universität wien

MASTERARBEIT / MASTER'S THESIS

„MS2 virus and the RNA/protein complementarity hypothesis“

verfasst von / submitted by

Florian Pötsch BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Univ.-Prof. Dr. Bojan Zagrovic BA

Acknowledgements

I would first like to thank my thesis advisor Univ.-Prof. Dr. Bojan Zagrovic, BA of the MFPL at the University of Vienna. The door to Prof. Zagrovic's office was always open whenever I ran into trouble or had a question about my research or writing. He consistently allowed me to explore different angles of my work, yet steered me in the right direction whenever I found myself drowned in a maze of conflicting possibilities.

I would also like to thank my colleges who taught me many valuable techniques which I could implement in my research project. Their passionate participation and input shed light on many variable ways I could otherwise not have thought of.

Finally, I must express my very profound gratitude to my parents and to my girlfriend for providing me with their unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Despite their importance, our understanding of RNA-protein interactions remains incomplete. This, in particular, concerns the interactions involving protein-coding mRNA sequences. In this regard, recent work has demonstrated a complementary relationship between nucleobase-density profiles of mRNA coding sequences and nucleobase-affinity profiles of their cognate protein sequences. This has been taken as a suggestion that mRNAs and their cognate proteins may directly interact in a co-aligned, complementary fashion, especially if unstructured. Here, we explore the RNA/protein complementarity hypothesis and study its limits of validity in a concrete biological system, the Enterobacteria phage MS2. Namely, the MS2 coat protein is known to bind in multiple locations to its own genomic RNA, providing a potential link with the complementarity hypothesis. First, we asked whether it is possible to detect interactions between the MS2 coat protein and its genome from sequence information only, following the methodical framework of the complementarity hypothesis. Second, we analyzed apparent periodicities in the interaction patterns between the MS2 RNA and coat protein via Fourier transform. Using the known nucleobase/amino-acid affinity scales, we were indeed able to identify 10 out of 13 possible detectable binding locations between the MS2 RNA and coat protein as reported experimentally. The complementarity hypothesis thus appears to provide a potentially promising approach for investigating and predicting specific RNA/protein interactions. However, as the relationship between individual RNA nucleobase profiles and protein nucleobase-affinity profiles remains unclear, further studies are needed in order to use it to design a robust, generally applicable tool for the analysis of RNA-protein interactions. Finally, our analysis did not detect any strong periodicities in the interaction patterns between the MS2 RNA and coat protein going beyond randomized controls.

Contents

List of Figures

Chapter 1

Introduction

The RNA world hypothesis is a widely accepted model aimed at explaining the early evolution of life [2]. According to the hypothesis, biology based on RNA molecules evolved first, while proteins and DNA entered at a later stage. A key challenge concerning the transition from the RNA world to modern biology concerns the evolution of translation i.e. establishment of a universal link between the RNA templates and the proteins they code for [3]. As one possibility, it has been proposed that the RNA templates themselves were originally able to function as scaffolds i.e. direct templates that were used for the synthesis and even folding of proteins [3]. Be it synthesis or folding of proteins by template RNAs, it is essential for such scenarios that some direct interaction between RNAs and proteins would take place. Importantly, the specificity in such interactions could have been in part defined by direct interaction preferences between RNA nucleotides and amino acids. Specifically, the stereochemical hypothesis of the origin of the genetic code suggests that the code evolved on the basis of direct binding preferences between codons and their cognate amino acids [4, 5]. While direct evidence in support for the stereochemical hypothesis has been rather scare, it is important to emphasize that the hypothesis has traditionally been examined almost exclusively in the context of individual codons and individual amino acids. However, any direct interactions between the two would likely be significantly potentiated in a polymeric context i.e. in the interactions between the complete mRNAs and their cognate proteins. The recently proposed cognate mRNA/protein complementarity hypothesis [6] is an attempt to explore this possibility and probe its potential consequences for the biology of today. It postulates that cognate mRNA and protein sequences may under some circumstances, be mutually physico-chemically complementary to each other and bind in a co-aligned fashion. Moreover, the generalized version of the complementarity hypothesis suggest that such interactions might be observed even beyond the cognate context i.e. between RNAs and proteins not related

by coding [7]. In particular, such interactions are expected to be relevant especially in the context of structural disorder i.e. in cases where either of the two partners, or both, are structurally disordered or unstructured. The established methods and computational programs for the prediction of RNA interactions (e.g. RCK, RNA-context, DeepBind, DisoRDPbind) [8, 9, 10, 11], rely largely on machine learning, or neural networks techniques. However, many of these methods suffer from limitation of training data, due in turn to the limited availability of relevant experimental data, or simply do not directly treat the physicochemical nature of the biopolymers involved. Although, there exist methods which use physicochemical logic in their approach i.e. catRAPID [12], one might argue that even there the underlying, generalizable principles may be difficult to discern. The complementarity hypothesis, on the other hand, relies purely on a clearly defined set of physicochemical principles and derives its predictions from the intrinsic affinities between nucleobases and amino acids for each other, which therefore may be able to tell us more about how, and why a given RNA and protein bind. This becomes especially pertinent in the case of interaction between unstructured elements. The lack of structure, in general, could also likely to be relevant when it comes to the question of how RNA-protein interactions evolved in primordial systems in which the unstructured elements where likely to be more abundant as compared to highly folded proteins and RNA [13]. Historically, the intrinsic binding affinities between nucleobases and amino acids have been studied in the context of a few specifically selected bases or amino acids [14, 15, 16], while only recently more comprehensive efforts have been undertaken. Specifically the comprehensive nucleobase/amino acid affinity scales that were used in the present study and covering all 4 standard RNA bases and all 20 standard amino acids were derived by using a distance dependent contact potential formalism [6] as applied to 300 high-resolution structures of different proteins-RNA complexes. The nucleobase/amino acid preferences were calculated for each individual RNA base, as well as, purines and pyrimidines using the following formalism: $\epsilon^{ij} = -ln\frac{N^{ij}_{obs}}{N^{ij}_{obs}} = -ln\frac{N^{ij}_{obs}}{X_i X_j N^{TOT}_{obs}}$ where N^{ij}_{obs} is the number of observed contacts between an amino-acid side chain of type i and a nucleobase of type j in experimental structures, and N^{ij}_{exp} is the expected number of such contacts. The latter is calculated as the product of molar fractions of amino acid i and base j among all observed contacts (X_i and X_j, respectively) and the total number of all observed contacts N^{TOT}_{obs} [17]. By using these computational derived nucleobases-amino acid affinities scales, as well as experimental scales, it has been shown that the nucleobase content of codons is directly related to the affinities of their cognate amino acid for precisely those nucleobases [7]. This has suggested

that the key feature of ancient translation may indeed have been a direct interaction between codons and amino acids they code for [4, 18, 19, 20]. Further analysis in this direction revealed that in some cases (e.g. high ADE content) it is actually the anti-codons that could preferentially interact with their cognate amino acids [7]. By extending the application of the above affinities to longer biopolymers such as mRNAs and their cognate proteins, it has been shown that, in humans, for example, the PUR density profiles of mRNAs match quantitatively the GUA-affinity profiles of cognate proteins with the median Pearson correlation coefficient of R=-0.80 [17, 21]. As these results were obtained primarily for primary sequence profiles, it is expected that any putative binding would occur in the context of dynamic, liquid-like, and multivalent complexes, such as in the case of IDPs or the unfolded states of folded proteins [22]. Finally, the complementarity hypothesis can be generalized to the level of non-cognate RNA and protein pairs: simply put, the hypothesis predicts favorable interactions between all RNAs and proteins with matching nucleobase-density and nucleobase-affinity profiles, regardless of whether they are related by a coding relationship or not.

Arguably the simplest biological system in which mRNAs reside in close proximity of their cognate proteins are positive sense RNA viruses. Simply put, the genome of such viruses directly encodes capsid proteins, which in turn encapsulate the genome. Importantly, capsid formation in many cases involves direct interactions between the viral RNA and capsid proteins. This provides an excellent test case for studying the generalized complementarity hypothesis in a biologically relevant context. Here, we have focused on the positive-sense RNA virus Enterobacteria phage MS2, one of the most widely studied viruses ever. MS2 has a small genome size of 3569 nucleotides that encode only 4 proteins (maturation protein, coat protein, lysis protein and replicase). Recent work has identified a 19nt stemloop structure in the genome of MS2, termed "packing signal", which directly binds to the coat protein dimer, and is responsible for initiating capsid formation. Moreover, it has been shown that there exist at least 14 different stemloop structures throughout the MS2 genome that are able and do indeed bind to the coat protein dimer [23]. Importantly, these binding events may not only be due to the secondary structure of the stemloops in question, but may also be related to the inherent features we would like to study. Here, we explore the possibility that a part of the specificity in RNA-protein interactions in the context of MS2 capsid formation may be related to the generalized complementarity hypothesis i.e. the possibility that the binding specificity may in part be detected at

the level of nucleobase density profiles of the MS2 genome at different locations and the nucleobase-affinity profiles of the MS2 coat proteins. In support of this possibility, we show that indeed in the regions where the coat protein is known to directly interact with the MS2 RNA, one also detects strongly complementary profiles of RNA nucleobase density and the coat protein nucleobase affinity profiles. Finally, given that the coat protein indeed binds multiple times to the genome, we use the formalism of Fourier transforms in order to explore the question of whether the signatures of complementary binding may be distributed periodically throughout the genome. Our analysis, however, suggests that the strength of any observed periodicities does not exceed that of randomized controls.

Chapter 2

Enterobacterio phage MS2

2.1 Background

The genome of the bacteriophage MS2 was the first genome to ever be sequenced completely [24] and is one of the smallest genomes known. It consists of a single stranded RNA [24] and encodes just four proteins: the maturation protein (A-protein), the lysis protein, the coat protein, and the replicase protein [25]. Enterobacteriophage MS2 belongs to a family of closely related bacterial viruses that includes bacteriophage f2, bacteriophage Q, R17, and GA [26].

2.2 Genome

2.2.1 Composition

The MS2 genome (NC_001417.2) consists of 3569 nucleotides, with a GC-content of 52%. The 5' end starts with a 130 nucleotide long leaderless non-coding RNA stretch followed by the coding sequences of the four proteins and a 171 nucleotide long non-coding RNA at the 3' end. The order of the CDSs of the four proteins is as follows: maturation protein (CDS from 130 to 1311), coat protein (1335 to 1727), lysis protein (1678 to 1905, overlapping with both the coat protein and the replicase), and the viral replicase beta subunit (1761 to 3398) Fig.2.1. The CDS of the maturation protein is the only exception with regards to the start codon, as it starts with a *gtg* and not the typical *atg*. The lysis CDS is another exception, as its CDS is *frameshifted +1* with regards to the other three proteins.

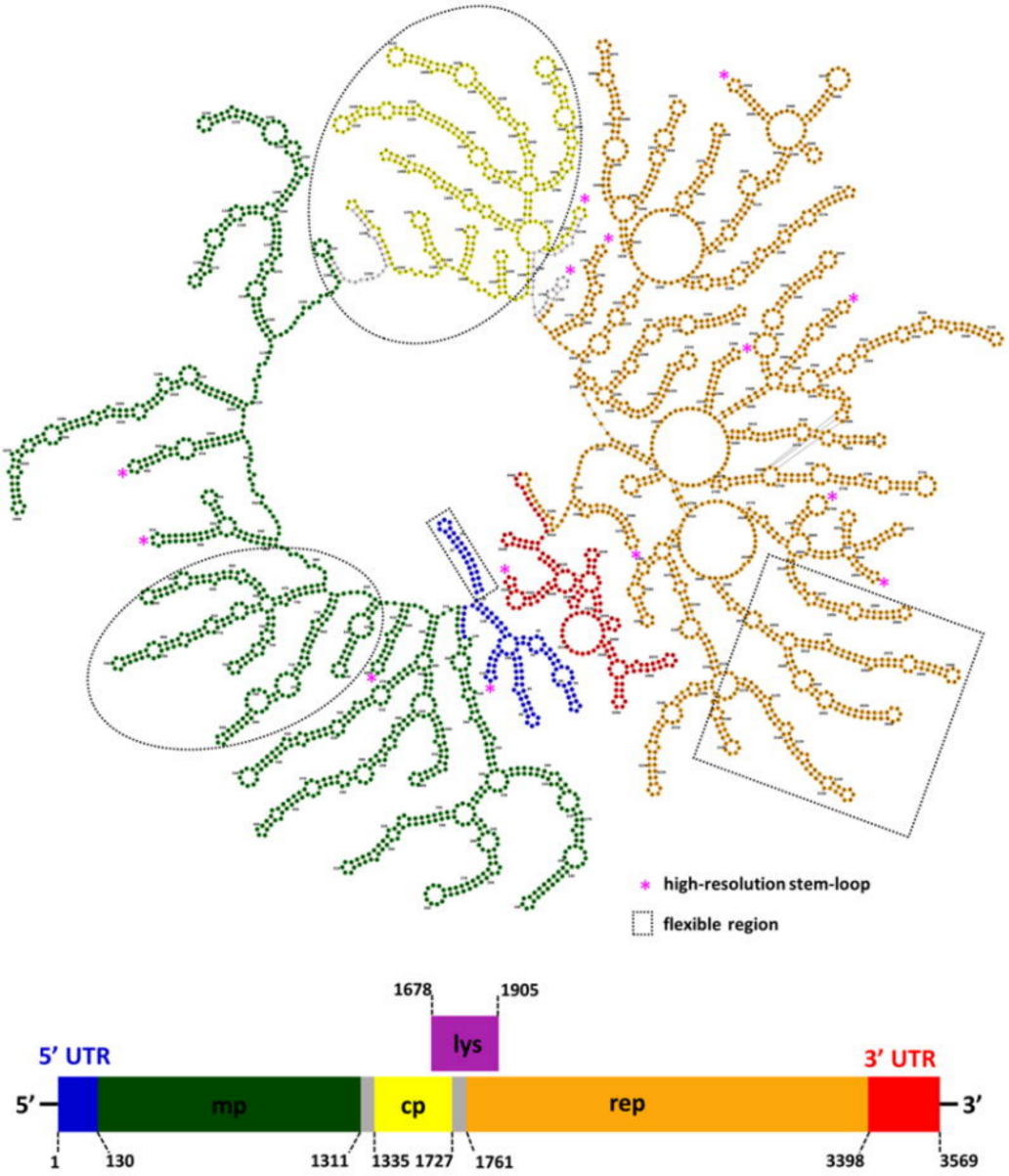


Figure 2.1: Secondary structure of the MS2 genome

The sequences are coloured according to the schematic diagram at the bottom, except for the lysis gene, which overlaps with the coat protein gene and the replicase gene. The star signs denote positions of the 16 high-resolution stem-loops. Segments enclosed with dotted boxes or ellipses are flexible [23].

2.2.2 Stemloops

Unlike dsDNA viruses which pump their genome into a preformed capsid [27, 28, 29], ssRNA viruses, such as bacteriophage MS2, co-assemble their capsid with the genome [30, 31, 32, 33]. *In situ* studies of the MS2 genome and its genome-delivery apparatus [23] by electron-counting cryo electron microscopy (cryoEM) at a resolution of 3.6Å showed that the MS2 RNA density is not uniformly distributed within the capsid and, furthermore, identified prominent major and minor grooves, hallmarks of double helices, indicating that most of the MS2 ssRNA is folded into stem-loops, with over 50 stemloops contacting the capsid via their loop regions. Among the stemloops identified in that particular study, 16 (15 contacting the coat protein and one the maturation protein) show clearly resolved individual nucleotides and even features that distinguish purines from pyrimidines. The higher resolution of the above 16 stemloops indicates stronger interactions with capsid proteins as compared to other stemloops, and thus perhaps a more important role in the capsid assembly. Three of those stemloops appearing at consecutive positions along the sequence (stemloops 1714-1737, 1746-1764, and 1766-1806), cluster together spatially and bind three neighboring coat protein dimers - a configuration desirable for nucleating capsid assembly. Indeed, the stemloop in the middle of the three, which encompasses the start codon of the replicase gene, was proposed to serve as a packaging signal involved in initiating capsid assembly [34, 35].

