_i) \cdot m_\mu \cdot m_\sigma)}\right). \tag{6.22}$$

3. To estimate the relation between percent GFP fluorescence and frameshifting efficiency, I included the eleven sequences, whose frameshifting efficiencies are established by Biswas et al. [17], Dulude et al. [7], and Léger et al. [68] and whose percent GFP values are given by Mikl et al. [9] (see Section 6.2.5). The stated frameshifting efficiencies I call FS_j with $j = 1, 2, ..., 11$. The data set with the measured percent GFP values of sequence j is called $M_{\text{FS},j}$. Thus, the third and last part of the likelihood has as input data all FS_j and $M_{\text{FS},j}$ and its purpose is to obtain a probability density for m_μ:

$$L_{3,M_{\text{FS}}}(m_\sigma, m_\mu) = \Pi_{j=1}^{11} \Pi_{m \in M_{\text{FS},j}} f_M(m, FS_j, m_\sigma, m_\mu) \tag{6.23}$$

with

$$f_M(m, FS_j, m_\sigma, m_\mu) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + FS_j \cdot m_\mu \cdot m_\sigma}} \exp\left(-\frac{(m - \mu_B - FS_j \cdot m_\mu)^2}{2(\sigma_B^2 + FS_j \cdot m_\mu \cdot m_\sigma)}\right). \tag{6.24}$$

In each f_M above I insert the mean value μ_B and standard deviation σ_B from Section 7.1. Finally, I multiply the product of the three parts of the likelihood with the prior function $\chi(\Delta G, \mu_S, \sigma_S^2, \Delta G_1, ..., \Delta G_{12})$ to get a the function that is proportional to the probability density of the whole set of parameters:

$$\begin{aligned} P_{M_{\tilde{V}}, M_{\text{WT}}, M_{\text{FS}}}(\Delta G, m_\sigma, m_\mu, \Delta G_1, ..., \Delta G_{12}) = &\chi(\Delta G, \mu_S, \sigma_S^2, \Delta G_1, ..., \Delta G_{12}) \cdot L_{1,M_{\tilde{V}}}(\Delta G, m_\sigma, m_\mu) \\ &\cdot L_{2,M_{\text{WT}}}(m_\sigma, m_\mu, \Delta G_1, ..., \Delta G_{12}) \cdot L_{3,M_{\text{FS}}}(m_\sigma, m_\mu). \end{aligned} \tag{6.25}$$

The prior function is defined as

$$\chi(\Delta G, \mu_S, \sigma_S^2, \Delta G_1, ..., \Delta G_{12}) = \begin{cases} 1, & \text{if } \mu_S \in [0\,\%, 100\,\%] \wedge \sigma_S^2 \geq 0 \wedge \Delta G \in I \wedge \Delta G_i \in I \text{ for } i = 1, ..., 12, \\ 0, & \text{else} \end{cases} \tag{6.26}$$

with $I = [-30\,\text{kJ/mol}, 30\,\text{kJ/mol}]$ as described in Section 6.2.1. I assume the set of parameters to be converged when the logarithm of p_{new} (see Section 4) is above 1100, because then, the trajectories of the parameters show a convergent behaviour. Once p_{new} has reached 1100, I execute 10,000 more iterations and obtain probability densities from the last 9,000 steps.

With this algorithm, I sample one probability density for the free-energy difference ΔG for each of the sequences tested by Mikl et al. [9] that are from the nine included viruses

introduced in Section 2.5 and that are obtained either from the wild type or by introducing variations only on the slippery sequence. In total, I determine free-energy differences for 568 different sequences. Each of the 568 probability densities of ΔG, I sample with two different sets of initial values, to confirm, that the free-energy differences converge to the same value. Thus, in total, I sample 1136 ΔG probability densities. The results are reported in Section 7.4.

6.3. Determination of the Differences between Frameshifting Free-Energy Differences of Sequences with Different Downstream Secondary Structures

In Section 2.4 I hypothesized that a pseudoknot reduces the free-energy difference ΔG between the 0 and the -1 frame. Thus, as mentioned in Section 3.2, the difference between the ΔG obtained from sequences that contain a stem loom and the ΔG of sequences with a pseudoknot the free-energy difference of sequences should then be larger than zero. To test my hypothesis, I determine, from the data of Mikl et al. [9], $\Delta\Delta G$ probability densities for many pairs of sequences exhibiting different secondary structures. In this context, I compare all sequences with a pseudoknot to each of the sequences with a stem loop (HIV and SIV). Thereby, I compare two viruses with different secondary structures at a time. To compare for example HIV (stem loop) and SARS (pseudoknot), I choose (from the data set M_{all}) pairs of sequences containing the same slippery sequence and the upstream and downstream region of the corresponding wild type of HIV and SARS, respectively. For each of the two compared sequences, I determine the probability density for the frameshifting free-energy difference ΔG as in Section 6.2.6. Then, I calculate, for each pair of sequences, the difference $\Delta\Delta G = \Delta G(\text{HIV}) - \Delta G(\text{SARS})$. Since the slippery sequence is the same in both sequences, the base-pair free-energy differences cancel out in the subtraction and only the free-energy difference resulting from the different secondary structures ($\Delta\Delta G$) remains. Repeating this procedure for all of the selected pairs of sequences yields a probability density of $\Delta\Delta G$ for the compared viruses. Additionally, I repeat this with all pairs of the listed viruses containing different secondary structures. The results are in Section 7.5.

To compare the effect of the secondary structure on different viruses, I combine all $\Delta\Delta G$ probability densities for each comparison between two viruses. This is possible, because I expect the $\Delta\Delta G$ to depend only on the secondary structure, thus, $\Delta\Delta G$ is expected to be the same for different slippery sequences. My method to combine the densities is the following: to each localized probability density (a density without an upper or lower boundary) I fit a Gaussian distribution. Then, I determine the product of all Gaussian distributions to obtain the joint probability density for each pair of viruses. The resulting

plot is in Section 7.5 (Fig. 20).

7. Results

This section reports on the results that are obtained by following the steps described in the methods in order to obtain the probability densities of the differences $\Delta\Delta G$ between the frameshifting free-energy differences of mRNA sequences containing different downstream secondary structures (either stem-loop or pseudoknot). As described in Section 6.2.6, in order to obtain the ΔG probability densities corresponding to each selected sequence, I use the Metropolis algorithm to sample from the probability density $P_{M_{\tilde{V}},M_{\mathrm{WT}},M_{\mathrm{FS}}}$ introduced in Eq. (6.25). When computing $P_{M_{\tilde{V}},M_{\mathrm{WT}},M_{\mathrm{FS}}}$ several parameters must be taken into account in addition to ΔG: μ_B, σ_B, μ_S, and σ_S. Thus, I first determine μ_B and σ_B from the measured background fluorescence, as described in Section 6.2.3. Next, I employ the methods explained in Section 6.2.4 to explore the relation between the variance σ_S^2 and the mean value μ_S. After that, I investigate the relation between μ_S and the frameshifting efficiency, as described in Section 6.2.5, to include it in $P_{M_{\tilde{V}},M_{\mathrm{WT}},M_{\mathrm{FS}}}$. Finally, I sample probability densities of ΔG of the considered sequences as described in Section 6.2.6. As explained in Section 6.3, I then determine probability densities for $\Delta\Delta G$.

7.1. Determination of the Mean Background Fluorescence μ_B and its Standard Deviation σ_B

As explained in Section 5.1 any measurement is affected by the presence of background fluorescence due to autofluorescence of the cell. In Section 6.2.2, I showed that this effect has to be taken into account when analyzing the signals. The two ways to determine mean value μ_B and standard deviation σ_B of the background fluorescence measurements given by Mikl et al. [9] are explained in Section 6.2.3: I either determine both values directly from the measurements or I exclude the outliers (see Section 6.1) first. A histogram of the background percent GFP fluorescence values is shown in Fig. 9a. Excluding the outliers from the background fluorescence yields Fig. 9b.

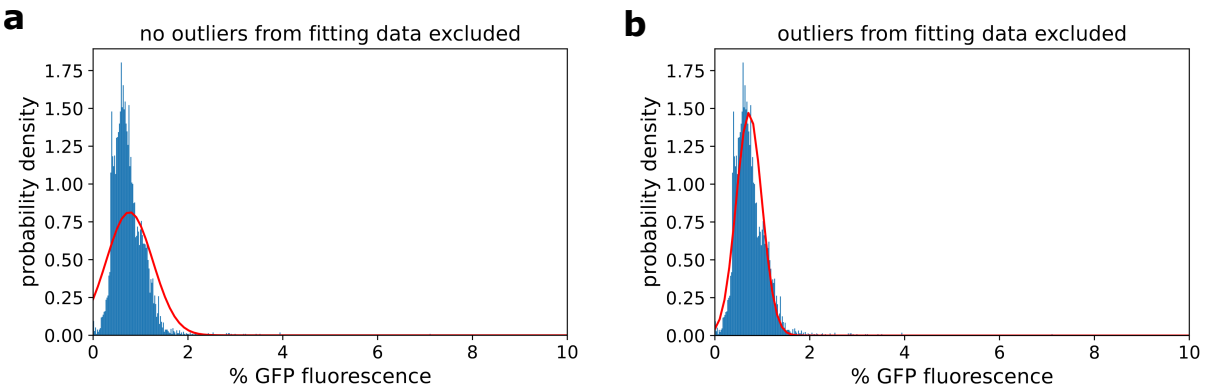

Figure 9: Histograms (blue) of the background percent GFP fluorescence of all variants (tested sequences) is depicted [9]. (**a**) A Gaussian (red) is plotted with mean value and standard deviation of all measurements, i.e. no outliers are excluded from the fitting data. (**b**) A Gaussian (red) is plotted with mean value and standard deviation of the remaining measurements after excluding outliers.

The Gaussian in Fig. 9b depicts a better approximation to the background noise than the Gaussian in Fig. 9a. That is why, I will use mean value and standard deviation of the selected measurements (method with interquartile range applied) in my further analysis:

$$\mu_B \approx 0.721\,\%\,, \tag{7.1}$$

$$\sigma_B \approx 0.271\,\%\,. \tag{7.2}$$

7.2. Determination of the Relation between Variance of GFP Fluorescence and Mean GFP Fluorescence

In addition to the background fluorescence, the measured values from Mikl et al. consist of the percent GFP fluorescence which results from a frameshift signal [9]. For this observable I also need standard deviation and mean value, which I introduced in Section 6.2.2. In this section, I derive a relation between the standard deviation σ_S and the mean value μ_S. To this aim, I first obtain the probability densities for both parameters as explained in Section 6.2.4. Then, I plot the variance σ_S^2 against the mean value μ_S. The resulting plot is shown in Fig. 10.

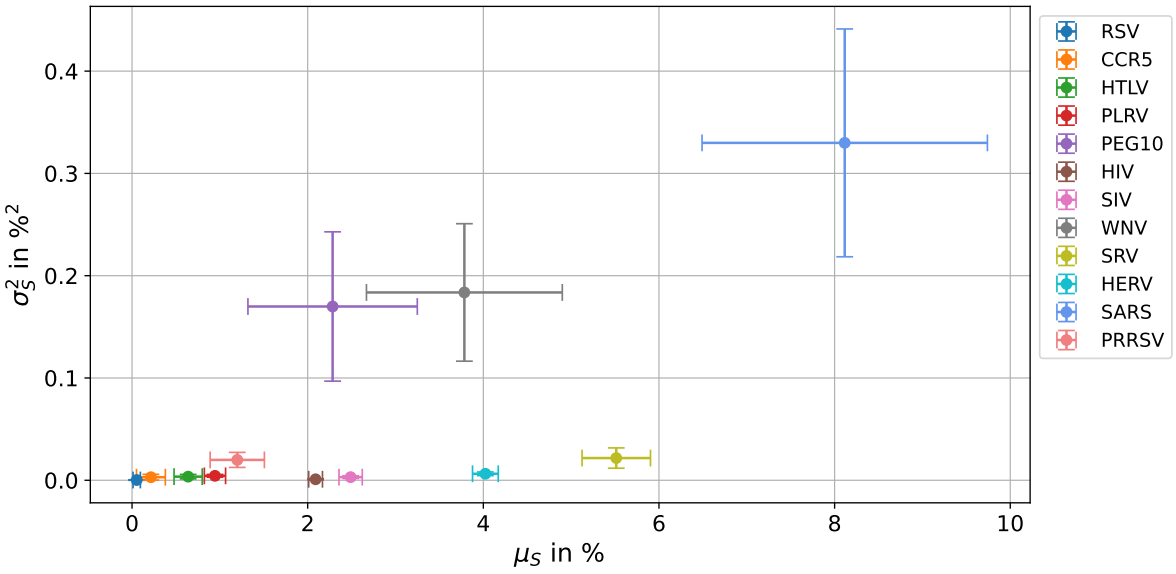

Figure 10: The variance σ_S^2 is plotted against the mean value μ_S for 15 different coding sequences of wild-type viruses (and human proteins). The data points represent the mean value and the error bars represent the standard deviation of the obtained probability densities for σ_S^2 and μ_S.

The plot does not show a clear dependency between variance and mean value. That is why, I assume the simplest case, which is a linear dependency:

$$\sigma_S^2 = m_\sigma \cdot \mu_S \,. \tag{7.3}$$

A probability density for the slope m_σ is derived by the Metropolis algorithm (see Section 6.2.6). To make sure, that for my purpose a linear dependency between σ_S^2 and μ_S is a better approximation than a linear dependency between σ_S and μ_S, I have run the algorithm with both relations. As mentioned in section Section 6.2.4, the assumption of a linear dependency between σ_S^2 and μ_S lead to higher values of $P_{M_{\tilde{V}},M_{\mathrm{WT}},M_{\mathrm{FS}}}$. Hence, I chose to include Eq. (7.3) in the algorithm in Section 6.2.6, instead of assuming a linear dependency between σ_S and μ_S.

7.3. Determination of the Relation between Mean GFP Fluorescence and Frameshifting Efficiency

In this work, the frameshifting efficiency is calculated from the thermodynamic model $f_{\mathrm{model}}(\Delta G)$ (Eq. (3.3)). However, the mean GFP percent fluorescence values μ_S do not directly reflect the frameshifting efficiency (see Section 5.2). To get a relation between the two observables, I plot μ_S for several variants against the frameshifting efficiencies of the corresponding variants which were independently measured by Biswas et al. [17], Dulude et

al. [7], and Léger et al [68] (Fig. 11). The methods to get probability densities of μ_S for the sequences with available frameshifting efficiencies are explained in Section 6.2.5. Notably, three different values for the frameshifting efficiency of wild-type HIV were reported in the literature [7, 17, 68], resulting in three data points with $\mu_S \approx 2.04\,\%$. Since wild-type sequences were measured in many repeats, the data from Mikl et al. contains many percent GFP values and their standard deviation of μ_S is rather small [9].

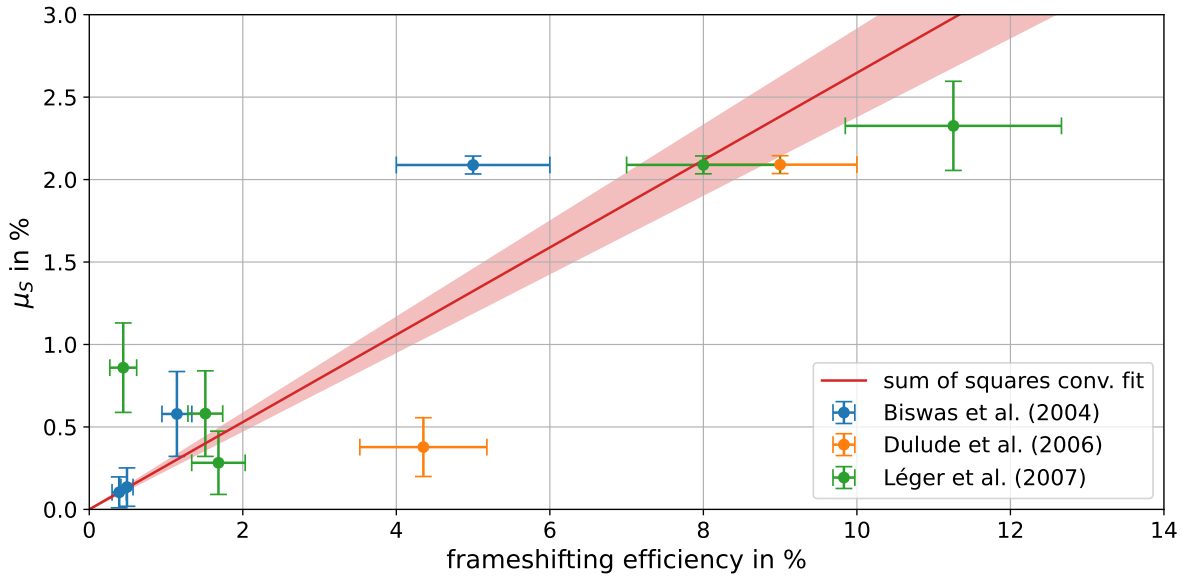

Figure 11: The mean value μ_S is plotted against the corresponding frameshifting efficiencies. For μ_S the data points represent the mean value and the error bars represent the standard deviations of the obtained probability densities. The error bars of the frameshifting efficiencies resemble the errors indicated in the respective paper. The papers are represented by different colors. The straight line of the sum of squares convergence fit of the data points is plotted in red. It is depicted plus minus its standard deviation.

The plot does not show a very clear relationship. As a first assumption, I consider a linear relationship as is displayed by the result of the sum of squares convergence fit in Fig. 11. The value resulting from the sum of squares convergence fit of the data points is $m_\mu \approx 0.225$. However, as explained in the methods, in order to make full use of the data (both wild-type and variant measurements), μ_S is varied in the following steps as a parameter, by including its linear relation to the frameshifting efficiency in the likelihood function:

$$\mu_S = m_\mu \cdot f_{\text{model}}(\Delta G)\,. \tag{7.4}$$

7.4. Determination of Free-Energy Differences with the Metropolis Algorithm

Using the algorithm from Section 6.2.6, I sample probability densities of ΔG for each sequence from Mikl et al. [9] that belongs to one of the the nine viruses mentioned in Section 2.5 and consists of either the wild type or variants of the slippery sequence, combined with wild-type upstream and downstream regions. In order to confirm that the free-energy differences converge to the same value, I sample each of the resulting 568 probability densities for ΔG within two independent calculations started from different sets of initial values. Thus, in total, I sample 1136 ΔG densities. Six probability densities for ΔG are plotted for three different sequences in Fig. 12 to exemplify the type of results that I obtained for the probability densities.

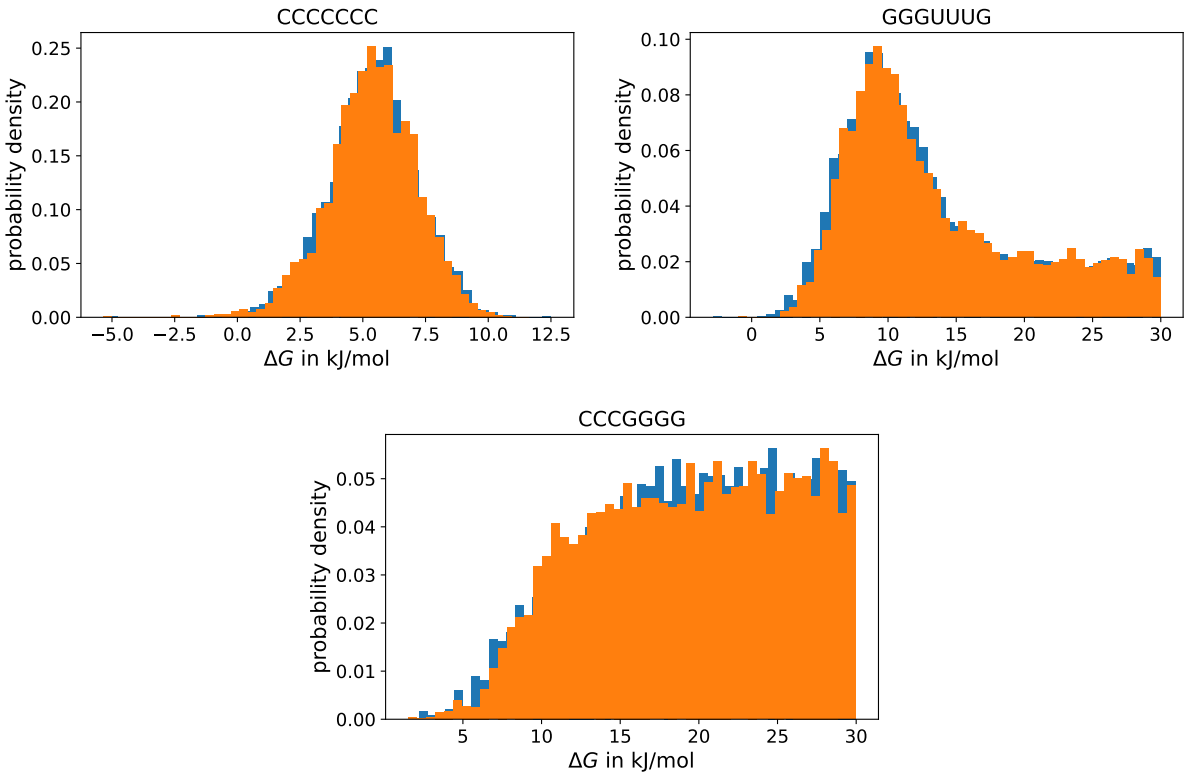

Figure 12: Examples of probability densities of ΔG for sequences with three different slippery sequences combined with the SARS wild-type upstream and downstream regions. Probability densities were obtained from two independent calculations (blue and orange).