2.3 Proteins

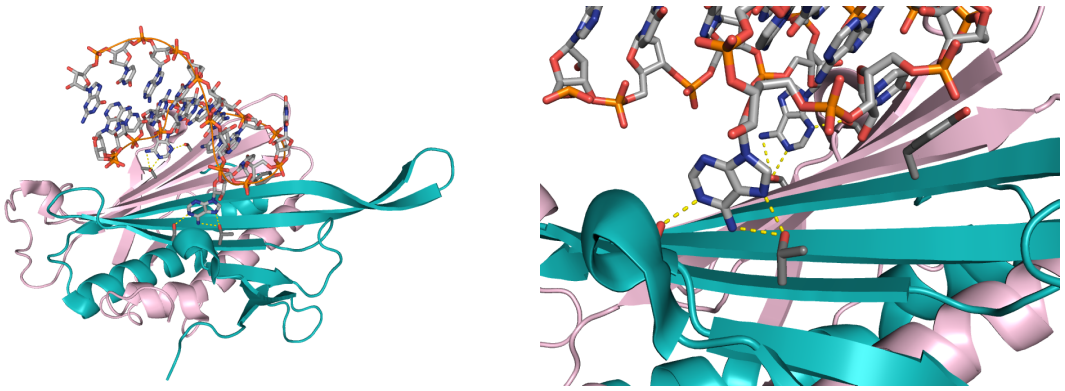
2.3.1 Maturation Protein (A-protein)

Virions of single-stranded RNA bacteriophages contain a single copy of the maturation protein (A), which is bound to the phage genome and is required for the infectivity of the particles. The A protein mediates the absorption of the virion to bacterial pili and the subsequent release and penetration of the genome into the host cell [36]. It influences the global arrangement of the virus coat dramatically: protein shells without A range between 31($\pm$1)Å and 37($\pm$1)Å, while protein shells with A have a thickness of 21($\pm$1)Å and 25($\pm$1)Å respectively, possibly by mediating the storage of energy or tension within the protein shell during virus assembly. This tension may later be used to eject the MS2 genomic RNA and A protein fragments into the host during infection [37]. The gene for the maturation protein is preceded by an untranslated leader of 130 nucleotides. The RNA in the region of the A protein CDS

folds into a cloverleaf shape i.e., three stem-loop structures enclosed by a long-distance interaction, which controls translation of A [38].

2.3.2 Coat protein

The MS2 coat protein Fig.2.2 is a member of a group of small proteins that bind to RNA in a multifunctional manner in related RNA bacteriophages. The coat protein binds and encapsidates the viral RNA but also plays a regulatory role. In the latter capacity, the protein affects translational repression of viral replicase synthesis by binding to the RNA operator of the replicase gene [39]. The MS2 coat protein is composed of 129 amino acids ($M_r = 13700$) and self-aggregates to form an icosahedral shell (180 subunits) which binds to and encapsidate the single stranded RNA MS2 genome. The coat protein consists of an N-terminal hairpin with β-strands A and B, a five-stranded anti-parallel β-sheet that forms the inner face of the capsid with strands C to G, and a C-terminal arm composed of two alpha-helices, A and B [40]. Late in the course of an infection, the coat protein binds to the translation initiation region of the replicase cistron and prevents ribosomes from initiating translation there [41].



(a) (b)

Figure 2.2: Coat protein dimer
a) A coat protein dimer binding to a RNA-stemloop structure. The two colors indicate the two individual coat proteins are depicted in different colors. b) Interactions of Serine 47 and Threonine 45 with the MS2 RNA.

2.3.3 Lysis Protein

The MS2 lysis protein (L) was the first-ever gene to be discovered as being embedded in two different genes, the coat- and replicase-encoding genes and in the +1 reading frame with respect to the two [42]. L is a 75-amino acid protein that has been reported to be present in the membrane fraction [43]. It causes lysis of *Escherichi coli* without inducing bacteriolytic activity or inhibiting net peptidoglycan (PG) synthesis [44]. For L to be able to lyse cells, it was proposed that it causes proton-motive-force depleting lesions in the inner membrane, thereby activating host autolytic enzymes such as lytic transglycosylases and D-D endopeptidases [45]. This function is embedded in the C-terminal peptide half of L, encoded by the genomic region that overlaps with the replicase gene [46]. Other results also suggest that the N-terminal peptide half, overlapping with to the coat gene, regulates its function and interacts with the host chaperon DNnaJ [44]. The translation of L is coupled to the coat protein via an RNA secondary structure, the L-hairpin, that masks the L ribosome binding site (RBS) and start codon [47]. The coupling comes about because ribosomes in the process of terminating coat translation disrupt the L-hairpin. This unmasks the L translational initiation signals and enables a low-frequency of translational initiation of the L reading frame [48].

2.3.4 Replicase

The RNA of the replicase of Bacteriophage MS2 consists of a 1637 nucleotide long stretch [24]. The synthesis of the replicase is controlled in two ways. First, translation of the replicase gene is inhibited by the MS2 coat protein, which binds to the replicase start region, thus acting as a translational repressor [49, 50, 51]. Second, replicase synthesis depends on the translation of the upstream coat protein cistron [52]. The ribosomal binding site of the replicase cistron is masked by long-distance basepairing to an internal coat cistron region. Activation of the replicase synthesis is thus sensitive to the frequency of upstream translation [53]. The replicase holoenzyme of Leviviridae consists of a phage-encoded replicase (the beta subunit) and, three host encoded proteins (alpha subunit: ribosomal potein S1; gamma subunit: EF-Tu; delta subunit: EF-Ts). The ribosomal protein S1 mediates binding of the holoenzyme to internal sites in the RNA, which allows the replicase to compete with translation for genomic RNA, since S1 also mediates the binding of the ribosome to the coat protein start site [54].

2.4 Virion

The virion of MS2 (Fig.2.3) consists of 180 subunits of the coat protein, one copy of the maturation protein and a single molecule of the positive sense RNA genome [35]. The capsid has a diameter of 280Å and exhibits a T = 3 icosahedral quasisymmetry. There are three coat protein subunits in the icosahedral asymmetric unit (named A, B and C). Because of extensive interactions of the coat protein subunits, the capsid can be considered to be assembled from 90 dimers [55]. There are two types of dimers in the capsid: AB dimers, which are located at icosahedral quasi-twofold axes, and CC dimers, which are located at icosahedral twofold axes. The only structurally important differences between the subunits are located in the FG loop, which connects strands F and G. The FG loops participate in the formation of three quasi-equivalent contacts. Two conformational variants of the FG loop are part of a sixfold arrangement at a threefold axis, whereas the remaining one is part of a pentameric interface [40].

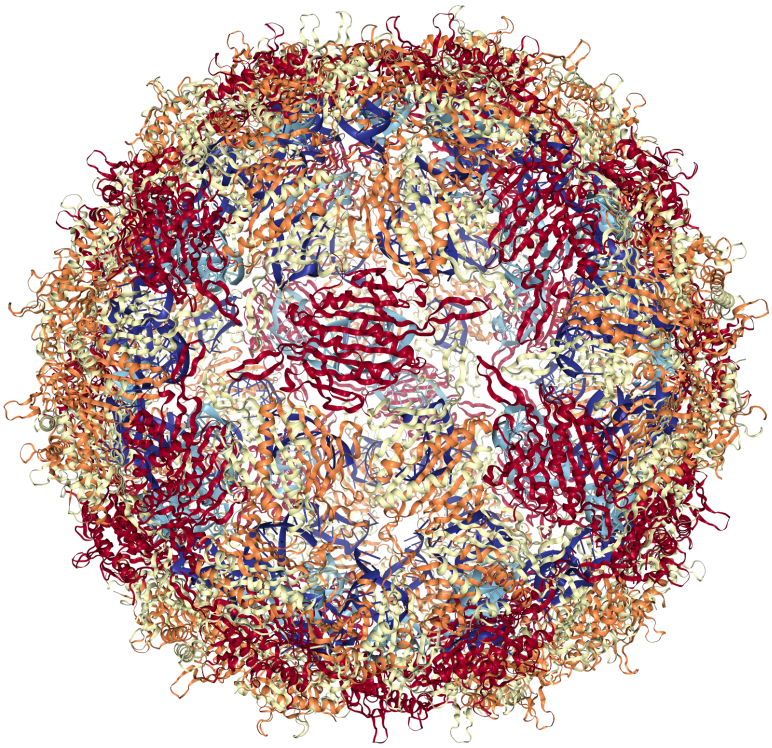


Figure 2.3: MS2 virion
The complete structure of the MS2 virion [56, 57, 58].

Chapter 3

Analysis of binding sites

3.1 Theory

Knowing that multiple MS2 coat protein dimers bind to stemloop structures of its own genomic RNA upon capsid formation [34, 35], it was our aim to assess to what degree the location and the physicochemical characteristics of the respective binding sites follow the logic of the origin of the genetic code and its generalization in the form of the complementarity hypothesis. As discussed in the introduction, the stereo-chemical hypothesis of the origin of the genetic code [59, 5, 60, 61] postulates that the code evolved as a consequence of direct binding preferences of amino acids for their cognate codons. The cognate mRNA/protein complementary interaction hypothesis [6], is a generalization of the stereo-chemical hypothesis which suggests that cognate mRNA and protein sequences may under some circumstances, in fact, be mutually physico-chemically complementary to each other and bind in a co-aligned fashion [62, 6, 21]. This implies that the MS2 coat protein should bind directly to the part of the MS2 genome that codes for it, and that this should be reflected in the matching nucleobase density profiles on the side of the capsid protein CDS and the corresponding nucleobase affinity profiles on the side of the protein itself. Moreover, it is possible that profile complementarity with the nucleobase affinity profiles of the capsid protein would extend to other known binding sites as well. These would then be taken as support of the idea that physicochemical profile matching can also be observed in a non-coding context. More practically, the idea is to apply the known nucleobase-preference scales to the MS2 coat protein sequence, derive the individual nucleobase-affinity profiles and compare them against the respective densities of the four RNA bases as well as pyrimidine (PYR) and purine (PUR) densities along every point of the MS2 genome. The result of the analysis would be Pearson correlation

coefficient profiles, which at every point reports on the degree of potential physico-chemical complementarity between the capsid protein sequence and the viral RNA. If the stereochemical hypothesis and its generalization hold, the regions where the coat protein is known to bind to the viral genome should exhibit strong negative Pearson correlation coefficients and *vice versa*. Concretely, our aim is to test the generalized complementarity hypothesis and explore its range of validity in the case of arguably the simplest biological entity in which the protein product is known to reside close to and interact with its own message i.e. a positive-sense RNA virus. Conversely, we hope to analyze to what extent the complementarity hypothesis could provide a novel mechanistic framework for explaining an important biological problem like virus capsid assembly.

3.2 Methods

3.2.1 Pearson Correlation Profiles

The 130 amino acid sequence of the coat protein peptide-chain was first substituted by the corresponding values from the specific amino acid/nucleobase affinity scales in question: the scales include the four RNA base (A, G, C, U) affinity scales as well as the PYR- and PUR affinity scales, which were all derived using a knowledge-based statistical potential formalism by Polyansky and Zagrovic [6] and the pyrimidine mimetic affinity (polar-requirement) scale, which was derived experimentally by Carl Woese and coworkers [4]. Window-averaging (window-size of 21) was applied to the obtained list of values, which reduced the length of the affinity profile down to 110, as the 10 amino-acids at each terminus of the protein could not be assigned a full window. Since neither the N-terminus nor C-terminus are known to influence binding of the coat protein to the genome, this trade-off of edge-perturbations against a higher resolution was deemed worthwhile. On the side of the viral genome, the density profiles of A, G, C, U, PYR and PUR were obtained by calculating the percentage of RNA bases per codon and applying further window-averaging (with window-size 21) over the obtained percentages. Effectively, a window-averaging with the size of 21 amino acids was introduced on the side of the coat protein, while a 63 nucleotide window-averaging was introduced on the genome side. The Pearson correlation coefficient reported at a given single nucleotide position in the genome is a result of placing the central value of the coat protein profile at the position of the nucleotide in question and calculating the level of correlation between the 110 values on the protein and the genome sides, matched in this way Fig.3.1. Since the number

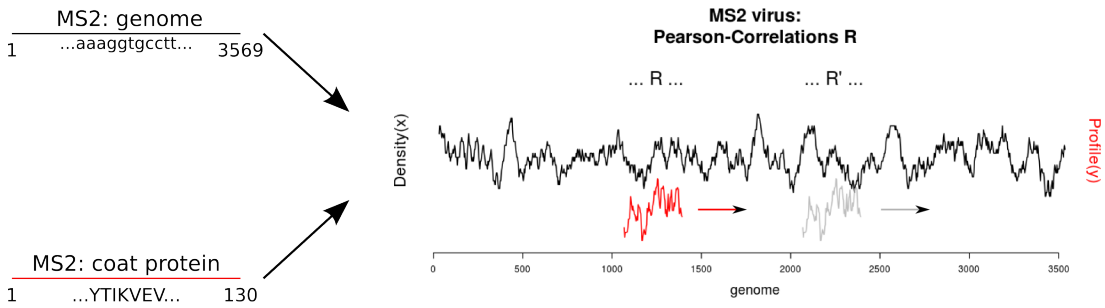


Figure 3.1: Concept
The key element of our analysis is comparison of the MS2 coat protein nucleobase affinity properties with the nucleobase density profiles of the MS2 genome at different locations.

of values in the coat protein is an even number, the calculation was performed twice, once for an overhang of 54 numbers on the left-hand side and 55 numbers on the right-hand side of the nucleotide and the second time *vice versa*, setting the reported correlation coefficient to the average of the two values. Due to this approach in combination with the window-averaging of 63 nucleotides on the genome, the first and the last 196 nucleotides of the genome are lost for profile calculation.

3.2.2 Conserved Binding Interactions

Serine 47 is an essential residue for the interaction between the coat protein and the genomic RNA, as it directly contacts the bases of the binding stemloops [23]. In order to analyze the relationship between SER47 and the bases it interacts with from the perspective of the generalized complementarity hypothesis, five locations in the genome were selected for closer inspection based on the following criteria. First, these regions had to correspond to strong negative peaks in the PYR-density/PYR-mimetic affinity profiles and second, these regions had to contain one or more of the fifteen stemloops that were structurally resolved in the cryo EM analysis by Dai and coworkers [23]. Profile comparison was carried out by aligning the coat protein polar-requirement profile and the genome at the position of the RNA base that is known to interact with SER47. The resulting Pearson correlation coefficient R was then compared with the optimal match (lowest Pearson R) in the region in question.