For each ΔG I determine, I visually inspected the results from the two independent calculations for convergence. Fig. 12 displays also the different types of the obtained probability densities: in many cases I did not get a localized density (Fig. 12 upper left), but densities which correspond to an upper or lower boundary (Fig. 12 upper right and bottom). Therefore, many of the $\Delta\Delta G$ densities in Section 7.5 are not localized.

7.5. Determination of the Differences between Frameshifting Free-Energy Differences of Sequences with Different Downstream Secondary Structures

In Section 6.3 I explained my methods to determine probability densities for $\Delta\Delta G$. I compare all seven viruses with a pseudoknot mRNA secondary structure to HIV and SIV whose mRNA forms stem loops. The resulting $\Delta\Delta G$ probability densities are shown as violin plots in Figs. 13 to 19.

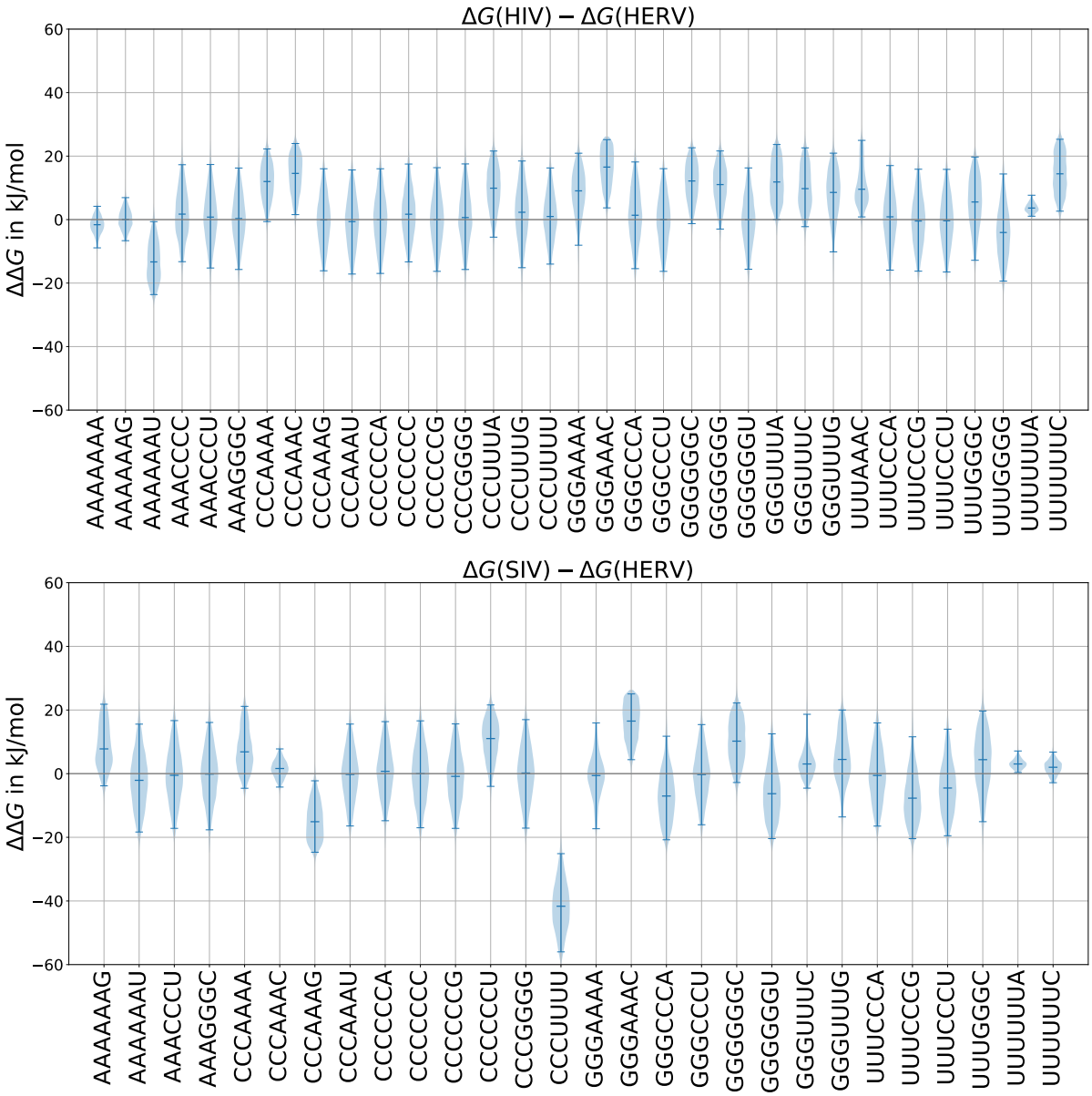

Figure 13: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and HERV and in the bottom plot of SIV and HERV.

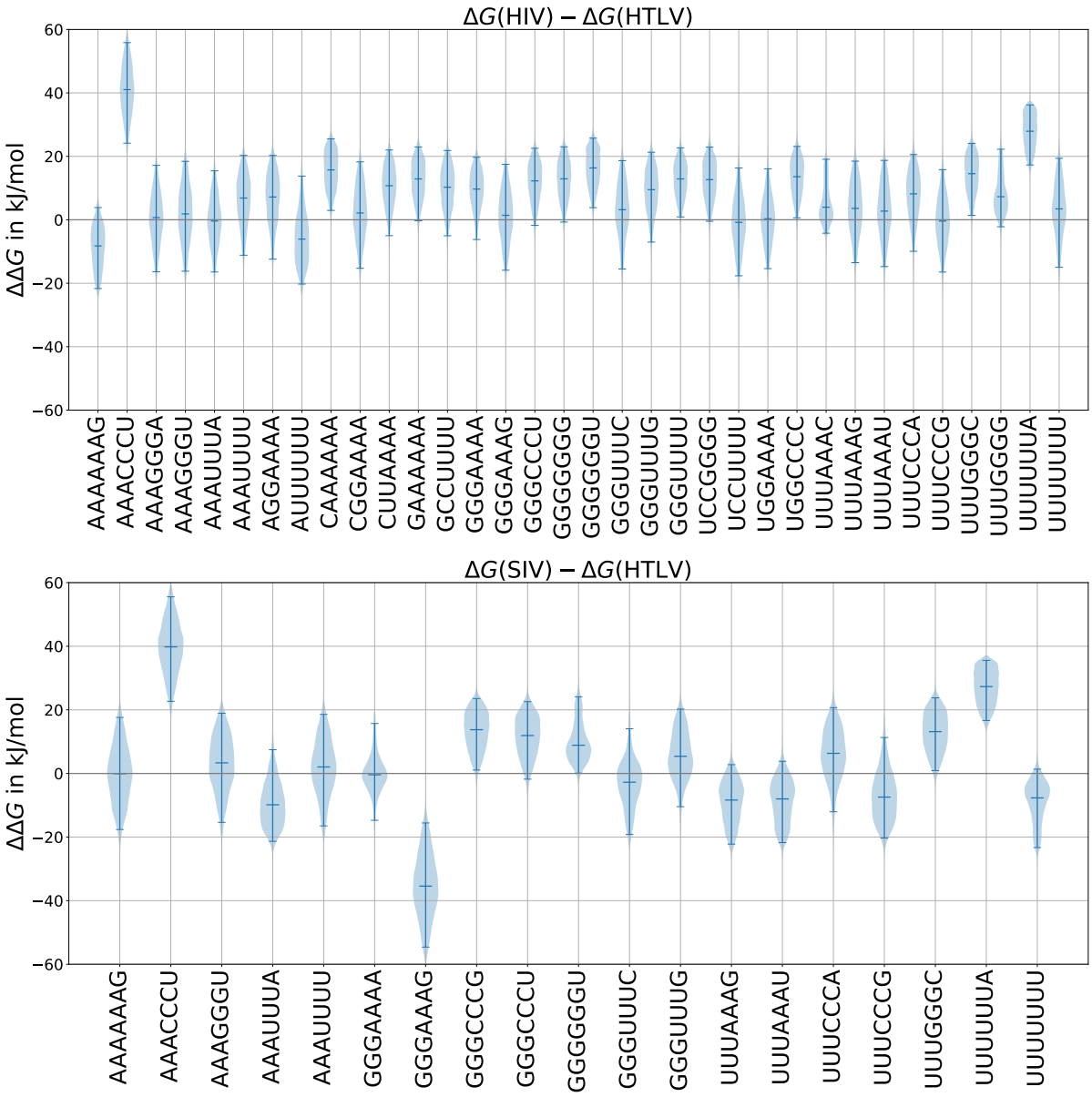

Figure 14: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and HTLV and in the bottom plot of SIV and HTLV.

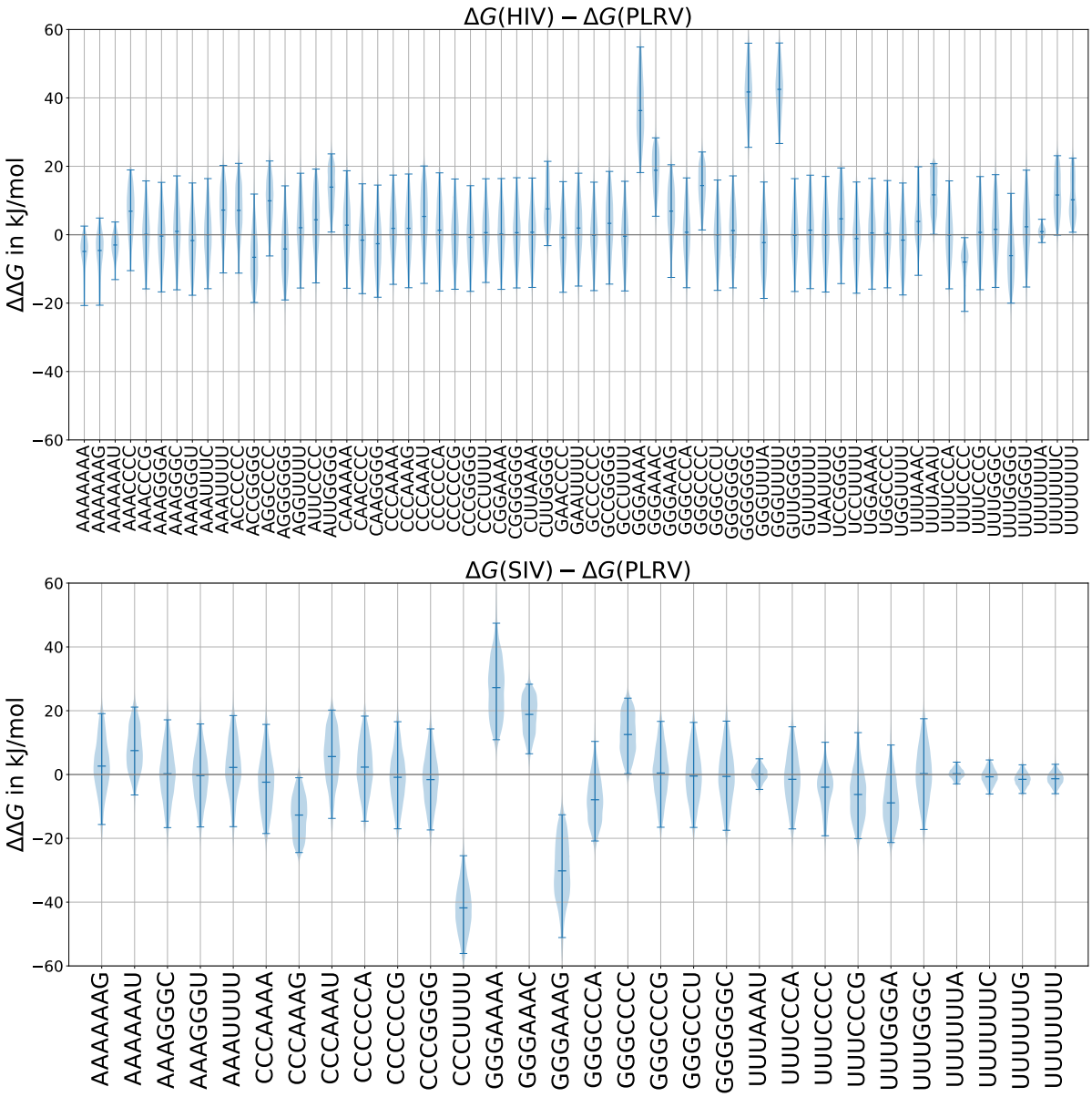

Figure 15: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and PLRV and in the bottom plot of SIV and PLRV.

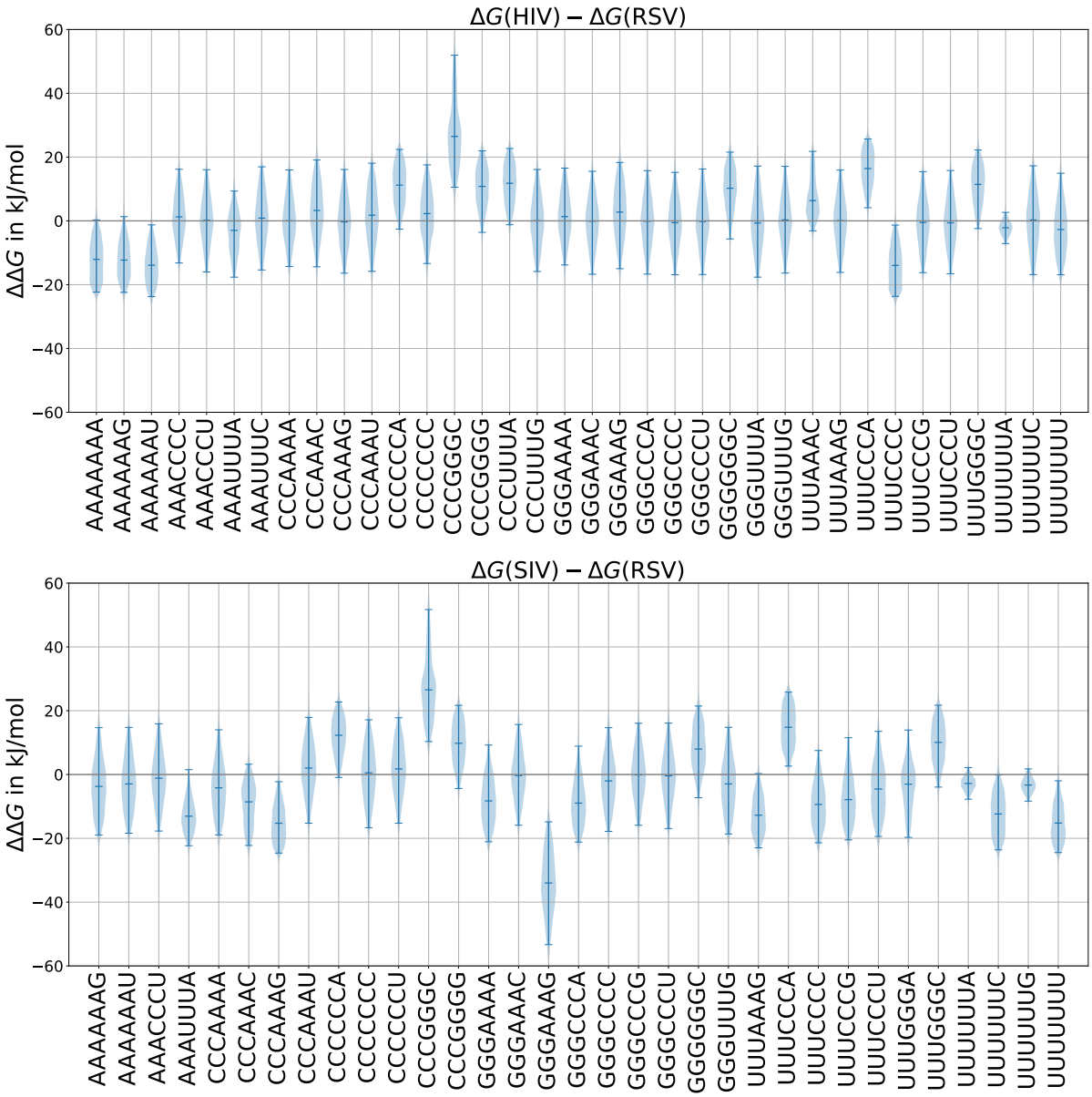

Figure 16: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and RSV and in the bottom plot of SIV and RSV.

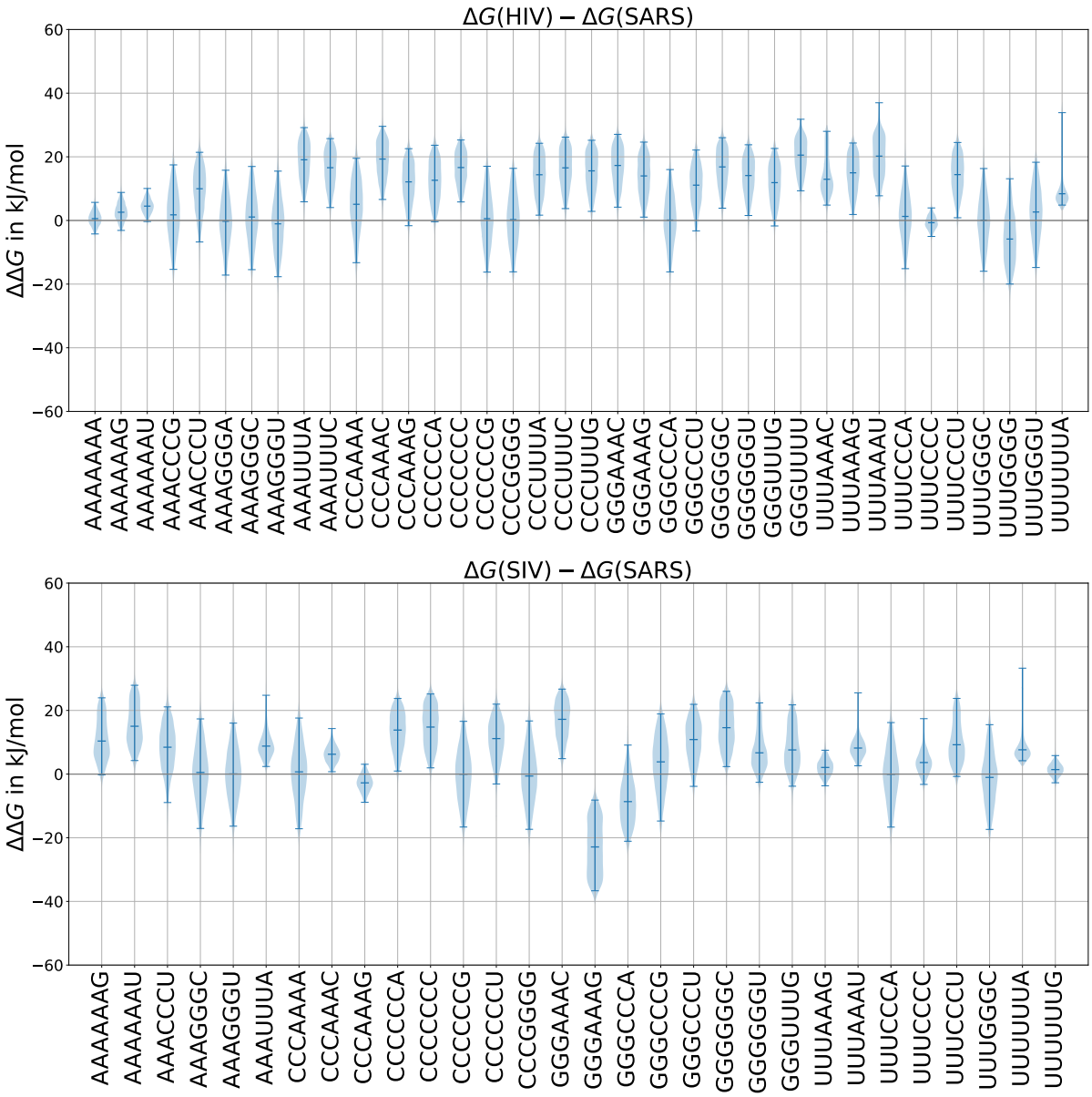

Figure 17: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and SARS and in the bottom plot of SIV and SARS.