3.3 Results

3.3.1 Pearson correlation profiles

3.3.1.1 General Properties

The most interesting finding to note is that regardless of which combination of genome nucleobase density profiles and coat protein nucleobase/amino acid affinity profiles one looks at, the resulting correlation profiles invariably exhibit a periodic alternation between negative and positive values of Pearson correlation coefficients R. Most of the R-values lie between ± 0.5, while the most prominent peaks reach close to ± 0.6 and ± 0.7, with several exceptions reaching the values around ± 0.8, mainly in the U-density profiles Fig:3.5. According to the complementarity hypothesis, the binding is expected to take place in regions with strong negative values of the Pearson R. For this reason, the peaks that were able to cross the threshold of -0.5 were treated as potential binding sites and further examined.

3.3.1.2 Predicted binding sites

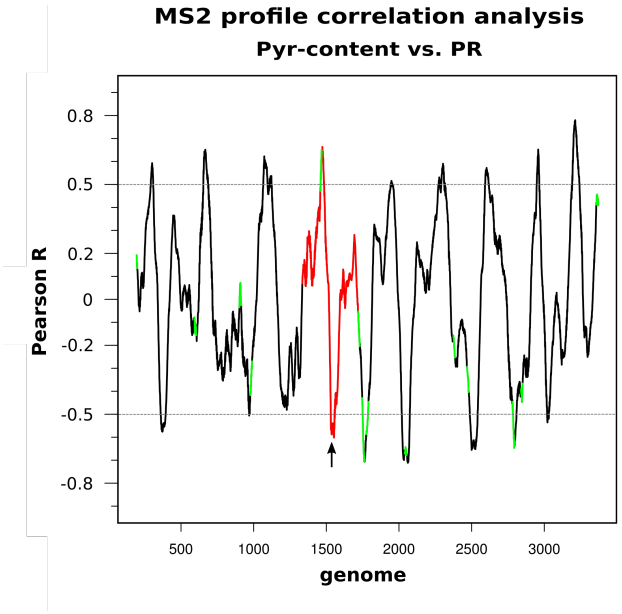
Comparing the binding sites predicted on the basis of the generalized complementarity hypothesis and using the experimentally known binding sites in the genome as a reference, there are two types of profiles that come closest to predicting the real binding sites. The first type of profiles are the diverse nucleobase densities when compared against the respective nucleobase/amino acid affinity scales, while the second type are those that were obtained by using the PYR-mimetic affinity scale form Woese and coworkers (the PR scale). The most significant result was obtained when comparing the RNA PYR-density profile and the coat protein PR profile (Fig.3.2) which was able to identify 6 known binding sites in peak areas below -0.5 and 3 additional known binding sites located close to the negative peak areas in R profiles, with the residues of the known binding sites contributing to the correlation of the minimum of the peak. It is interesting to note that in the latter case, the binding site that lies in the CDS of the coat protein aligns well with a positive peak area, which would indicate an area of non-binding.

As in principle, each RNA base type contributes to the binding of the coat protein dimer to the genome, it is reasonable that the binding sites should be predictable only when looking at all the four different RNA base profiles in combination. Remarkably, this indeed is the case as a combination of all four base profiles (RNA base density versus respective base preference) together with the PYR-PR profile are able

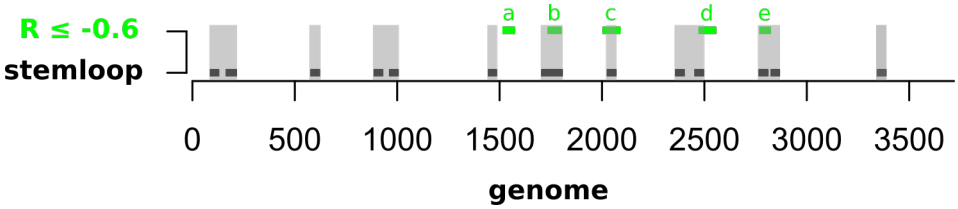
to predicted 10 out of the 15 known binding sites, as seen in Fig.3.3. Comparing the known binding sites i.e. the stemloops[23] against the co-purified cDNA fragmnets from CLIP-Seq data [17], small deviations can be seen. However, our analysis was able to capture both of their characteristics very well. Due to the limitations of the sliding technique used, when it comes to window-averaging, it is impossible to detect the first and the last binding stemloop as these regions are lost in the calculations and can, thus, not be evaluated. If these two structures are subtracted, it is clear that our simplistic analysis that follows the logic of the complementary hypothesis is able to identify 10 binding regions out of 13 known binding regions.

3.3.1.3 Conserved Binding Interactions

In contrast to the above results, one could not observe any strong indication of the interaction of SER47 and the respective RNA nucleobase in individual stemloops by aligning the genomic PYR-density profile and the coat protein PYR-mimetic (PR) affinity profile. The distance between SER47 and the interacting residue ranges from 7 to 26 residues and even though the coat protein profile overlaps with the residue in the genome, meaning it contributes to the respective Pearson correlation at the minimum, we could not identify any regular pattern concerning their positions Fig.3.4. Comparing the correlation coefficients as well as the general shape of the density against the shape of the coat protein profile, for both the SER47-aligned and the centered approach, it is clear that both options give less negative and, thus, weaker correlation coefficients than optimal. Neither the shape of the density at those locations nor the comparison to the coat protein profile reveal any reoccurring or prominent trends.



(a)



(b)

Figure 3.2: Pearson Correlation: PYR-content/PR

a) Pearson correlation coefficient profile of MS2 genome PYR density profile vs. MS2 coat protein PYR-mimetic affinity profile (PR profile) The coding region of the coat protein is indicated in red, while the green stretches in the map correlate to the locations of the binding stemloops in the genomic RNA. The arrow depicts the area where the Pearson R value is calculated by comparing the middle of the coat protein CDS against the middle of the coat protein amino acid sequence. b) locations of the 5 peaks from a) that have a corresponding Pearson R value $\leq$ -0.6, against the location of the coat protein binding RNA stemloops [23].

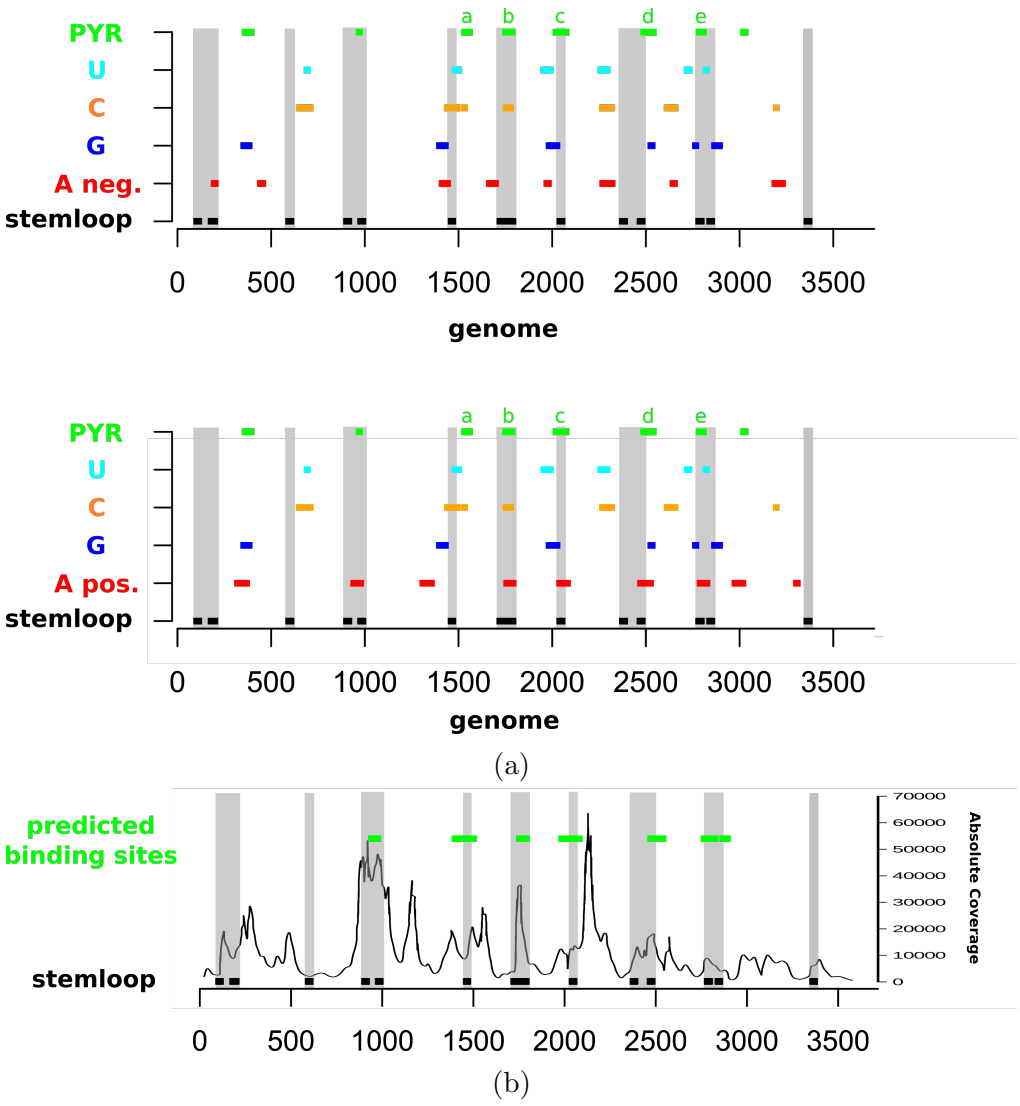


Figure 3.3: Predicted binding regions

a) The individual regions predicted to be bindings sites based on correlation profiles of the four RNA-bases using their respective preference scale, as well as the PYR-content against the PYR-mimetic affinity scale from Woese (PR scale) [4]. The bars indicate regions in the profiles with Pearson R values below or equal to -0.5, with A neg. ≤ -0.5 and A pos. ≥ 0.5. We show both values for A because of the peculiar behaviour of A-affinity profiles as discussed in the Introduction. The black bars indicate the regions of the 15 stemloops that are known to bind to a coat protein dimer. The letters a-e) indicate the five highest peaks and were selected for further investigation as seen in Figure A.6. b) The summation of the predicted binding regions which overlay with the known binding sites, based on the four nucleobase affinities (A pos.) and the PYR-mimetic affinity. There are three consecutive stemloops located between positions 1714 to 1806 in the genome (containing the packing signal), which appear as a single black bar. The absolute coverage profile (black) provides the experimentally (CLIP-Seq sequencing) deduced abundances of cDNA fragments (MS2 virus genome) that co-purified with the MS2 coat protein [63].

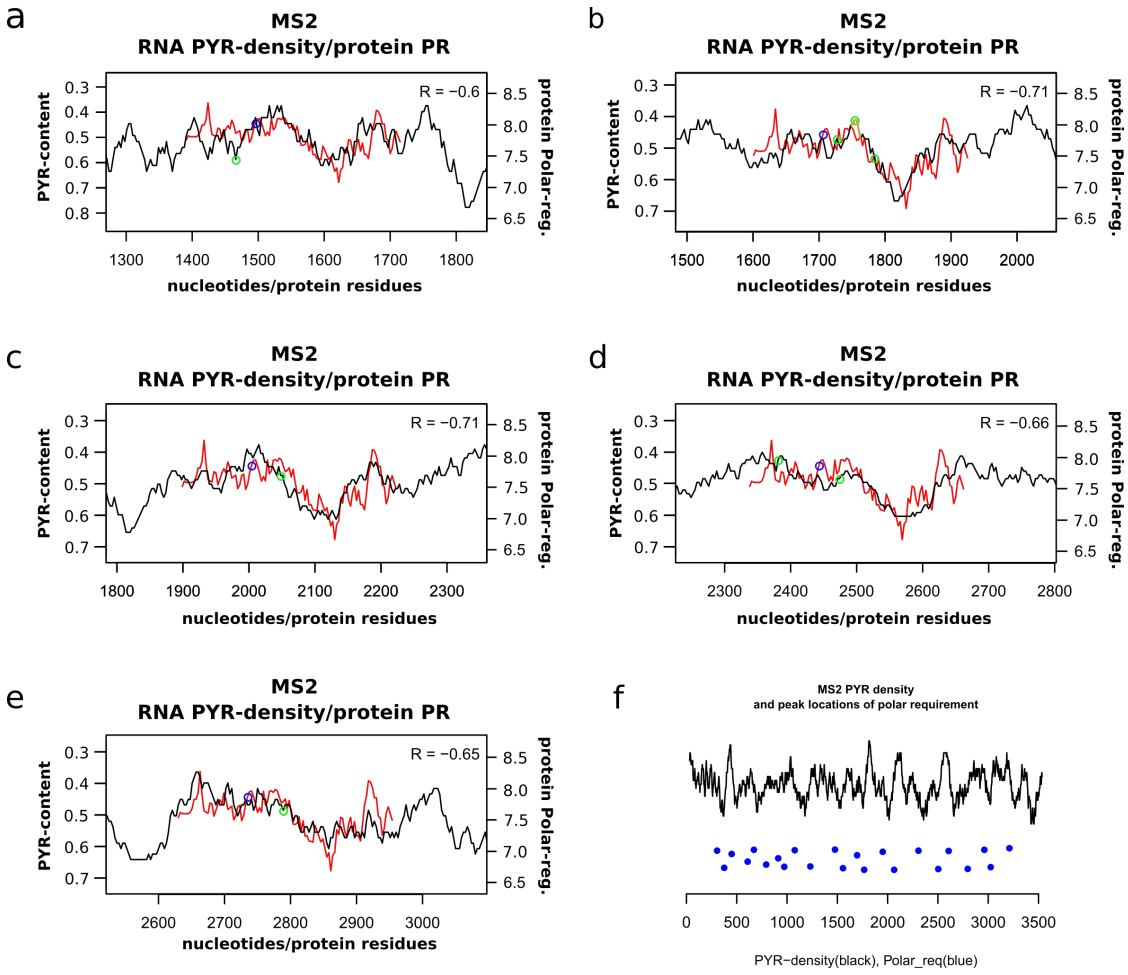


Figure 3.4: Conserved binding interactions

a-e) The five most prominent peak locations in the genome were selected from the PYR-PR correlation. RNA PYR density in black and the coat protein PR profile in red. The coat protein PR profile was placed at the location of the nucleotide that gives the corresponding minimal R value. SER 47 is marked as a blue circle in the protein profile and the residues that SER 47 binds to in the genome are marked as a green circle. c) The orange stretch in the PYR-density is the stemloop-region of the packing signal, which is thought to be responsible for the initiation of encapsidation. f) PYR-density of the MS2 genome and the location of the individual peaks from the Pearson correlation file.