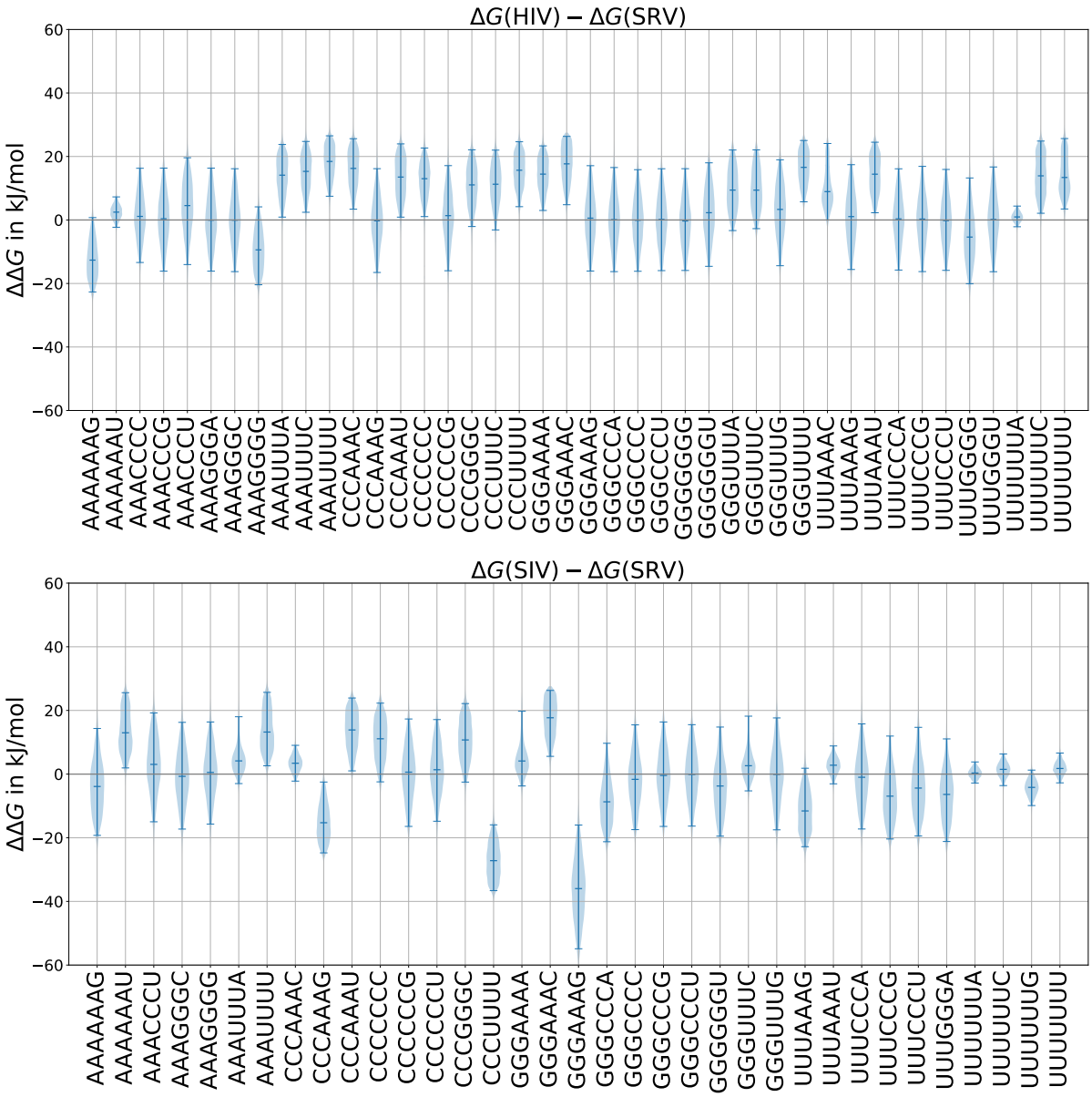

Figure 18: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and SRV and in the bottom plot of SIV and SRV.

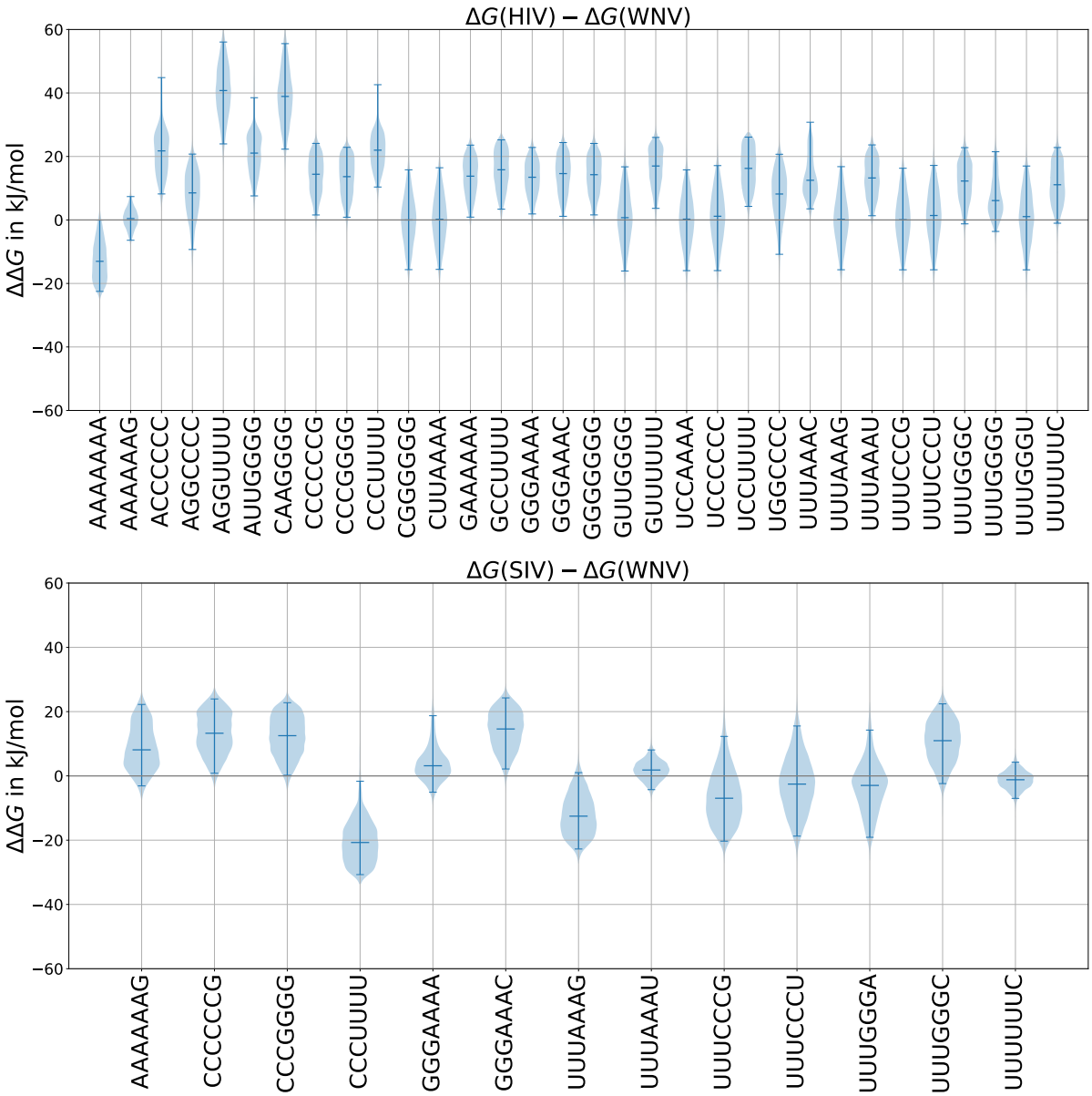

Figure 19: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. In the upper plot the slippery sequences are surrounded by the sequence context of HIV and WNV and in the bottom plot of SIV and WNV.

The probability densities in Figs. 13 to 19 have rather large 95 % confidence intervals and many are located around 0 kJ/mol. However, the viruses compared with HIV display more $\Delta\Delta G$ densities above 0 kJ/mol than with SIV. When compared with SIV, the densities appear to be balanced around 0 kJ/mol. Furthermore, there are plots that exhibit one or two very low $\Delta\Delta G$ probability densities with median below -20 kJ/mol. These densities correspond always either to the slippery sequence „GGGAAAG“ or „CCCUUUU“.

To show a more direct comparison between the effect of the secondary structure on different viruses, I combine all $\Delta\Delta G$ probability densities for each comparison between

two viruses from Figs. 13 to 19 by multiplying their probability densities as explained in Section 6.3. The resulting plot is depicted in Fig. 20.