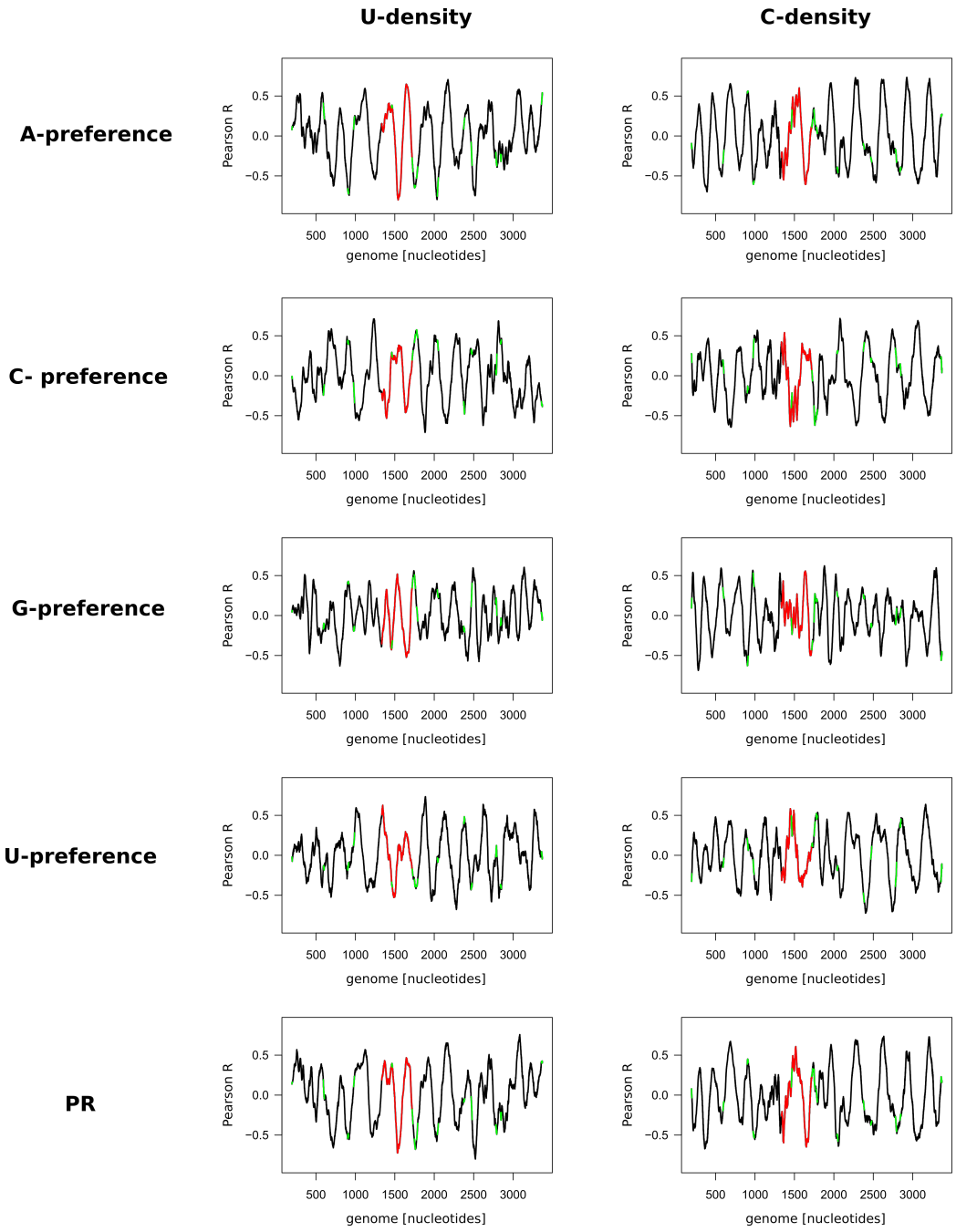


Figure 3.5: Pearson Correlation Profiles: U/C

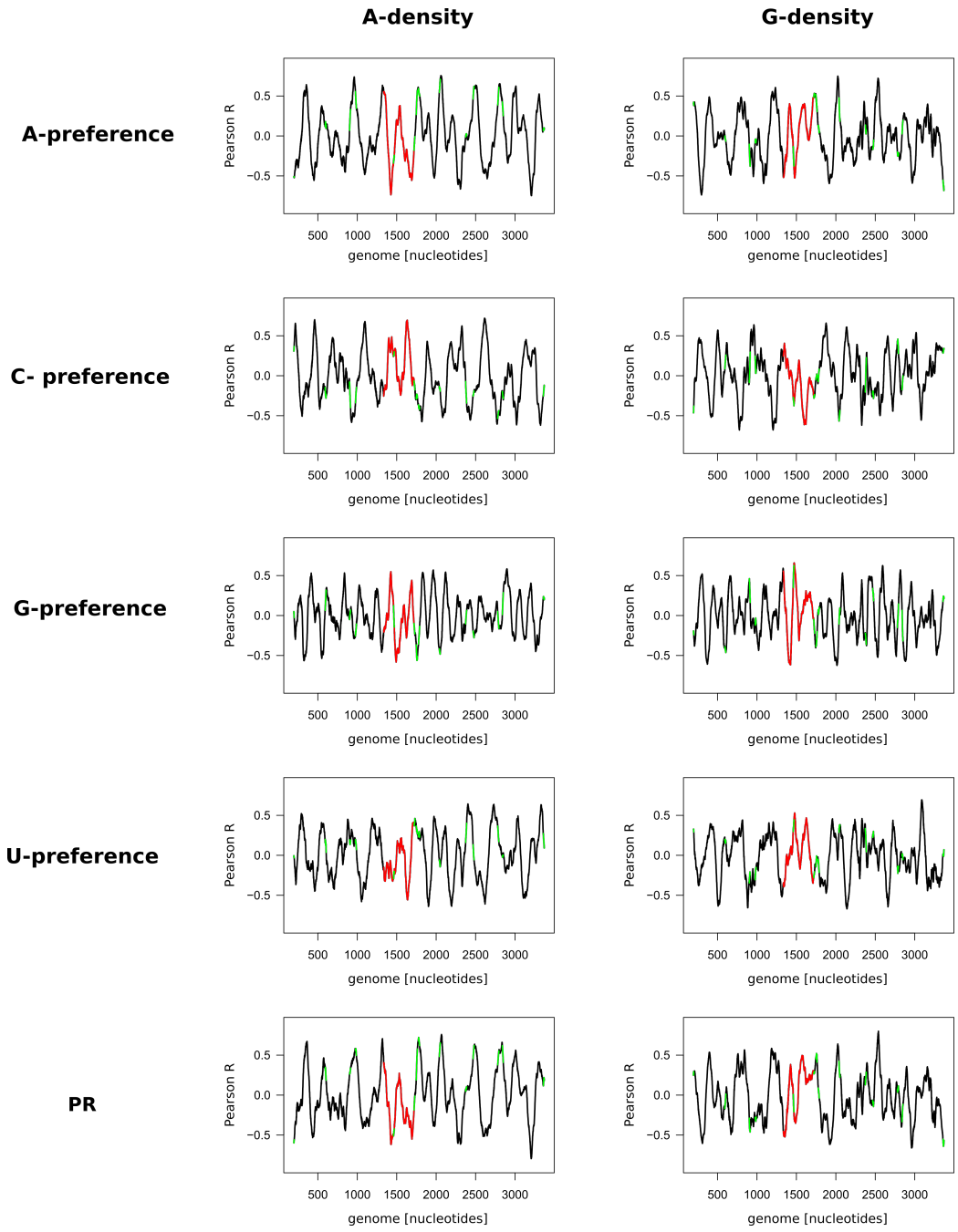


Figure 3.6: Pearson Correlation Profiles: A/G

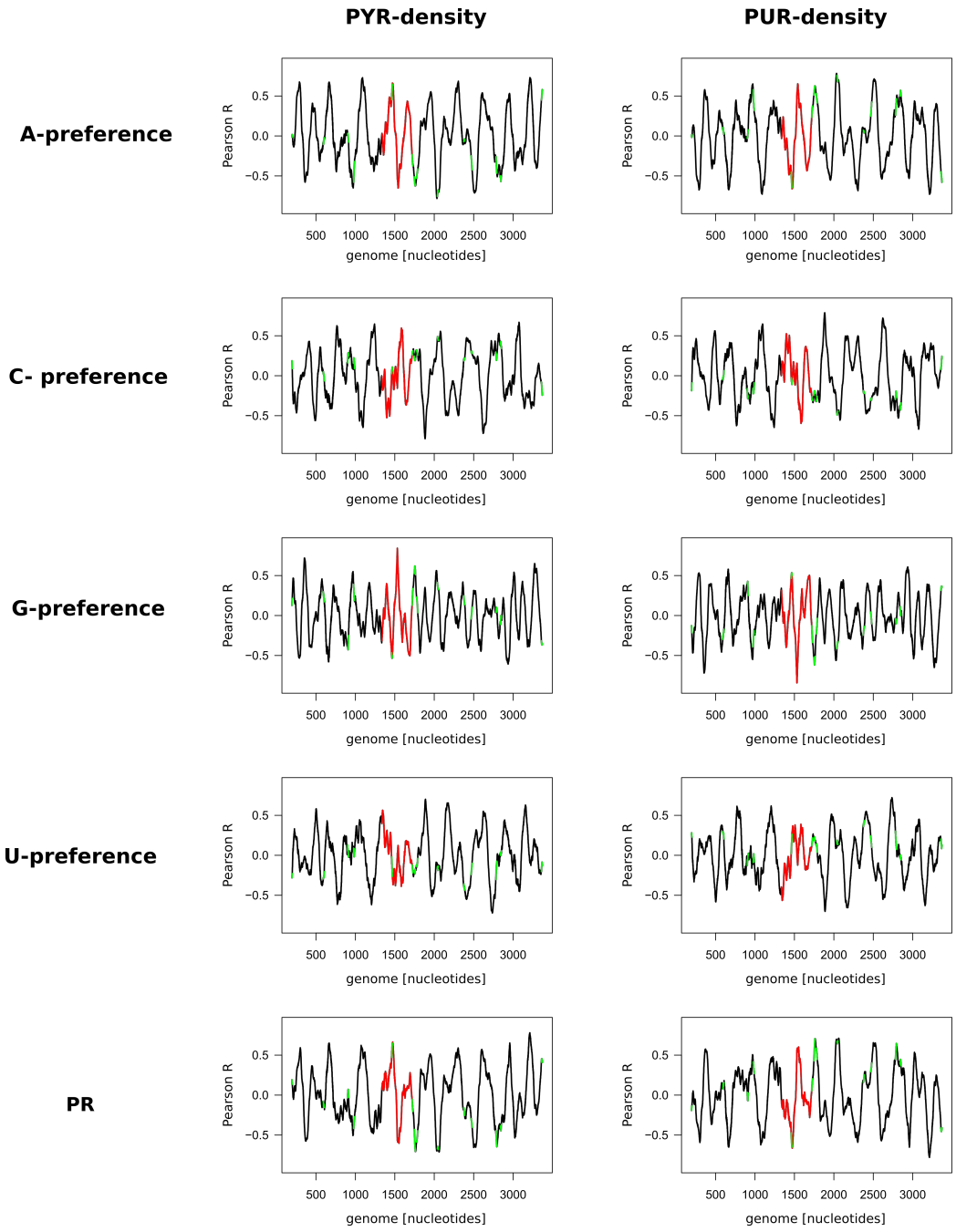


Figure 3.7: Pearson Correlation Profiles: PYR/PUR

Chapter 4

Fourier Transform and Periodicity

4.1 Introduction

The different Pearson correlation R profiles discussed above all exhibit a seeming periodicity in their features. As mentioned above, the coat protein dimers bind to similar secondary structures i.e. stemloops throughout the whole viral genome. This gives rise to a possibility that these regions of interaction may not only be similar in their secondary structure and composition, but that they could also be periodically arranged in the MS2 viral genome. To address this hypothesis and ask if the obtained periodicity exhibits any relation to the properties of the MS2 genome and its four proteins, we have carried out a Fourier transform analysis of different profiles discussed above.

4.2 Theory [1]

Fourier analysis is the study of the way different mathematical functions can be represented or approximated by sums of simpler trigonometric functions. The decomposition of a waveform, which can be a function of time, space or other variables, into a sum of simple sinusoids of different frequencies and amplitudes is called the Fourier transformation. The Fourier analysis is an indispensable analytical tool in a wide range of scientific applications including, physics, signal processing, digital image processing, statistics, diffraction, geometry, protein structure analysis and many other areas.

4.2.1 (Continuous) Fourier transform

In most cases, the term *Fourier transform* refers to a transformation of a function with a continuous real argument (signal) into a continuous function of frequency, known as a frequency distribution. The Fourier transform of a function *f(t)* is defined as:

$$\hat{f}(s) = \int_{-\infty}^{\infty} f(t) \times e^{-2\pi ist} dt \tag{4.1}$$

It is assumed that *f(t)* is defined for all real numbers *t*. For any $s \in \mathbf{R}$, integrating *f(t)* against $e^{-2\pi ist}$ with respect to t produces a *complex valued* function of s, that is, the Fourier transform $\hat{f}(s)$ is a complex-valued function of $s \in \mathbf{R}$. If t has the dimension of time, then to make st dimensionless in the exponential, $e^{-2\pi ist}$ s must have the dimension of 1/time. The domain of the Fourier transform is the set of real numbers s. One says that $\hat{f}$ is defined in the *frequency domain* and that the original signal *f(t)* is defined in the *time domain* (or the *spatial domain*, depending on the context). For a (nonperiodic) signal defined on the whole real line, we generally do not have a discrete set of frequencies as in the periodic case, but rather a *continuum* of frequencies. The set of all frequencies is the *spectrum* of *f(t)*. The squared magnitude $|\hat{f}(s)|^2$ is called the *power spectrum* (especially in connection with its use in communications) or the *spectral power density* (especially in connection with its use in optics) or the *energy spectrum* (in other physics applications). An important relation between the energy of the signal in the time domain and the energy spectrum in the frequency domain is given by Parseval's identity for Fourier transforms:

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{f}(s)|^2 ds \tag{4.2}$$

4.2.2 Discrete Fourier transform (DFT)

4.2.2.1 Definition

In contrast to the Continuous Fourier transform the discrete Fourier transform deals with signals and functions, that are both limited in time and band, with the knowledge that this can only approximately be true. It is assumed that *f(t)* is zero outside of $0 \leq t \leq L$. It is also assumed that the Fourier transform *F f(s)* is zero, or effectively zero (beyond our ability to measure i.e. of negligible energy) outside of $0 < s < 2B$. Instead of thinking in terms of sampled values of a continuous signal and sampled value of its Fourier transform, one may think of the discrete Fourier transform as an operator that accepts as input a discrete list of N values and returns as output a

discrete list of N values.
For the definition of the discrete Fourier transform let $\mathbf{f} = (\mathbf{f}[0], \mathbf{f}[1], ..., \mathbf{f}[N-1])$ be an N-tuple. The *discrete Fourier transform (DFT)* of $\mathbf{f}$ is the N-tuple $\mathbf{F} = \mathbf{F}[0], \mathbf{F}[1], \mathbf{F}[N-1])$ defined by:

$$\mathbf{F}[m] = \sum_{n=0}^{N-1} \mathbf{f}[n] e^{-2\pi i m n/N}, m = 0, 1, ..., N-1 \tag{4.3}$$

4.2.2.2 Periodicity

The definition of the DFT suggests some additional structure concerning the outputs and inputs. The output values $\mathbf{F}[m]$ are defined initially only for m=0 to m=N-1 , but their definition as

$$\mathbf{F}[m] = \sum_{k=0}^{N-1} \mathbf{f}[k] \omega^{-km} \tag{4.4}$$

implies a periodicity property. Since

$$\omega^{-k(m+N)} = \omega^{-km} \tag{4.5}$$

one has

$$\sum_{k=0}^{N-1} \mathbf{f}[k] \omega^{-k(m+N)} = \sum_{k=0}^{N-1} \mathbf{f}[k] \omega^{-km} = \mathbf{F}[m] \tag{4.6}$$

If one considers the left-hand side as the DFT formula producing an output, then that output would be $\mathbf{F}\ [m + N]$. More generally, and following the same kind of calculation, we would have

$$\mathbf{F}[m + nN] = \mathbf{F}[m] \tag{4.7}$$

for any integer n. Thus, instead of just working with $\mathbf{F}$ as an N-tuple it's natural to "extend" it to be a periodic sequence with a period of N.