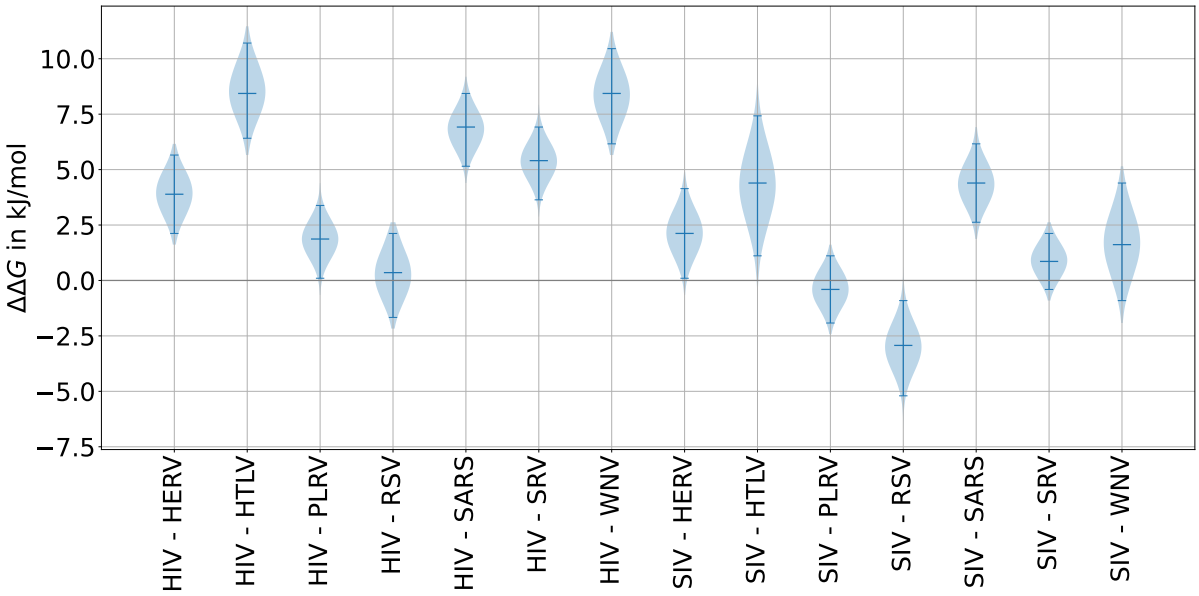

Figure 20: For each pair of viruses, all $\Delta\Delta G$ probability densities are combined to one probability density. The median of the $\Delta\Delta G$ density, as well as the 95 % confidence interval, is marked.

As in Figs. 13 to 19 also in Fig. 20, HIV displays more probability densities above 0 kJ/mol than SIV, whose densities are again more balanced around 0 kJ/mol.

To investigate this difference between HIV and SIV I show the $\Delta\Delta G$ probability densities for HIV and SIV in comparison in Fig. 21. The selected sequences are the sequences for which Mikl et al. [9] provided measurements and, additionally, that are considered in Figs. 13 to 19.

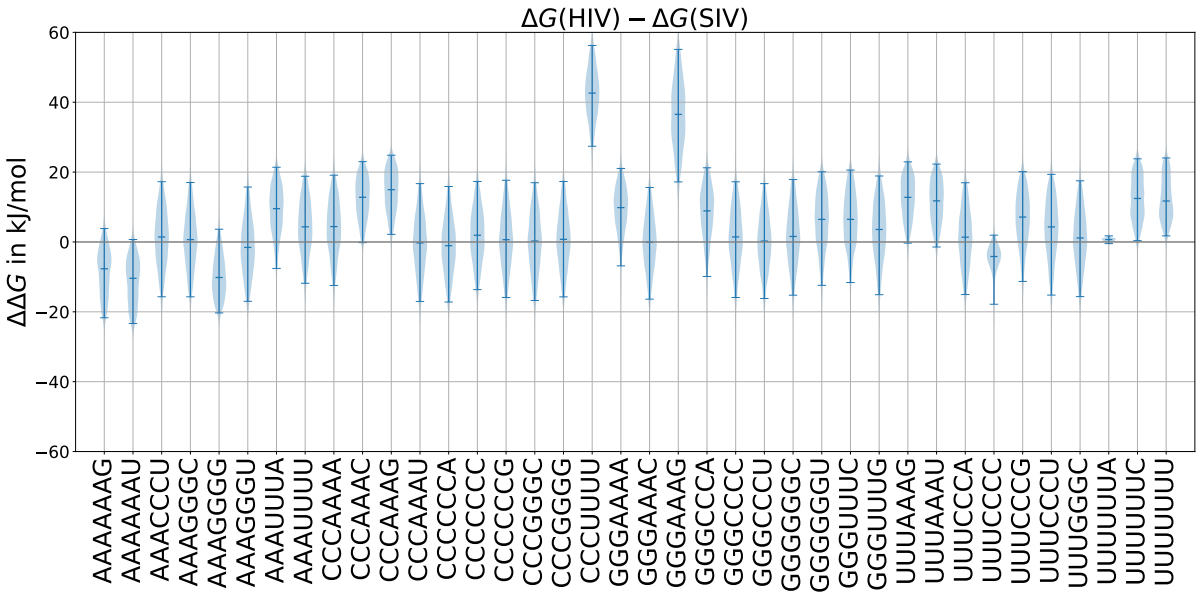

Figure 21: The probability densities of $\Delta\Delta G$ are shown for different slippery sequences (light blue area). The medians and the 95 % confidence intervals of the densities are indicated by horizontal lines. The slippery sequences are surrounded by the sequence context of HIV and SIV.

Fig. 21 depicts most $\Delta\Delta G$ probability densities around 0 kJ/mol. However, there are two slippery sequences that stand out for being located significantly higher: „GGGAAAG“ and „CCCUUUU“. Both were already mentioned above for showing probability densities that are significantly shifted to lower free-energy differences for each pair of virus that contains SIV. The probability density which corresponds to the slippery sequence „UUUUUUA“ shows a very low uncertainty, because it represents the wild-type sequence of HIV as well as of SIV and is therefore based on many measurements.

8. Discussion

Based on the results shown above, in particular the difference in frameshifting free-energy differences between sequences displaying two different downstream secondary structures, here I discuss the influence of the stem-loop and pseudoknot on the free energies associated to frameshifting efficiencies. According to my hypothesis, a pseudoknot might reduce the free-energy difference between the 0 and the -1 frame. Hence, the free-energy difference ΔG of a sequence with a stem loop should be larger than the free-energy difference ΔG of a sequence with a pseudoknot. Consequently, when considering the model in Eq. (3.5), a $\Delta\Delta G > 0$ would support my hypothesis. In my results from Section 7.5, the $\Delta\Delta G$ probability densities obtained by comparison of sequences that contain a pseudoknot with variants of HIV sequences are consistent with the hypothesis, since they are

mostly located above 0 kJ/mol. In particular, for six out of seven virus pairs, the 95 % confidence interval is above 0 kJ/mol. In contrast, for the comparison with variants of SIV sequences, only three virus pairs, shows a confidence interval that does not contain 0 kJ/mol. Therefore, the comparison with SIV sequences does not strongly support the hypothesis that a pseudoknot reduces the free-energy difference between the frames, but it also does not contradict the hypothesis. Further, regarding the probability densities obtained from SIV sequences in Figs. 13 to 19, it is worth noticing that mainly the results from two of the slippery sequences contribute to the fact that the densities in Fig. 20 are lower than the ones obtained when comparing with HIV sequences. Indeed, the densities associated with „GGGAAAG“ and „CCCUUUU“ display $\Delta\Delta G$ densities with a median lower than -20 kJ/mol. Taking a look in the provided data from Mikl et al. gives a reason for this occurrence. For each selected sequence with either „GGGAAAG“ or „CCCUUUU“ as a slippery sequence, only one percent GFP expression value is provided. In addition, this single value is significantly larger (85.39 % and 23.00 %) than most of the other GFP expression values measured for the other sequences [9]. A large percent GFP expression value results in a large frameshifting efficiency and therefore, in a very low ΔG, which is visible in Figs. 13 to 19 and also in Fig. 21. Since there is only one value available for both sequences, excluding outliers with the interquartile range as in Section 6.1 could not be applied. The measurement method for the percent GFP values from Mikl et al. allows a high through-put, but is in turn less accurate than other methods [9]. Thus, it is questionable why these percent GFP values are that large. Indeed, neither of the sequences corresponds to the wild type, where the slippery sequence is expected to be optimized to yield the highest frameshifting efficiency. Consequently, this may be a measurement error. For most of the tested sequences, that do not correspond to the wild type, there is only one percent GFP expression value measured by Mikl et al. [9]. Hence, for most sequences, I could not exclude outliers with the interquartile range as in Section 6.1 and, thus, it cannot be ruled out that there are more measurement errors as the ones I expect for the two sequences I considered above. Given sufficient statistics to obtain a model for this additional error, it could be included in the likelihood function and give a more robust estimation of the uncertainty of the $\Delta\Delta G$ values.

In general, the $\Delta\Delta G$ probability densities in Fig. 20 could be explained by relating them to the position and size of the pseudoknot in the sequence. RSV does not portray a trend of $\Delta\Delta G$ densities above 0 kJ/mol. An explanation could be the shortness of its spacer region (one nucleotide), which could lead the pseudoknot to have a position other than at the entrance of the mRNA entry channel when frameshifting takes place. Consequently, another mechanism than the pseudoknot resisting the unfolding and reducing the free-energy difference could occur. After RSV, the viruses HERV, PLRV, and SRV show $\Delta\Delta G$ densities with the lowest free-energy differences. This might be due to the size of their

pseudoknots: HERV, PLRV, and SRV exhibit a rather small pseudoknot containing 34, 26, and 37 nucleotides, respectively. For comparison, the pseudoknots of SARS, HTLV, and WNV are 68, 72, and 62 nucleotides long, respectively (compare Fig. 6). A small pseudoknot might generate a less strong back-pull compared to a large pseudoknot, due to the smaller number of base pairs that need to be dissolved upon unfolding. According to my hypothesis a less strong back-pull would result in a lower free-energy reduction, which would explain the lower $\Delta\Delta G$ densities of HERV, PLRV, and SRV. Additionally, this is in agreement with the obtained densities for HTLV, SARS, and WNV, which display the highest $\Delta\Delta G$ densities in Fig. 20 and have (excluding RSV) the largest pseudoknots (72, 68, and 62 nucleotides, respectively).

However, it is still unclear, why the viruses, when compared with HIV, show densities that are localized much higher than when compared with SIV. According to my hypothesis the stem loop should not affect the free-energy difference between the 0 and the -1 frame. However, an explanation could base on a kinetic effect. In particular, translocation could not be stalled sufficiently by the SIV stem loop (e.g. due to a potentially less stable SIV stem loop), such that there is less time for the ribosome to overcome the free-energy barrier between the frames. Consequently, the frameshifting efficiency and therefore the GFP expression would be decreased. A decreased measured GFP expression would result in $\Delta\Delta G$ probability densities that are shifted to lower free-energy differences and explain Fig. 20. Nonetheless, Fig. 21 shows no general trend in the positive direction. Most $\Delta\Delta G$ probability densities are balanced around 0 kJ/mol. Thus, I do not assume a kinetic effect and the fact of SIV $\Delta\Delta G$ densities being located lower in Fig. 20 might be only based on the few sequences that are located above 0 kJ/mol in Fig. 21.

There may be other elements contributing to the frameshifting efficiency. Until now, I did not consider the upstream sequences of the viruses. There is experimental evidence, that the slippery sequence and a downstream secondary structure affect the frameshifting efficiency the most. However, it cannot be ruled out, that the upstream sequence does not play a role in the frameshifting efficiency. SARS-CoV-1 for example, has an upstream secondary structure: the virus exhibits a so-called attenuator loop upstream of its slippery sequence. It is called attenuator loop, because it is proposed to play an attenuating role in frameshifting, meaning that it decreases the frameshifting efficiency [69].

Additionally, there is experimental evidence that a stop codon near the slippery sequence affects frameshifting. Bhatt et al. proposed that the presence of a stop codon near the frameshifting site of SARS-CoV-2 increases the frameshifting efficiency [22].

Thus, including also other contributing elements into my analysis might produce more accurate results. What also would enhance the accuracy of my results, is to modify the likelihood function from Section 6.2.6 to make use of all the information more efficiently. Instead of sampling a ΔG probability density for each sequence separately, the Metropolis

algorithm could sample the ΔG of all of the sequences at the same time. Therefore, it would be possible to compute the ΔG probability densities for all sequences after one single run. In this way, when sampling the parameters, I would have the information not only from one specific variant $\tilde{V}$, but also from the data sets of all other considered variants. Additionally, such a modification would allow to cross-validate the results by leaving out a subset of the data and predicting it from the remaining data set.

9. Conclusion

In this thesis I explored the effect of downstream secondary structures on the thermodynamics of frameshifting, in particular, I aimed at comparing the effect of stem loop and pseudoknot on the free energies. For this purpose I employed Bayesian statistics on the measurements from Mikl et al. [9] and used the Metropolis algorithm to estimate free-energy differences between the 0 and the -1 frame for many selected sequences. Afterwards, I determined probability densities for the differences between frameshifting free-energy differences of sequences with different secondary structures, in order to test my hypothesis of a pseudoknot reducing the free-energy difference by generating a back-pull towards the -1 frame. While conclusive results would surely require a more precise data set, more included parameters, as the upstream sequence or the position of a stop codon, and a further modification of the likelihood function, my results support the indication that pseudoknots reduce the free-energy difference in contrast to stem loops and, therefore, the pseudoknots enhance frameshifting efficiencies. This gives a first glimpse of the possibly important influence of the secondary structure on the thermodynamics of frameshifting, and, therefore, on the reproduction of viruses overall.

References

[1] David L Nelson and Michael M Cox. *Lehninger principles of biochemistry*. 7th ed. W. H. Freeman, Jan. 2017. ISBN: 978-1-4641-2611-6.

[2] William H Elliott and Daphne C Elliott. *Biochemistry and molecular biology*. 3rd ed. New York, NY: Oxford University Press, 2005. ISBN: 0-19-927199-2.

[3] Shannon Yan, Jin-Der Wen, Carlos Bustamante, and Ignacio Tinoco Jr. "Ribosome excursions during mRNA translocation mediate broad branching of frameshift pathways". In: *Cell* 160.5 (2015), pp. 870–881.

[4] Jonathan D Dinman. "Control of gene expression by translational recoding". In: *Advances in protein chemistry and structural biology* 86 (2012), pp. 129–149.

[5] CG Kurland. "Translational accuracy and the fitness of bacteria". In: *Annual review of genetics* 26.1 (1992), pp. 29–50.

[6] Neva Caliskan, Frank Peske, and Marina V Rodnina. "Changed in translation: mRNA recoding by- 1 programmed ribosomal frameshifting". In: *Trends in biochemical sciences* 40.5 (2015), pp. 265–274.

[7] Dominic Dulude, Yamina A Berchiche, Karine Gendron, Léa Brakier-Gingras, and Nikolaus Heveker. "Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1". In: *Virology* 345.1 (2006), pp. 127–136.

[8] Lars V Bock, Neva Caliskan, Natalia Korniy, Frank Peske, Marina V Rodnina, and Helmut Grubmüller. "Thermodynamic control of -1 programmed ribosomal frameshifting". In: *Nature communications* 10.1 (2019), pp. 1–11.

[9] Martin Mikl, Yitzhak Pilpel, and Eran Segal. "High-throughput interrogation of programmed ribosomal frameshifting in human cells". In: *Nature communications* 11.1 (2020), pp. 1–18.

[10] John W Baynes and Marek H Dominiczak. *Medical Biochemistry*. 5th ed. London, England: Elsevier Health Sciences, 2019. ISBN: 978-0-7020-7299-4.

[11] Daniel N Wilson and Jamie H Doudna Cate. "The structure and function of the eukaryotic ribosome". In: *Cold Spring Harbor perspectives in biology* 4.5 (2012).

[12] Song Cao and Shi-Jie Chen. "Predicting ribosomal frameshifting efficiency". In: *Physical biology* 5.1 (2008).

[13] Florian Klepper, Eva-Maria Jahn, Volker Hickmann, and Thomas Carell. "Synthesis of the transfer-RNA nucleoside queuosine by using a chiral allyl azide intermediate". In: *Angewandte Chemie International Edition* 46.13 (2007), pp. 2325–2327.

[14] Jack Parker. “Errors and alternatives in reading the universal genetic code”. In: *Microbiological reviews* 53.3 (1989), pp. 273–298.

[15] Ewan P Plant, Rasa Rakauskaite, Deborah R Taylor, and Jonathan D Dinman. “Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins”. In: *Journal of virology* 84.9 (2010), pp. 4330–4340.

[16] Zenta Tsuchihashi. “Translational frame shifting in the Escherichia coli dnaX gene in vitro”. In: *Nucleic acids research* 19.9 (1991), pp. 2457–2462.

[17] Preetha Biswas, Xi Jiang, Annmarie L Pacchia, Joseph P Dougherty, and Stuart W Peltz. “The human immunodeficiency virus type 1 ribosomal frameshifting site is an invariant sequence determinant and an important target for antiviral therapy”. In: *Journal of virology* 78.4 (2004), pp. 2082–2087.

[18] Jason W Harger, Arturas Meskauskas, and Jonathan D Dinman. “An ’integrated model’ of programmed ribosomal frameshifting”. In: *Trends in biochemical sciences* 27.9 (2002), pp. 448–454.

[19] Ian Brierley and Francisco J Dos Ramos. “Programmed ribosomal frameshifting in HIV-1 and the SARS–CoV”. In: *Virus research* 119.1 (2006), pp. 29–42.

[20] Ewan P Plant, Gabriela C Pérez-Alvarado, Jonathan L Jacobs, Bani Mukhopadhyay, Mirko Hennig, and Jonathan D Dinman. “A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal”. In: *PLoS biology* 3.6 (2005).

[21] Chen Bao, Sarah Loerch, Clarence Ling, Andrei A Korostelev, Nikolaus Grigorieff, and Dmitri N Ermolenko. “mRNA stem-loops can pause the ribosome by hindering A-site tRNA binding”. In: *Elife* 9 (2020).

[22] Pramod R Bhatt, Alain Scaiola, Gary Loughran, Marc Leibundgut, Annika Kratzel, Romane Meurs, René Dreos, Kate M O’Connor, Angus McMillan, Jeffrey W Bode, et al. “Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome”. In: *Science* 372.6548 (2021), pp. 1306–1313.

[23] Seyedtaghi Takyar, Robyn P Hickerson, and Harry F Noller. “mRNA helicase activity of the ribosome”. In: *Cell* 120.1 (2005), pp. 49–58.

[24] Kenneth Murphy and Casey Weaver. *Janeway’s immunobiology*. 9th ed. New York, NY: Garland science, 2016. ISBN: 978-0-8153-4505-3.

[25] Viktoriya S Anokhina and Benjamin L Miller. “Targeting ribosomal frameshifting as an antiviral strategy: from HIV-1 to SARS-CoV-2”. In: *Accounts of Chemical Research* 54.17 (2021), pp. 3349–3361.

[26] Stuart LeGrice and Matthias Gotte, eds. *Human Immunodeficiency Virus Reverse Transcriptase*. New York, NY: Springer, 2013. ISBN: 978-1-4614-7290-2.

[27] World Health Organization. *HIV/AIDS. Key facts*. accessed 2022-08-18. URL: `https://www.who.int/news-room/fact-sheets/detail/hiv-aids`.

[28] Martine Peeters and Valerie Courgnaud. "Overview of primate lentiviruses and their evolution in non-human primates in Africa". In: *HIV sequence compendium* 2 (2002).

[29] Patrik Medstrand and Dixie L Mager. "Human-specific integrations of the HERV-K endogenous retrovirus family". In: *Journal of virology* 72.12 (1998), pp. 9782–9787.

[30] Jan Paces, Adam Pavlícek, and Václav Paces. "HERVd: database of human endogenous retroviruses". In: *Nucleic acids research* 30.1 (2002), pp. 205–206.

[31] Eiji Matsuura, Yoshihisa Yamano, and Steven Jacobson. "Neuroimmunity of HTLV-I infection". In: *Journal of Neuroimmune Pharmacology* 5.3 (2010), pp. 310–325.

[32] Fernando A Proietti, Anna Bárbara F Carneiro-Proietti, Bernadette C Catalan-Soares, and Edward L Murphy. "Global epidemiology of HTLV-I infection and associated diseases". In: *Oncogene* 24.39 (2005), pp. 6058–6068.

[33] Michael Taliansky, Mike A Mayo, and Hugh Barker. "Potato leafroll virus: a classic pathogen shows some new tricks". In: *Molecular plant pathology* 4.2 (2003), pp. 81–89.

[34] Robin A Weiss and Peter K Vogt. "100 years of Rous sarcoma virus". In: *Journal of Experimental Medicine* 208.12 (2011), pp. 2351–2355.

[35] Paul A Rota, M Steven Oberste, Stephan S Monroe, W Allan Nix, Ray Campagnoli, Joseph P Icenogle, Silvia Penaranda, Bettina Bankamp, Kaija Maher, Min-hsin Chen, et al. "Characterization of a novel coronavirus associated with severe acute respiratory syndrome". In: *science* 300.5624 (2003), pp. 1394–1399.

[36] World Health Organization. *Severe Acute Respiratory Syndrome (SARS)*. accessed 2022-08-19. URL: `https://www.who.int/health-topics/severe-acute-respiratory-syndrome#tab=tab_1`.