4.2.3 Fast Fourier transform (FFT)

The FFT is an algorithm for computing the DFT with fewer than N^2 multiplications i.e. it brings down the required number of computation steps from the initial N^2 to $Nlog_2N$ multiplications for the sorting, and the number of additions needed in the algorithm down to $3log_2N$. The FFT starts by separating $\mathbf{f}[n]$ into two sequences with even and odd indices (0 is even), each of length $N/2$. The general shape of the factorization to get DFT_N (the solution of a normal DFT) via $DFT_N/2$ is:

$$\underline{\mathbf{F}}_{\mathbf{N}} = \begin{Bmatrix} I_{N/2} & \Omega_{N/2} \\ I_{N/2} & -\Omega_{N/2} \end{Bmatrix} \begin{Bmatrix} \underline{\mathbf{F}}_{\mathbf{N/2}} & 0 \\ 0 & \underline{\mathbf{F}}_{\mathbf{N/2}} \end{Bmatrix} \begin{Bmatrix} \text{sort the even} \\ \text{and odd indices} \end{Bmatrix} \tag{4.8}$$

where $I_{N/2}$ is the $N/2 \times N/2$ identity matrix. 0 is the zero matrix (of size $N/2 \times N/2$ in this case). $\Omega_{N/2}$ is the diagonal matrix with entries $1, \omega_N^{-1}, \omega_N^{-1}, ..., \omega_N^{N/2-1}$ (ω being the vector complex exponential, with $\omega = 1, \omega, \omega^2, ..., \omega^{N-1}$, where $\omega = e^{2\pi/N}$) down the diagonal. $\underline{\mathbf{F}}_{\mathbf{N/2}}$ is the DFT of half the order, and the permutation matrix puts the $N/2$ even indices first and the $N/2$ odd indices second. The next step is to repeat the algorithm, each time halving the size of the DFT. For the next level "down" this would be,

$$\underline{\mathbf{F}}_{\mathbf{N/2}} = \begin{Bmatrix} I_{N/4} & \Omega_{N/4} \\ I_{N/4} & -\Omega_{N/4} \end{Bmatrix} \begin{Bmatrix} \underline{\mathbf{F}}_{\mathbf{N/4}} & 0 \\ 0 & \underline{\mathbf{F}}_{\mathbf{N/4}} \end{Bmatrix} \begin{Bmatrix} \text{sort } N/2\text{-lists to} \\ \text{two } N/4\text{-lists} \end{Bmatrix} \tag{4.9}$$

and inserting this result into the previous equation, the operations become "nested" (recursive):

$$\underline{\mathbf{F}}_{\mathbf{N}} = \begin{Bmatrix} I_{N/2} & \Omega_{N/2} \\ I_{N/2} & -\Omega_{N/2} \end{Bmatrix}$$
$$\begin{Bmatrix} \begin{Bmatrix} I_{N/4} & \Omega_{N/4} \\ I_{N/4} & -\Omega_{N/4} \end{Bmatrix} \begin{Bmatrix} \underline{\mathbf{F}}_{\mathbf{N/4}} & 0 \\ 0 & \underline{\mathbf{F}}_{\mathbf{N/4}} \end{Bmatrix} \begin{Bmatrix} N/2\text{-lists to} \\ N/4\text{-lists} \end{Bmatrix} & 0 \\ 0 & \begin{Bmatrix} I_{N/4} & \Omega_{N/4} \\ I_{N/4} & -\Omega_{N/4} \end{Bmatrix} \begin{Bmatrix} \underline{\mathbf{F}}_{\mathbf{N/4}} & 0 \\ 0 & \underline{\mathbf{F}}_{\mathbf{N/4}} \end{Bmatrix} \begin{Bmatrix} N/2\text{-lists to} \\ N/4\text{-lists} \end{Bmatrix} \end{Bmatrix}$$
$$\begin{Bmatrix} \text{sort } N/2\text{-lists to} \\ \text{two } N/4\text{-lists} \end{Bmatrix} \tag{4.10}$$

To repeat this and keep halving the size of the DFT, N needs to be a power of 2. If the initial signal does not equal to a power of 2, *Zero Padding* is introduced, which consist of adding of zeros at the end of the initial signal until N is a power of 2. The construction then continues "going down" levels until it reaches from $\mathbf{F}_N$ to the DFT of order 1, $\mathbf{F}_1$, which takes a single input and returns it unchanged. After the halving is over the remaining work consist of initial sorting and reassembling. Thus reading from right to left, the initial inputs $(\mathbf{f}[0], \mathbf{f}[1], ..., \mathbf{f}[N-1])$ are first sorted and then passed back up through a number of reassembly matrices, ultimately winding up as the outputs $(\mathbf{F}[0], \mathbf{F}[1], ..., \mathbf{F}[N-1])$. The entire process from $\mathbf{f}$'s to $\mathbf{F}$'s is called the Fast Fourier Transform because of the reduction in the number of operations.

4.3 Methods

Discrete Fourier Transforms were calculated using an FFT algorithm implemented in Python for both original data sets, consisting of Pearson correlation profiles and the PYR-density profile, and different artificial data sets, consisting of Pearson correlation profiles that were produced by randomizing or shuffling either the MS2 genome sequence, its coat protein amino acid sequence or changing the number of amino acids of the coat protein. Due to the computational burden, only profiles obtained by comparing PYR-content of the genome against PYR-mimetic affinity profiles (PR profiles) of the coat protein were analyzed for both type of data sets. In the course of these calculations, different window-sizes for averaging, as described in the Methods section, were introduced for both the original and artificial data sets. Window-sizes for the Pearson correlation profiles ranged from 5 amino acids (for the coat protein) and 15 nucleotides (for the MS2 genome) "05_15" to 60 amino acids and 180 nucleotides "60_180", in steps of +5 amino acids i.e. 15 nucleotides, keeping the ratio of 1 amino acid against 1 codon. For the PYR-density profile, the additional window-sizes of 300, 450, 600, 750 and 900 nucleotides were introduced. The randomization for the artificial data set was implemented using two different shuffling procedures: first, the genome sequence and second, the coat protein sequence were shuffled (with 10^4 calculations for each). Furthermore, there were four additional alterations on the side of the coat protein that were investigated separately: 1) shortening of the coat protein sequence to 65 amino acids (with amino acids subtracted from the original from both ends symetrically), 2) extending of the coat protein to a length of 190 amino-acids (the first half of the initial peptide-chain was added to the end of the original chain), and 3)-4) two extensions to the length of 160 amino acids, one time adding the first 30 amino-acids of the peptide to the end and the second time adding randomly chosen amino acids to the original chain. *Spectral power densities* were extracted from the calculated Fourier transforms and used to identify the 5 highest periodicity numbers for each transform. These periodicity numbers were then further used to calculate the length (given in the number of amino acids) of the fragment with the highest periodicity. For the artificial data sets, the average over all individual *spectral power densities* was calculated for each randomization/shuffling method.

4.4 Results

4.4.1 Primary data-set

The periods with the highest spectral power density for the Pearson R profiles for PYR/PR comparisons fall in the range from 132.4 to 132.5 amino acids for all different windows used for averaging, except for the two smallest windows (05_15 and 10_30), where the period lengths are 14.5 and 42.5 amino acids, respectively Fig.4.1. However, even in these two cases, a periodic fragment of a length around 132.5 amino acids can be found among the top spectral power density peaks Fig.4.2. The period lengths close to 132 amino acids are obtained by a periodic number of 8, with slight differences from this value expected due to *Zero Padding*.
Interestingly, the MS2 genome pyrimidine content Fourier transform, as obtained by using different averaging windows exhibits a similar behaviour as the previous Pearson R profiles above, as periodic lengths of around 132.5 amino acids can be found as either corresponding to the highest peak in the power spectrum or being among one of the 5 highest peaks Fig.4.3.

4.4.2 Artificial data sets

Two representative averaging windows (10_30 and 21_63) were chosen in order to analyze the effect of window-averaging on the above findings. These specific window sizes were chosen since the spectra corresponding to different profiles show the greatest change when going from the smallest window of 05_15 to the window of 20_60, with practically negligible differences for bigger window-sizes. The window 21_63 was chosen instead of 20_60 due to the fact that it was used in all previous analyses. Given the length of the coat protein, the window-sizes are capped at 60_180 except for the shortened set where the cap is at 30_90.

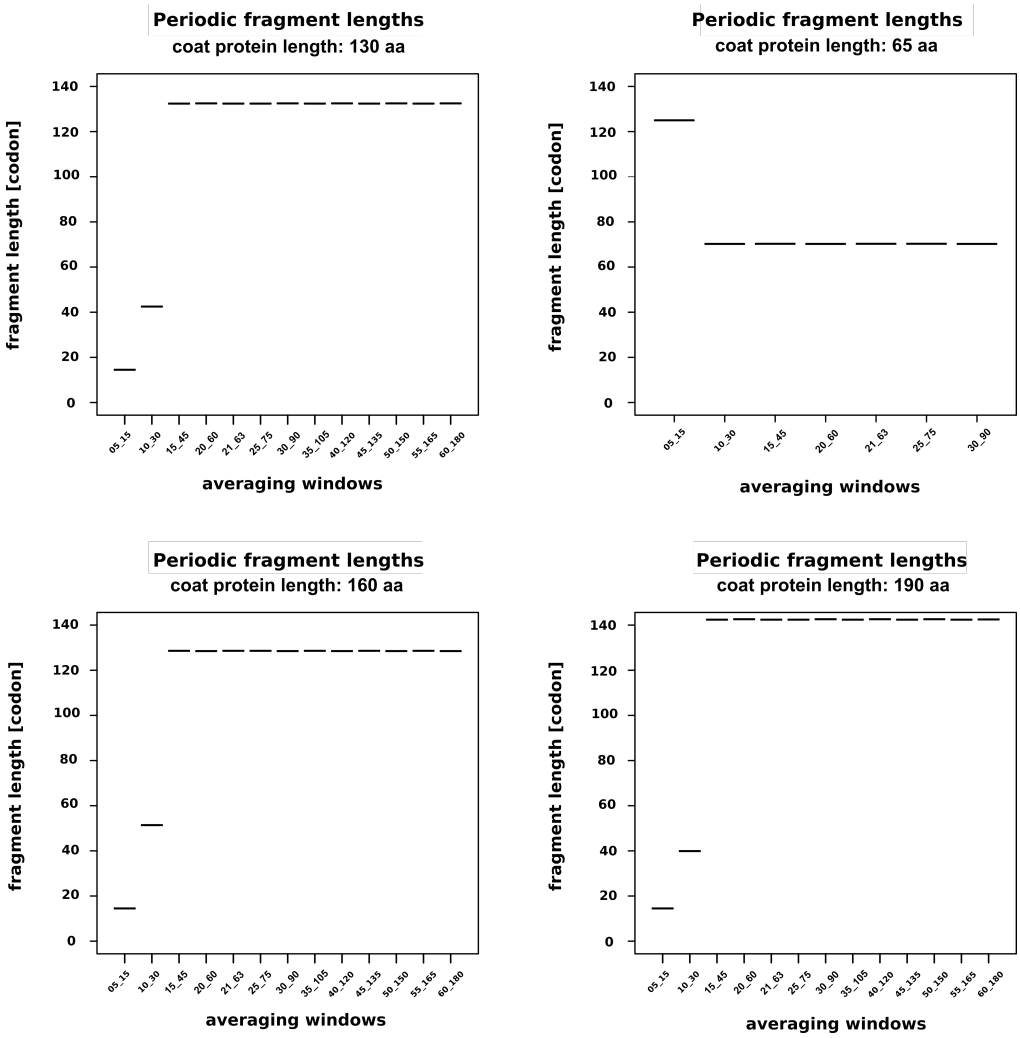


Figure 4.1: Periodicity fragments: highest spectral density

The figures show the length of the periodic fragment, corresponding to the periodicity peak with the highest spectral power that was obtained by Fourier transform, when using different averaging windows. The left top shows the values for an unchanged coat protein length, the individual values are 14.5, 42.5 and 132.5 ±0.1. The right top correlates to values when the length of the coat protein was reduced to 65 aa, while values are 125.0 and 70.25 ±0.06 for the others. The left bottom is calculated with a coat protein length of 160 and the values do not differ for either of the two different extensions that were used for this particular length (the values are 14.5, 51.4 and 128.6 ±0.2). The figure on the right bottom shows the values for the extension of the coat protein to a length of 190 aa with the corresponding values of 14.4, 39.9 and 142.4 ±0.2.

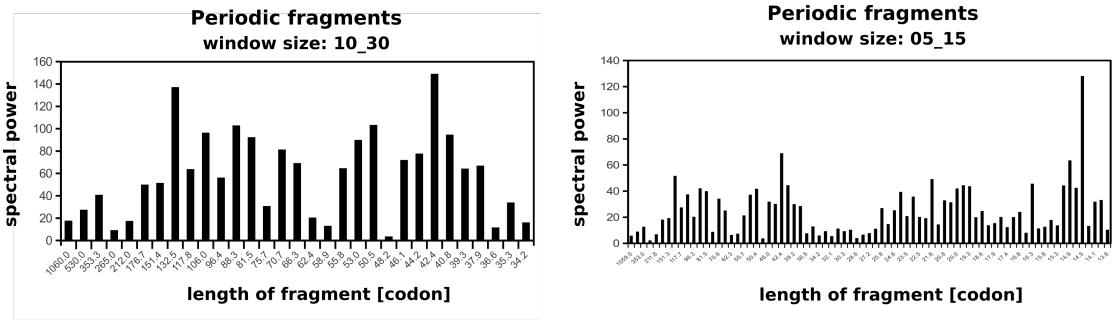


Figure 4.2: Fragment Spectra: 05_15 & 10_30
Spectra obtained by Fourier transform for the window sizes 05_15 and 10_30.

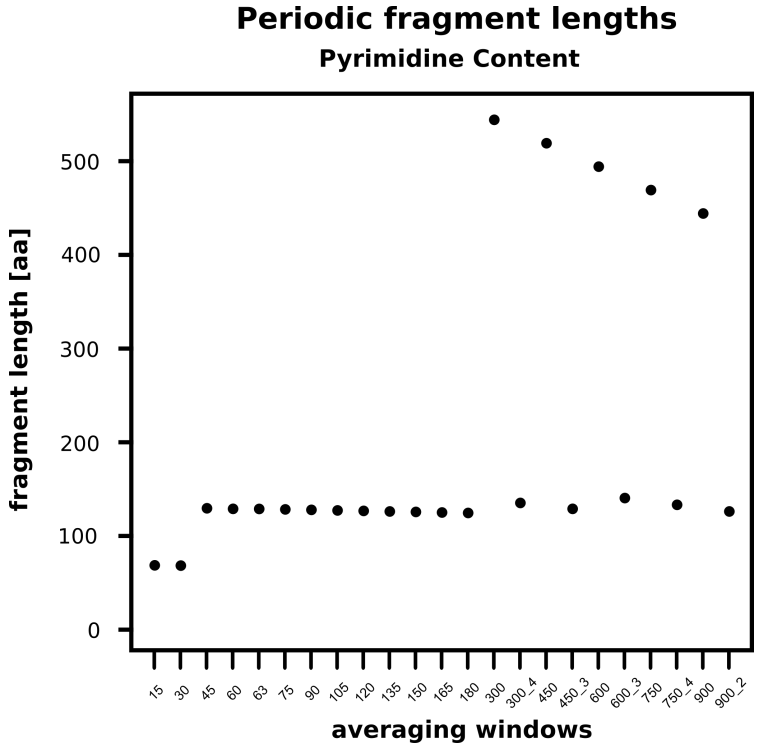


Figure 4.3: Periodic fragments: Pyrimidine content
The values of the highest spectral power for different window-sizes. Starting from a window-size of 300, a second value is included as well. The number after the window-size, in the label on the x-axis, is the rank of the peak in the individual spectra.

4.4.2.1 Investigating the original coat protein length

When shuffling the residues of the coat protein, the highest peak in all 10^4 power spectra does not exceed the value of 132.5 amino acids. Smaller averaging windows in general exhibit periodic fragments with a shorter length, while for larger window-sizes (over 21_63), the fraction of spectra with the highest peak exhibiting a value of around 132.5 amino acids increases. Averaging of different periodic numbers over the complete FFT spectra (per given averaging window size) shows that the highest peak in the original FFT spectra is also the most abundant peak in all of the shuffled coat protein FFT spectras Fig.4.5. In contrast to this, shuffling of the genome reveals a largely different effect. First, the periodic fragment lengths, as obtained by the highest peaks, are typically larger or smaller than 132.5. The distribution of periodic numbers, for individual window-sizes larger than 15_45, resembles a left-steep and right-skewed Poisson-distribution and this trend gets stronger the larger the window size Fig.4.4. Second, the distributions of the periodic numbers for individual window sizes are more evenly distributed with one or two clearly defined maxima Fig.4.5. If, however, the period numbers from the non-shuffled spectra are compared against the averaged period number spectra, it appears that the non-shuffled values are still in close proximity to the most abundant values from the shuffled ones.

4.4.2.2 Analysis of shortened coat protein sequences

Calculations performed with the shortened coat protein show a slightly different trend in that the smallest averaging window (05_15) results in the longest periodic fragment (125.0 amino acids), while all other windows give a shorter periodic fragment (70.25 $\pm$0.06 amino acids) Fig.4.1. There are also differences for the shuffling of the coat protein sequence, as the values that exceed the non-shuffled short coat protein spectra lengths can be found, even though they only make up a small fraction of the total values Fig.4.7. The averaged periodic number spectra reveal multiple peaks with an irregular distribution Fig.4.6.

In accordance with the previous results for shuffled genomes, the short coat protein spectra closely resemble the left-steep and right-skewed Poisson-distribution: the resemblance can be seen starting from a window-size of 10_30 and upwards Fig.4.7. Since the distributions closely match the previous results, the distribution of periodic numbers is also expected to follow the previously mentioned trend, as it indeed does Fig.4.6.

4.4.2.3 Long coat protein length

The extension of the coat protein sequence to 190 amino acids and the two different extensions to 160 amino acids lead to results that are similar to that acquired with the original coat protein length. The only slight changes are that the length of the periodic fragments are slightly different: for both of the 160 extensions the values are off by approximately ±5 amino acids, and for the 190 adjustment, for the window-sizes 20_60 and longer, the length of the periodic fragment is 142.5 ±0.2 Fig.4.1. The only exception to this is a single value for the window 15_45 in the 160 extension with the randomly added amino acids, as not the first but the second highest peak in the spectrum corresponds to a value of 128.5 amino acids.
Shuffling of either the coat protein sequence or the genome leads to similar results as the distribution described in the above *Original coat protein length* section, with the only difference being the exact values and the fact that in the shuffled coat spectrum a small fraction is able to exceed the initial highest value of the non-shuffled one.

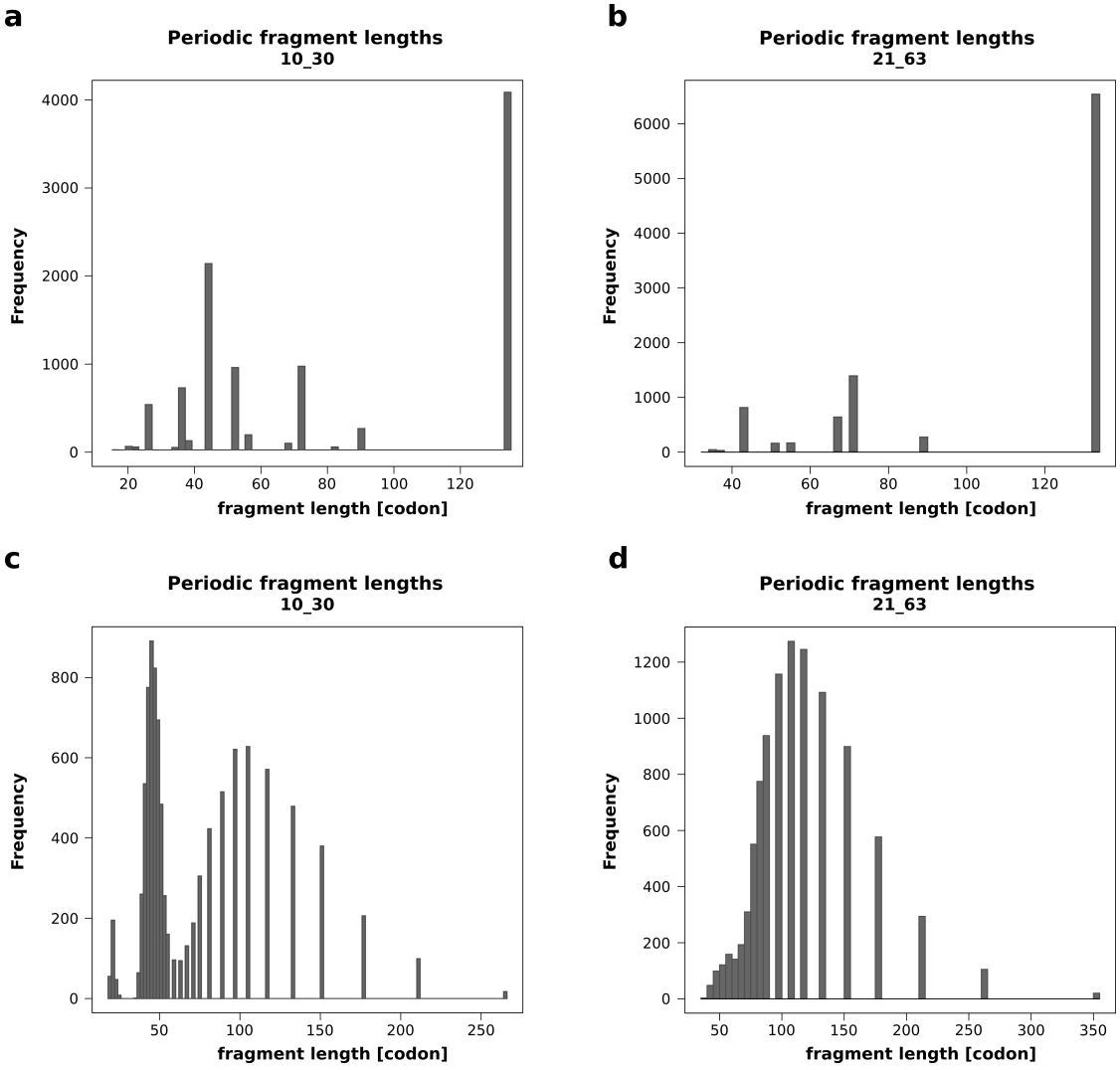


Figure 4.4: Periodic fragments: original coat protein length

a-b) Shuffling of the residues in the original coat protein, c-d) shuffling of the genome. The frequency gives the number of obtained fragments from the generated 10^4 shuffled residue-sequences/genomes. The windows that were used were 10_30 for figures a, c and 21_63 for the figures b, d.

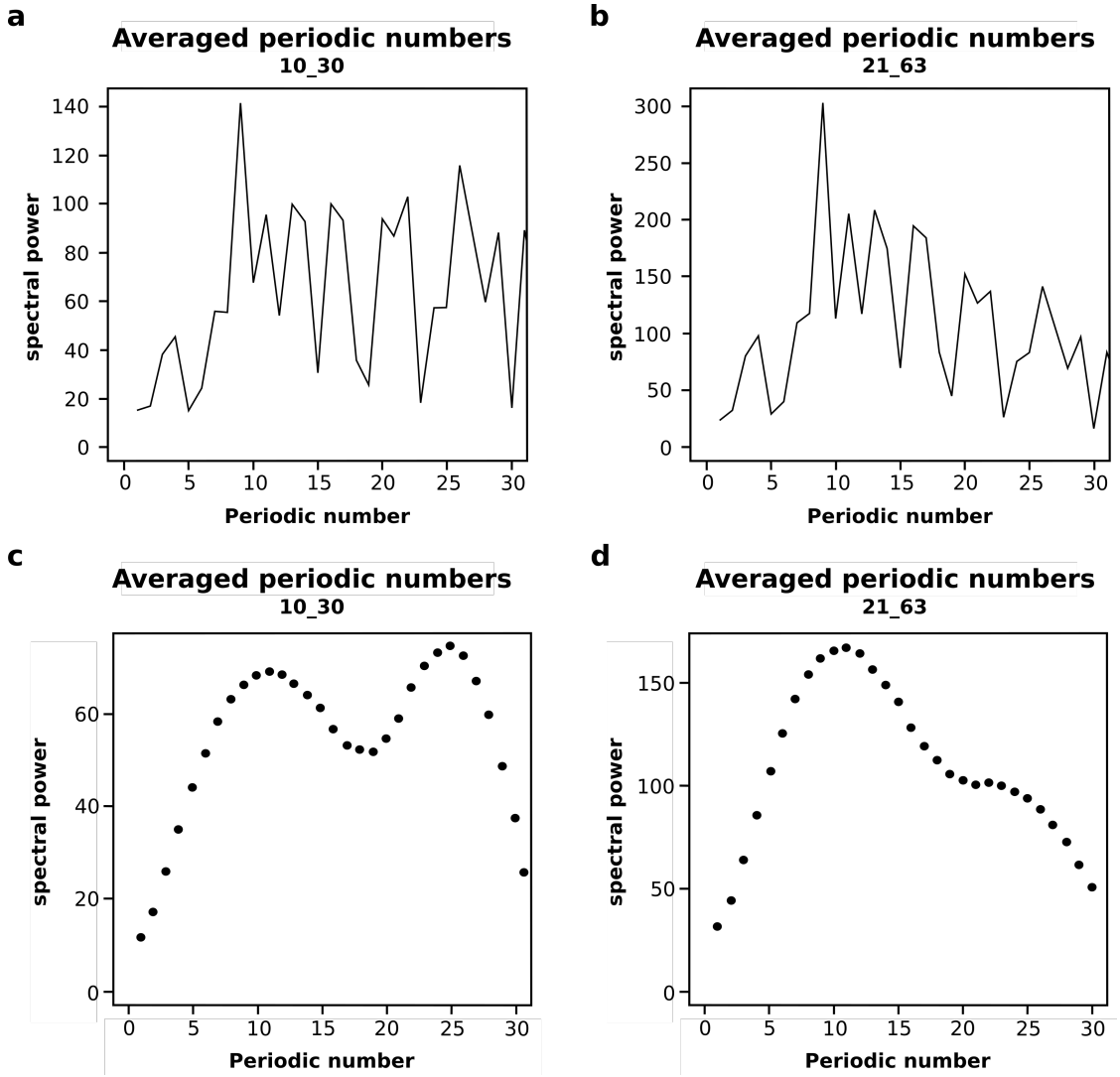


Figure 4.5: Periodic numbers: original coat protein length

a-b) Periodic numbers from the coat protein residue shuffling; c-d) periodic numbers from the shuffling of the genome. The values for the spectral power are averaged over all 10^4 shuffled spectra. The windows that were used in the process of calculation were 10_30 for figures a, c and 21_63 for the figures b, d.

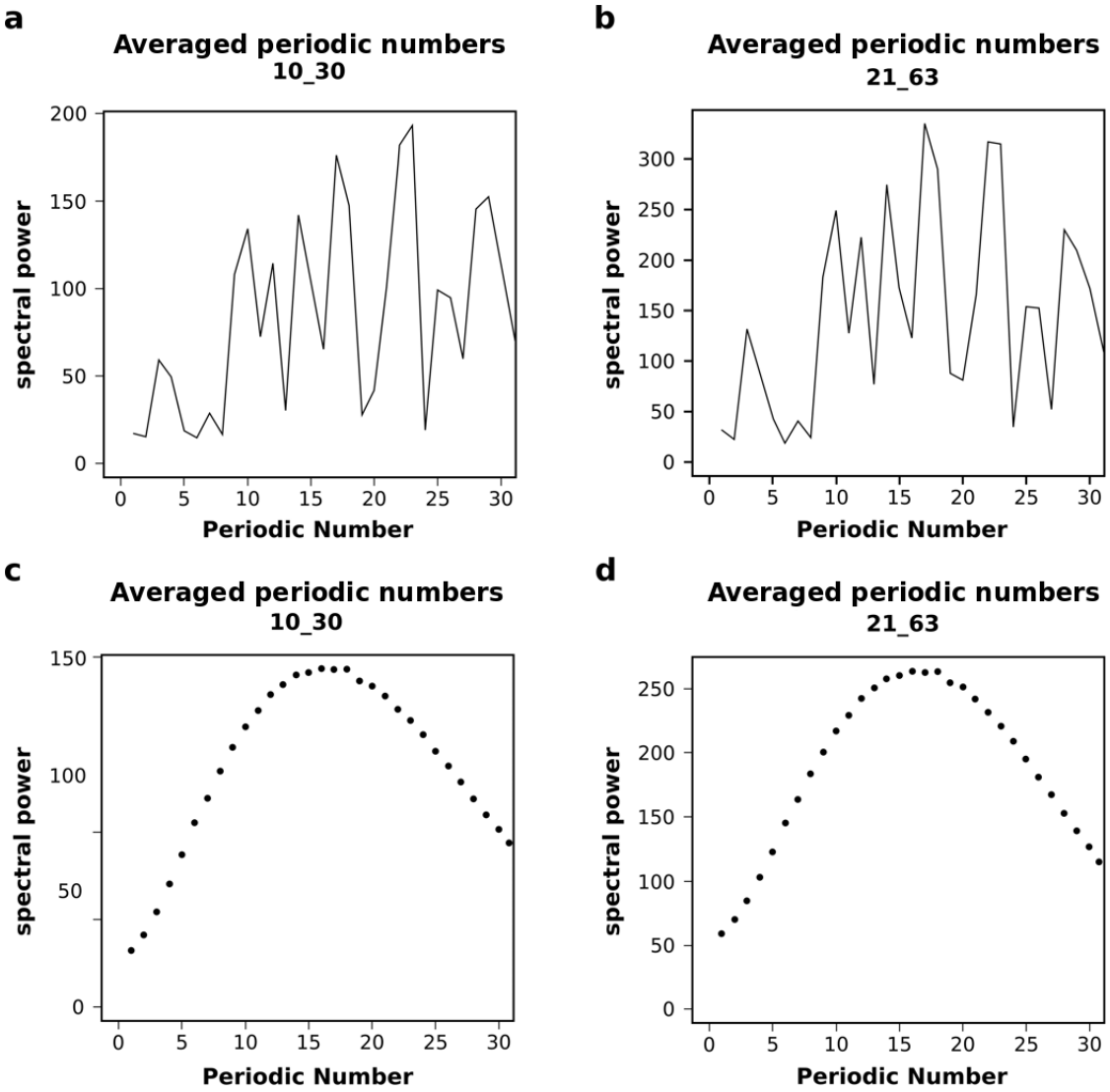


Figure 4.6: Periodic numbers: short coat protein length

a-b) Periodic numbers from the shortened coat protein residue-sequence shuffling; c-d) periodic numbers from the shuffling of the genome. The values for the spectral power are averaged over all 10^4 generated shuffled spectra. The windows that were used were 10_30 for figures a, c and 21_63 for the figures b, d.