[37] Sara Ibrahim Omar, Meng Zhao, Rohith Vedhthaanth Sekar, Sahar Arbabi Moghadam, Jack A Tuszynski, and Michael T Woodside. "Modeling the structure of the frameshift-stimulatory pseudoknot in SARS-CoV-2 reveals multiple possible conformers". In: *PLoS computational biology* 17.1 (2021).

[38] Michael D Power, Preston A Marx, Martin L Bryant, Murray B Gardner, Philip J Barr, and Paul A Luciw. "Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome retrovirus". In: *Science* 231.4745 (1986), pp. 1567–1572.

[39] Edwin ten Dam, Ian Brierley, Stephen Inglis, and Cornelis Pleij. "Identification and analysis of the pseudoknot-containing gag-pro ribosomal frameshift signal of simian retrovirus-1". In: *Nucleic acids research* 22.12 (1994), pp. 2304–2310.

[40] Grant L Campbell, Anthony A Marfin, Robert S Lanciotti, and Duane J Gubler. "West nile virus". In: *The Lancet infectious diseases* 2.9 (2002), pp. 519–529.

[41] Xuanmao Jiao, Omar Nawab, Tejal Patel, Andrew V Kossenkov, Niels Halama, Dirk Jaeger, and Richard G Pestell. "Recent Advances Targeting CCR5 for Cancer and Its Role in Immuno-Oncology". In: *Cancer research* 79.19 (2019), pp. 4801–4807.

[42] Ashton Trey Belew, Arturas Meskauskas, Sharmishtha Musalgaonkar, Vivek M Advani, Sergey O Sulima, Wojciech K Kasprzak, Bruce A Shapiro, and Jonathan D Dinman. "Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway". In: *Nature* 512.7514 (2014), pp. 265–269.

[43] Yousuf A Khan, Gary Loughran, Anna-Lena Steckelberg, Katherine Brown, Stephen J Kiniry, Hazel Stewart, Pavel V Baranov, Jeffrey S Kieft, Andrew E Firth, and John F Atkins. "Evaluating ribosomal frameshifting in CCR5 mRNA decoding". In: *Nature* 604.7906 (2022).

[44] Taylor A Evans and Jennifer Ann Erwin. "Retroelement-derived RNA and its role in the brain". In: *Seminars in Cell & Developmental Biology*. Vol. 114. Elsevier. 2021, pp. 68–80.

[45] Tyler Jacks, Michael D Power, Frank R Masiarz, Paul A Luciw, Philip J Barr, and Harold E Varmus. "Characterization of ribosomal frameshifting in HIV-1 gag-pol expression". In: *Nature* 331.6153 (1988), pp. 280–283.

[46] Edwin B Ten Dam, Cornelius WA Pleij, and Leendert Bosch. "RNA pseudoknots: translational frameshifting and readthrough on viral RNAs". In: *Virus genes* 4.2 (1990), pp. 121–136.

[47] Zhihua Du, Jason A Holland, Mark R Hansen, David P Giedroc, and David W Hoffman. "Base-pairings within the RNA pseudoknot associated with the simian retrovirus-1 gag-pro frameshift site". In: *Journal of molecular biology* 270.3 (1997), pp. 464–470.

[48] Yang-Gyun Kim, Stefan Maas, Stephanie C Wang, and Alexander Rich. "Mutational study reveals that tertiary interactions are conserved in ribosomal frameshifting pseudoknots of two luteoviruses". In: *RNA* 6.8 (2000), pp. 1157–1165.

[49] Alicja B Kujawa, Gabriele Drugeon, Danuta Hulanicka, and Anne-Lise Haenni. "Structural requirements for efficient translational frameshifting in the synthesis of the putative viral RNA-dependent RNA polymerase of potato leafroll virus". In: *Nucleic Acids Research* 21.9 (1993), pp. 2165–2171.

[50] Pradeep S Pallan, William S Marshall, Joel Harp, Frederic C Jewett, Zdzislaw Wawrzak, Bernard A Brown, Alexander Rich, and Martin Egli. "Crystal structure of a luteoviral RNA pseudoknot and model for a minimal ribosomal frameshifting motif". In: *Biochemistry* 44.34 (2005), pp. 11315–11322.

[51] Eliza Thulson, Erik W Hartwick, Andrew Cooper-Sansone, Marcus AC Williams, Mary E Soliman, Leila K Robinson, Jeffrey S Kieft, and Kathryn D Mouzakis. "An RNA pseudoknot stimulates HTLV-1 pro-pol programmed- 1 ribosomal frameshifting". In: *Rna* 26.4 (2020), pp. 512–528.

[52] Ezequiel Balmori Melian, Edward Hinzman, Tomoko Nagasaki, Andrew E Firth, Norma M Wills, Amanda S Nouwens, Bradley J Blitvich, Jason Leung, Anneke Funk, John F Atkins, et al. "NS1' of flaviviruses in the Japanese encephalitis virus serogroup is a product of ribosomal frameshifting and plays a role in viral neuroinvasiveness". In: *Journal of virology* 84.3 (2010), pp. 1641–1647.

[53] Bo Wu, Haibo Zhang, Ruirui Sun, Sijia Peng, Barry S Cooperman, Yale E Goldman, and Chunlai Chen. "Translocation kinetics and structural dynamics of ribosomes are modulated by the conformational plasticity of downstream pseudoknots". In: *Nucleic acids research* 46.18 (2018), pp. 9736–9748.

[54] Emily IC Nikolić, Louise M King, Marijana Vidakovic, Nerea Irigoyen, and Ian Brierley. "Modulation of ribosomal frameshifting frequency and its effect on the replication of Rous sarcoma virus". In: *Journal of virology* 86.21 (2012), pp. 11581–11594.

[55] Ernő Keszei. *Chemical Thermodynamics: An Introduction*. Berlin, Heidelberg: Springer, 2012. ISBN: 978-3-642-19863-2.

[56] Mary Kathryn Cowles. *Applied Bayesian statistics*. Springer texts in statistics. New York, NY: Springer, Jan. 2013. ISBN: 978-1-4614-5695-7.

[57] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, et al. "Bayesian statistics and modelling". In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–26.

[58] William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. 3rd ed. New Jersey: John Wiley & Sons, 2017. ISBN: 978-1-118-09315-8.

[59] Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of monte carlo methods*. New Jersey: John Wiley & Sons, 2011.

[60] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[61] Roger Y Tsien. "The green fluorescent protein". In: *Annual review of biochemistry* 67.1 (1998), pp. 509–544.

[62] Julien Picot, Coralie L Guerin, Caroline Le Van Kim, and Chantal M Boulanger. "Flow cytometry: retrospective, fundamentals and recent instrumentation". In: *Cytotechnology* 64.2 (2012), pp. 109–130.

[63] Monica Monici. "Cell and tissue autofluorescence research and diagnostic applications". In: *Biotechnology annual review* 11 (2005), pp. 227–256.

[64] J P Verma. "Application of factor analysis: To study the factor structure among variables". In: *Data Analysis in Management with SPSS Software*. India: Springer India, 2013, pp. 359–387.

[65] Patrick Billingsley. *Probability and Measure*. 3rd ed. Wiley Series in Probability & Mathematical Statistics. Nashville, TN: John Wiley & Sons, May 1995. ISBN: 0-471-00710-2.

[66] Markus Haase. *Functional analysis*. Graduate studies in mathematics. Providence, RI: American Mathematical Society, Sept. 2014. ISBN: 978-0-8218-9171-1.

[67] Ryuichi Ono, Shin Kobayashi, Hirotaka Wagatsuma, Kohzo Aisaka, Takashi Kohda, Tomoko Kaneko-Ishino, and Fumitoshi Ishino. "A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21". In: *Genomics* 73.2 (2001), pp. 232–237.

[68] Mélissa Léger, Dominic Dulude, Sergey V Steinberg, and Léa Brakier-Gingras. "The three transfer RNAs occupying the A, P and E sites on the ribosome are involved in viral programmed-1 ribosomal frameshift". In: *Nucleic acids research* 35.16 (2007), pp. 5581–5592.

[69] Mei-Chi Su, Chung-Te Chang, Chiu-Hui Chu, Ching-Hsiu Tsai, and Kung-Yao Chang. "An atypical RNA pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of SARS coronavirus". In: *Nucleic acids research* 33.13 (2005), pp. 4265–4275.

A. Calculation of the Convolution

In this section, my aim is to calculate the convolution

$$f_M = f_B * f_S \tag{A.1}$$

with the convolution theorem [66]

$$\mathcal{F}\{f_B * f_S\}(z) = \mathcal{F}\{f_B\}(k) \cdot \mathcal{F}\{f_S\}(k)\,, \tag{A.2}$$

where $\mathcal{F}$ denotes a Fourier transform. A Fourier transform is defined as follows [65]:

$$\mathcal{F}\{f\}(k) = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{-ikx}\mathrm{d}x\,. \tag{A.3}$$

Hence, the Fourier transform of a Gaussian f like f_B or f_S is

$$\mathcal{F}\{f\}(k) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2} - ikx}\mathrm{d}x\,. \tag{A.4}$$

By rewriting the exponent

$$-\frac{(x-\mu)^2}{2\sigma^2} - ikx = -\frac{(x-\xi)^2}{2\sigma^2} - i\mu k - \frac{\sigma^2 k^2}{2} \tag{A.5}$$

with $\xi = \mu - i\sigma^2 k$ I get

$$\mathcal{F}\{f\}(k) = \mathrm{e}^{-i\mu k - \frac{\sigma^2 k^2}{2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-\xi)^2}{2\sigma^2}}\mathrm{d}x \tag{A.6}$$

$$= \mathrm{e}^{-i\mu k - \frac{\sigma^2 k^2}{2}}\,. \tag{A.7}$$

The second equal sign is valid since the PDF following the $\cdot$ is normalized to one by definition. Multiplying the Fourier transforms point-wise yields

$$\mathcal{F}\{f_M\}(k) = \mathcal{F}\{f_B\}(k) \cdot \mathcal{F}\{f_S\}(k) = \mathrm{e}^{-i(\mu_B+\mu_S)k - \frac{(\sigma_B^2+\sigma_S^2)k^2}{2}}\,. \tag{A.8}$$

Applying the inverse Fourier transform

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{F}\{f\}(k)\mathrm{e}^{ikx}\mathrm{d}k \tag{A.9}$$

gives the required PDF

$$f_M(m) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_B^2 + \sigma_S^2}} \mathrm{e}^{-\frac{(m-\mu_B-\mu_S)^2}{2(\sigma_B^2+\sigma_S^2)}} \,. \tag{A.10}$$