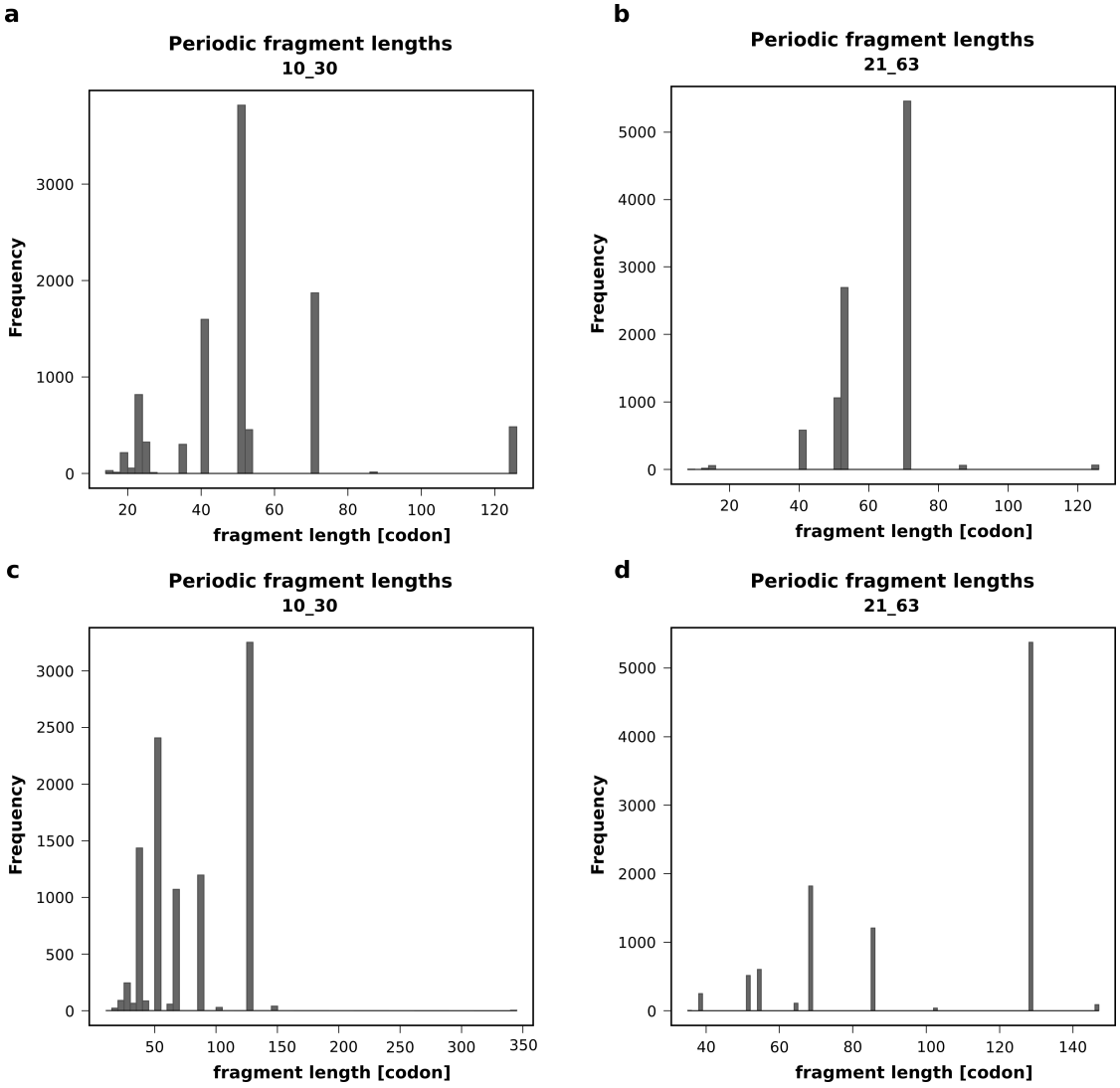


Figure 4.7: Periodic fragments: short coat protein length

a-b) Shuffling of the residues in the shortened coat protein; c-d) shuffling of the genome. The frequency gives the number of obtained fragments from the generated 10^4 shuffled residue-sequences/genomes.The windows that were used were 10_30 for figures a, c and 21_63 for the figures b, d.

Chapter 5

Discussion

5.1 Implications for the generalized complementarity hypothesis

Inter-molecular complementary interactions at the level of primary sequences of biological macromolecules have traditionally been analyzed primarily in the context of the interactions involving single DNA or RNA strands. Such complementary interactions play a central mechanistic role in a large number of different cellular processes and include interactions between DNA or RNA strands themselves or both DNA/RNA strands with antisense RNAs, miRNAs, siRNAs, nc-RNAs, riboswitches, tRNA/rRNAs and other nucleic acids [64, 65, 66, 67, 68, 69]. In many of these cases, the strand complementarity is the key principle behind interaction specificity. On the other hand, a putative complementaritiy between RNAs and proteins remains, as proposed recently [7, 21, 22], remains to be explored in more detail.

To study the limits of validity of the RNA/protein complementarity hypothesis in a concrete biological system, we have selected for our analysis the Enterobacteria phage MS2, as it is one of best studied positive-sense RNA viruses known. Importantly, its minimal, ssRNA genome makes it an amenable test case. As shown in this study, the mRNA/protein complementarity hypothesis and its generalization to non-cognate sequences provides a productive framework for thinking about the codon/amino acid relationship in the course of evolution, but also provides a potentially novel principle for analyzing the interaction between RNA and proteins in highly evolved, present-day systems. Importantly, the application of the generalized complementarity principle to the MS2 genome and its coat protein led to the identification of 10 out of 13 possible detectable binding sites known from experiment Fig.3.3. Even when comparing

our predicted data against other experimentally obtained data [17], we were able to capture the main features of the experimentally determined binding sites. Considering that each of the four RNA nucleobase profiles points at somewhat different binding sites, however, suggests that the simplistic picture in which individual profiles are treated separately in the course of binding predictions may be incomplete. Moreover, we are currently unable to furnish a satisfactory explanation as to why the anti-matching areas in the adenine profiles, which would indicate the unfavorable interactions normally, are able to pick up experimentally proven binding interactions between the coat protein and the virus genome. Although not every peak in the Pearson R profiles points to a binding site, it is known that the binding of the coat protein to the genome and the subsequent assembly and packing of the virion does not take place at one of the 15 identified stemloops only. There rather exist multiple stemloops, scattered throughout the genome, that are able to bind the coat protein dimer, with the only difference being that the binding is not as strong as in those conserved 15 stemloops [23]. This is consistent with the fact that the Pearson correlations coefficients R corresponding to those peaks are smaller in magnitude than those with strong interactions. This is, however, only one of the potential explanations and further study in this direction is needed. Overall, there exist multiple indications that cognate interactions play a central role in MS2 encapsidation. For example, the *in vivo* studies, performed with single molecule fluorescence, of the packing mechanism of two single stranded viruses, one being MS2, revealed that the hydrodynamic collapse, in the course of capsidation, is specific to viral RNAs making cognate interactions with their respective coat protein. Furthermore, only cognate interactions, in this study, yielded capsids of the correct size and symmetry ($T = 3$ for MS2), whereas non-cognate assembly reactions proved to be of a relatively low yield and produce a high proportion of misassembled species [70]. All of the results presented here were obtained by linear alignment of a single coat protein sequence against the viral genome, and one needs to consider that such an event is highly unlikely in a natural environment as secondary and tertiary structures of both the genome and the coat protein would have a strong influence on the actual binding. After all, the experimental structure of the virus reveals that the genomic RNA contacts the capsid proteins in a highly structured context Fig.2.2. One potential explanation of the successes of our simplistic, primary-sequence-based analysis might be that the co-aligned cognate binding in the unstructured state actually enables the proper folding of the involved partners and the subsequent encapsidation. In other words, it is possible that the multivalent, dynamic interactions in the unstructured state, which may be

adequately captured in a primary-sequence analysis, modulate the folding landscape of both the RNA and the proteins, while at the same time guiding the process of capsid assembly. This and the fact that the coat protein is thought to bind as a dimer and not as a single protein as it was treated here could explain why there are deviations between the profiles and the actual known binding sites. Specifically, this may be the reason why it is difficult to detect the conserved points of contact between nucleobases and amino acids known to be essential for binding such as SER47.
On the other hand, it should be pointed out that coat proteins of bacteriophages such as MS2 have a unique fold among non-enveloped spherical (+)-stranded RNA viruses. The MS2 coat protein is entirely ordered and forms a rigid dimer, whereas a significant portion of the coat proteins of both nonenveloped and enveloped spherical (+)-stranded RNA viruses are structurally flexible and posses intrinsically disordered sequences [71, 72, 73, 74]. The flexible regions, mostly N-terminal arms, tend to be located within the internal cavity of the capsid and mediate interaction with RNA [75, 76, 77, 78, 79, 71]. As the complementarity hypothesis is able to pick up interactions even for a rigid protein, such as the MS2 coat protein and the viral MS2 genome, an extension of the work to other (+)-sense RNA viruses is worth considering. The motivation behind this would be that cognate interactions may primarily occur at the level of interactions between unstructured regions of RNAs and proteins [22]. Despite the above shortcomings, the complementarity hypothesis and the used scales appear to provide a potentially promising tool for investigating binding sites or more general RNA/protein interactions *a priori*, even without the knowledge of secondary and tertiary structures of the partners involved.

As our principal result, we have shown that in the regions where the MS2 coat protein binds directly to the MS2 RNA, one also observes strong complementarity between RNA nucleobase density profiles and the coat protein nucleobase affinity profiles. Importantly, strong profile complementarity is seen not only in the region of the MS2 coat protein gene, but also in multiple other locations, including the genes of other MS2 proteins. In other words, it appears as if the MS2 coat protein gene has some features in common with the genes of other MS2 proteins, since one and the same protein (the MS2 coat proteins) exhibits strong complementary matching with all of them in different regions. First, these results support the possibility that the viral genome may have evolved through a duplication of the coat protein CDS, followed by a subsequent evolution of other MS2 proteins from the coat protein CDS. In other words, the nucleobase affinity profiles of different MS2 proteins tend to be similar due

to gene duplication of the coat protein CDS. Of course, this possibility cannot at this stage be differentiated from the reverse possibility that the coat protein CDS actually derived from the CDS of some other MS2 protein. Alternatively, the MS2 genome i.e. the coding regions of the four MS2 proteins evolved in a convergent way such that they exhibit similar nucleobase/amino acid affinities for the coat protein because of a common evolutionary pressure. Specifically, the viral genome needs to be highly compressed in the enclosed space of the virion. The symmetric organization of the coat protein [40] in the context of the viral capsid and the interactions of the viral genome with the inside of the capsid could provide the evolutionary pressure that could lead to higher specificity for the interactions between the viral genome and the coat protein. Future research should shed more light on these possibilities.

5.2 Periodicity

Periodicity is a common feature in biological systems at different levels, ranging from the oscillations of transcription factors as a regulatory element in gene expression [80], to the circadian changes of behavior and physiology, driven by molecular clocks endogenous to most organisms [81], to the simple periodicity in DNA sequences, which is caused by the triplet nature of the genetic code [82]. It is thus of an ever-growing interest to identify periodic patterns in biological systems and the implications they could have as well as identify new ones. On the other hand, not all patterns carry a biological meaning and frequently apparent periodicities are introduced by the way we look at a problem or probe for it. We have hypothesized that there may exist a specific periodicity in the MS2 genome that stems from the specific interactions of the MS2 RNA and the coat protein. Moreover, we have hypothesized that the periodic fragment length of 132.5 amino-acids, which is close to the number of amino acids in the coat protein, could hint at a unique relationship between the coat protein and the genome. However, our analysis could not show sufficient evidence in support of this hypothesis. The first indication for this is, that the order of the amino acids in the sliding window does not have any major effect on the periodic fragment lengths, which would have to be true if the hypothesis were to be correct. Second, there should also be a significant difference based on what type of amino acids are added or subtracted from the coat protein. However, if both of the 160 amino acid extensions, which only differ in the amino acids that were added, are compared against each other with regards to their periodic fragment lengths, no indication of this was found. The shuffling of the genome has a more profound effect on the respective fragment lengths, at least for

the fragment with the highest spectral power density. However, a closer look at the averaged powers of the individual periodic numbers coming from different spectra and their Gaussian distributions, the fact that for the extended, yet not shuffled variants the periodic numbers resemble the periodic numbers obtained by shuffling, i.e. the similar distributions, and the relative closeness of the absolute value of the highest peak (extended, yet not shuffled) as compared to the most abundant peak of the averaged spectra suggests that, for this specific nucleobase composition of the genome, the nucleobase order does not make a difference and the typical fragment length can be expected regardless of sequence. As neither the order of amino acids in the sliding window nor the order of the genome are relevant for the periodic fragment lengths obtained by Fourier analysis, the only factor that could contribute to it is the length of the coat protein that was used as the sliding window. This suggests that the bias one introduces by applying a particular sliding window length has a much stronger effect and the Fourier transform filters out exactly this perturbation. Independently of this, a second relevant factor is window-averaging. Window-averaging is often used in bioinformatics and the consequences of it are more often than not neglected. The impact for window-averaging is less for the larger window-sizes, but if the windows are small, as is often the case, the impact on the results can be significant and should be considered. The scope of our present analysis is not sufficient to provide a comprehensive answer as to how exactly window averaging biases the obtained results and what the right way to deal with it would be, but they do point at the fact that window-averaging does make a difference and should be considered carefully. We see this as an important direction to follow in the future.

Chapter 6

Zusammenfassung

Trotz der enormen Bedeutung von RNA-Protein Interaktionen bleibt unser Verständnis dieser jedoch unvollständig. Im Detail betrifft dies vor allem Interaktionen von Protein codierenden mRNA Sequenzen. Diesbezüglich jedoch haben wissenschaftliche Arbeiten, auf diesem Gebiet, die Komplementäre Verwandtschaft von mRNA codierenden Sequenzen zu Nukleobasen Affinitätsprofilen von deren codierenden Protein Sequenzen nachgewiesen. Dieses wurde als Grundlage genommen für die Vermutung, dass mRNAs und die Proteine für die diese codieren in einer direkten, gerichteten komplementären Weise miteinander interagieren, besonders wenn die Partner unstrukturiert sind. In dieser Arbeit untersuchten wir die Hypothese der RNA/Protein Komplementarität in Bezug auf ihre Gültigkeit, im Rahmen eines konkreten biologisch relevanten Systems, dem Enterobacteria Virus MS2. Von diesem Virus MS2 ist bekannt, dass sein eigenes Coat-Protein in mehreren Stellen an die eigene genomische RNA bindet und diese Eigenschaft könnte eine wesentliche Verbindung mit der vorher erwähnten Hypothese aufweisen. Die erste Frage, die wir uns stellten war, ob es möglich ist Interaktionen vom MS2 Coat-Protein und dem Genome des Virus selber festzustellen, gestützt nur auf die Analyse der Primären Sequenzen mit Hilfe der Komplementaritaetshypothese. Die zweite Frage war, ob man feststellen könnte, mithilfe der Fourier Transform, ob eine definitive Periodizität zwischen den Interaktionsmustern von der viralen RNA und dessen Coat-Protein vorhanden ist. Unter der Benützung von bereits wissenschaftlich bekannten Nukleobasen/Aminosäuren Affinitäten waren wir in der Lage 10 der 13 potentiell detektierbaren, experimentell bestätigten Bindungsstellen vorherzusagen. Die Komplementaritaetshypothese scheint daher eine sehr vielversprechende Methode zu bieten für das Erforschen und Vorhersagen von RNA/Protein Interaktionen. Da jedoch die genaue Beziehung zwischen den individuellen RNA Basen und deren Profilen noch nicht genau geklärt ist, ist es bisher noch nicht möglich gewesen ein robustes und

universelles System zu bieten, welches die RNA-Protein Interaktionen im generellen beschreibt und weitere Forschung in diesem Bereich ist von Nöten. Des weiteren zeigte unse Analyse keine dedektierbare Periodizität in den Interaktionsmustern zwischen der MS2 RNA und dem Coat-Protein, welche über zufällige Kontrollen hinausgehen würde.

Bibliography

[1] Prof. Brad Osgood. *The Fourier Transform and its Applications*. Electrical Engineering Department Stanford University. 2007.

[2] Bernhardt HS. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biology Direct.*, 23(7), 2012.

[3] Woese CR. Translation: in retrospect and prospect. *RNA*, 7(8):1055–1067, 2001.

[4] CR Woese, DH Dugre, WC Saxinger, and Dugre SA. The molecular basis for the genetic code. *Proc Natl Acad Sci USA.*, 55(4):966–974, 1966.

[5] Koonin E.V. and Novozhilov A.S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life.*, 61:99–111, 2009.

[6] Anton A. Polyansky and Bojan Zagrovic. Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res.*, 41(18):8434–8443, 2013.

[7] Zagrovic Bojan, Bartonek Lukas, and Polyansky Anton A. RNA-protein interactions in an unstructured context. *FEBS Letters*, 2018.

[8] Kazan H, Ray D, Chan ET, Hughes TR, and Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*, 6, 2010.

[9] Orenstein Y, Wang Y, and Berger B. RCK: accurate and efficient inference of sequence- and structure-based proteinRNA binding models from RNAcompete data. *Bioinformatics*, 32:i351–i359, 2016.

[10] Alipanahi B, Delong A, Weirauch MT, and Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33:931–838, 2015.

[11] Peng Z, Wang C, Uversky VN, and Kurgan L. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. *Methods Mol Biol.*, 1484:187–203, 2017.

[12] Agostini F, Zanzoni A, Klus P, Marchese D, Cirillo D, and Tartaglia GG. catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, 29:2928–2930, 2013.

[13] Noller HF. Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol.*, 4:1–U20, 2012.

[14] Cheng AC, Chen WW, Fuhrmann CN, and Frankel AD. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J Mol Biol*, 327:781–796, 2003.

[15] Ebrahimi A, HabibiKhorassani M, Gholipour AR, and Masoodi HR. Interaction between uracil nucleobase and phenylalanine amino acid: the role of sodium cation in stacking. *Theor Chem Acc*, 124:115–122, 2009.

[16] Peterson TL Rutledge LR, NavarroWhyte L and Wetmore SD. Effects of extending the computational model on DNA-protein T-shaped interactions: the case of adenine-histidine dimers. *Phys Chem A*, 115:12646–12658, 2011.

[17] Polyansky Anton A. and Zagrovic Bojan. Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res.*, 41(18):84348443, 2013.

[18] Woese CR. Evolution of the genetic code. *Naturwissenschaften*, 60:447–459, 1973.

[19] Koonin EV and Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, 61:99–111, 2009.

[20] Mathew DC and LutheySchulten Z. On the physical basis of the amino acid polar requirement. *J Mol Evol*, 66:519–528, 2008.

[21] Polyansky A.A., Hlevnjak M., and Zagrovic B. Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biol.*, 10:1248–1254, 2013.

[22] Beier A, Zagrovic B, and Polyansky AA. On the contribution of protein spatial organization to the physicochemical interconnection between proteins and their cognate mRNAs. *Life (Basel)*, 4:788799, 2014.

[23] Dai Xinghong, Li Zhihai, Lai Mason, Shu Sara, Du Yushen, Z. Hong Zhou, and Sun Ren. In situ structures of the genome and genome-delivery apparatus in an ssRNA virus. *Nature*, 514(7635):112–116, 2017.

[24] W. Fiers, Duerinck F. Contreras, R., G. Haegeman, D. Iserentant, J. Merregaert, W. M. Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, 1976.

[25] J. van Duin and N. Tsareva. *In Calendar, R. L. The Bacteriophages (Second ed.)*, chapter 15, pages 175–196. Oxford University Press, 2006.

[26] C. Z. Ni, C. A. White, R. S. Mitchell, J. Wickersham, R. Kodandapani, D. S. Peabody, and K. R. Ely. Crystal structure of the coat protein from the GA bacteriophage: model of the unassembled dimer. *Protein Sci.*, 5(12):2485–2493, 1996.

[27] Jiang Wen, Chang Juan, Jakana Joanita, Weigele Peter, King Jonathan, and Chiu Wah. Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature*, 312:612–616, 2006.

[28] GC Lander, L Tang, SR Casjens, EB Gilcrease, P Prevelige, A Poliakov, CS Potter, B Carragher, and JE Johnson. The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science*, 312:1791–1795, 2006.

[29] Carlos Enrique Catalano. *Viral Genome Packaging Machines: Genetics, Structure, and Mechanism*. Springer, 2005.

[30] G Basnak, VL Morton, O Rolfsson, NJ Stonehouse, AE Ashcroft, and PG Stockley. Viral genomic single-stranded RNA directs the pathway toward a T=3 capsid. *J Mol Biol*, 395:924–936, 2010.

[31] S Sun, VB Rao, and MG Rossmann. Genome packaging in viruses. *Curr Opin Struct Biol.*, 20:114–1120, 2010.

[32] JD Perlmutter and MF Hagan. Mechanisms of virus assembly. *Annu Rev Phys Chem.*, 66:217–239, 2015.

[33] A Klug. The tobacco mosaic virus particle: structure and assembly. *Philos Trans R Soc Lond B Biol Sci.*, 354:531–535, 1999.

[34] K Valegard, JB Murray, PG Stockley, NJ Stonehouse, and L Liljas. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature*, 371:623–626, 1994.

[35] GW Witherell, JM Gott, and OC Uhlenbeck. Specific interaction between RNA phage coat proteins and rna. *Prog Nucleic Acid Res Mol Biol.*, 40:185–220, 1991.

[36] J Rumnieks and K. Tars. Crystal Structure of the Maturation Protein from Bacteriophage Q. *J. Mol. Biol.*, 429:688–696, 2017.

[37] DA Kuzmanovic, I Elashvili, C Wick, C O'Connell, and S. Krueger. The MS2 coat protein shell is likely assembled under tension: a novel role for the MS2 bacteriophage A protein as revealed by small-angle neutron scattering. *J. Mol. Biol.*, 355:1095–1111, 2006.

[38] H Groeneveld, K Thimon, and J. van Duin. Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *RNA*, 1:79–88, 1995.

[39] Romaniuk Paul J., Lowary Peggy, Nan Wu Huey, Stormo Gary, and C. Uhlenbeck Olke. RNA binding site of R17 coat protein. *Biochemistry*, 26:1563–1568, 1987.

[40] Plevka Pavel, Tars Kaspars, and Liljas Lars. Structure and stability of icosahedral particles of a covalent coat protein dimer of bacteriophage MS2. *Protein Sci.*, 18:1653–1661, 2009.

[41] CZ Ni, R Syed, R Kodandapani, J Wickersham, DS Peabody, and KR. Ely. Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. *Structure*, 3:255–263, 1995.

[42] Atkins JF, Steitz JA, Anderson CW, and Model P. Binding of mammalian ribosomes to MS2 phage RNA reveals an overlapping gene encoding a lysis function. *Cell*, 18(2):247–256, 1979.

[43] MN Beremand and T. Blumenthal. Overlapping genes in RNA phage: a new protein implicated in lysis. *Cell*, 18(2):257–266, 1979.

[44] Chamakura Karthik R., S. Tran Jennifer, and Ry Young. MS2 lysis of *Escherichia coli* Depends on Host Chaperone DnaJ. *J Bacteriol*, 199(12), 2017.

[45] B. Walderich, A. Ursinus-Wssner, J. van Duin, and JV. Hoeltje. Induction of the autolytic system of Escherichia coli by specific insertion of bacteriophage MS2 lysis protein into the bacterial cell envelope. *J Bacteriol*, 170(11):5027–5033, 1988.

[46] Berkhout B, de Smit M H, Spanjaard R A, Blom T, and J van Duin. The amino terminal half of the MS2-coded lysis protein is dispensable for function: implications for our understanding of coding region overlaps. *EMBO J.*, 4(12):33153320, 1985.

[47] BF Schmidt, B Berkhout, GP Overbeek, A van Strien, and J. van Duin. Determination of the RNA secondary structure that regulates lysis gene expression in bacteriophage MS2. *J. Mol. Biol.*, 195(3):505–516, 1987.

[48] B Berkhout, BF Schmidt, A van Strien, J van Boom, J van Westrenen, and J. van Duin. Lysis gene of bacteriophage MS2 is activated by translation termination at the overlapping coat gene. *J. Mol. Biol.*, 195(3):517–524, 1987.

[49] T Sugiyama and D. Nakada. Control of translation of MS2 RNA cistrons by MS2 coat protein. *Proc. Natl. Acad. Sci. USA*, 57:1744–1750, 1976.

[50] H Robertson, RE Webster, and ND. Zinder. Bacteriophage coat protein as repressor. *Nature*, 218:533–536, 1968.

[51] K Eggen and D. Nathans. Regulation of protein synthesis directed by coliphage MS2 RNA. ii. in vitro repression by phage coat protein. *J. Mol. Biol.*, 39:293–305, 1969.

[52] HF Lodish and Zinder ND. Mutants of the bacteriophage f2. 8. Control mechanisms for phage-specific syntheses. *J. Mol. Biol.*, 19:333–348, 1966.

[53] B Berkhout and J. van Duin. Mechanism of translational coupling between coat protein and replicase genes of RNA bacteriophage MS2. *Nucleic Acids Res.*, 13:6955–6967, 1985.

[54] RA Kastelein, E Remaut, W Fiers, and J. van Duin. Lysis gene expression of RNA phage MS2 depends on a frameshift during translation of the overlapping coat protein gene. *Nature*, 295(5844):35–41, 1982.

[55] K Valegrd, L Liljas, K Fridborg, and T Unge. The three-dimensional structure of the bacterial virus MS2. *Nature*, 345:36–41, 1990.

[56] van den Worm S.H., Stonehouse N.J., Valegrd K., Murray J.B., Walton C., Fridborg K., Stockley P.G., and Liljas L. Crystal structures of MS2 coat protein mutants in complex with wild-type RNA operator fragments. *Nucleic Acids Res.*, 26(5):1345–1351, 1998.

[57] AS Rose, AR Bradley, Y Valasatava, JM Duarte, A Prli, and PW Rose. Web-based molecular graphics for large complexes. *ACM Proceedings of the 21st International Conference on Web3D Technology (Web3D '16)*, pages 185–186, 2016.

[58] AS Rose and PW Hildebrand. NGL Viewer: a web application for molecular visualization. *Nucl Acids Res.*, 43(W1):W576–W579, 2015.

[59] Woese C.R. The genetic code: the molecular basis for genetic expression. *New York.: Harper & Row*, 1967.

[60] Yarus M. Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J Mol Evol.*, 47(1):109–117, 1998.

[61] Yarus M., Widmann J.J., and Knight R. RNA-amino acid binding: a stereochemical era for the genetic code. *J. Mol. Evol.*, 69(5):406429, 2009.

[62] Hlevnjak M, Polyansky A.A., and Zagrovic B. Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. *Nucleic Acids Res.*, 40(18):8874–8882, 2012.

[63] Rolfsson , Middleton S, Manfield IW, and et al. Direct Evidence for Packaging Signal-Mediated Assembly of Bacteriophage MS2. *J Mol Biol.*, 428(2):431–448, 2016.

[64] Katayama et al. Antisense transcription in the mammalian transcriptome. *Science*, 309(5740):1564–1566, 2005.

[65] He Y, Vogelstein B, Velculescu VE, Papadopoulos N, and Kinzler KW. The antisense transcriptomes of human cells. *Science*, 322(5909):1855–7, 2008.

[66] Faghihi MA et al. Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol*, 11(5), 2010.

[67] Nobuyoshi Kosaka et al. Trash or Treasure: extracellular microRNAs and cell-to-cell communication. *Front Genet.*, 4(173), 2013.

[68] Khade PK, Shi X, and Joseph S. Steric complementarity in the decoding center is important for tRNA selection by the ribosome. *J Mol Biol*, 425(20):3778–3789, 2013.

[69] Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, and Breaker RR. Genetic control by a metabolite binding mRNA. *Chem Biol.*, 9(9), 2002.

[70] Borodavka A, Tuma R, and Stockley PG. A two-stage mechanism of viral RNA compaction revealed by single molecule fluorescence. *RNA Biol.*, 10(4):481–9, 2013.

[71] Peng Ni and C. Cheng Kao. Non-encapsidation Activities of the Capsid Proteins of Positive-strand RNA viruses. *Virology.*, 446(0), 2013.

[72] Valegrd K, Liljas L, Fridborg K, and Unge T. The three-dimensional structure of the bacterial virus MS2. *Nature.*, 345(6270):36–41, 1990.

[73] Roland Ivanyi-Nagy, Jean-Pierre Lavergne, Caroline Gabus, Damien Ficheux, and Jean-Luc Darlix. RNA chaperoning and intrinsic disorder in the core proteins of Flaviviridae. *Nucleic Acids Res.*, 36(3):712–725, 2008.

[74] Lars Liljas. *Functional Role of Structural Disorder in Capsid Proteins.* Wiley Online Library, 2011.

[75] Fisher A and Johnson J. Ordered duplex RNA controls capsid archtecture in an icosahedral animal virus. *Nature.*, 361:176179, 1993.

[76] Harrison SC, Olson AJ, Schutt CE, and Winkler FK. Tomato bushy stunt virus at 2.9 resolution. *Nature.*, 276:368375, 1978.

[77] Lucas RW, Larson SB, and McPherson A. The crystallographic structure of brome mosaic virus. *J Mol Biol.*, 317:95108, 2002.

[78] Boulant S, Vanbelle C, Ebel C, Penin F, and Lavergne JP. Hepatitis C Virus Core Protein Is a Dimeric Alpha-Helical Protein Exhibiting Membrane Protein Features. *J Virol.*, 79:1135311365, 2005.

[79] Choi H, Tong L, Minor W, Dumas P, Boege U, Rossmann MG, and Gerd W. Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion. *Nature.*, 354:3743, 1991.

[80] CKeng Boon Wee, Wee Kheng Yio, Uttam Surana, and Keng Hwee Chiam. Transcription Factor Oscillations Induce Differential Gene Expressions. *Biophysical Journal*, 102(11):2413–2423, 2012.

[81] Coon SL et al. Circadian changes in long noncoding RNAs in the pineal gland. *Proc Natl Acad Sci USA.*, 109(33):13319–24, 2012.

[82] Stephen T Eskesen, Frank N Eskesen, Brian Kinghorn, and Anatoly Ruvinsky. Periodicity of DNA in exons. *BMC Mol Biol.*, 5(12), 2004.