# 1 Annotation

Bacteriophage (phages) - are viruses, that selectively attack bacterial cells. This specificity is due to the dependence of the phage on the cellular mechanism of the host, which the phage needs for replication and synthesis of viral components. An example of such a dependence can be observed in T4 phage and E. coli[1] or cyanobacteria with its phage [2].

In this work, we hypothesized that this kind of relationship between phage and host bacterium can be observed at the genetic level. Since in many known cases attacks by bacterial phages begin with the capture of the bacterial RNA polymerase, we looked for common motifs in the promoter regions of phages and bacteria.

As a result, we found that in almost half of the cases, phages and hosts have common motifs in these areas. Also, in most cases, we found common motifs between different phages attacking the same bacteria. Thus, when looking for interactions between a phage and a bacterium, it is a good start to look for common motifs in the promoter regions of the phage and the bacterium.

# Contents

# 2 Introduction

## 2.1 Bacteriophages

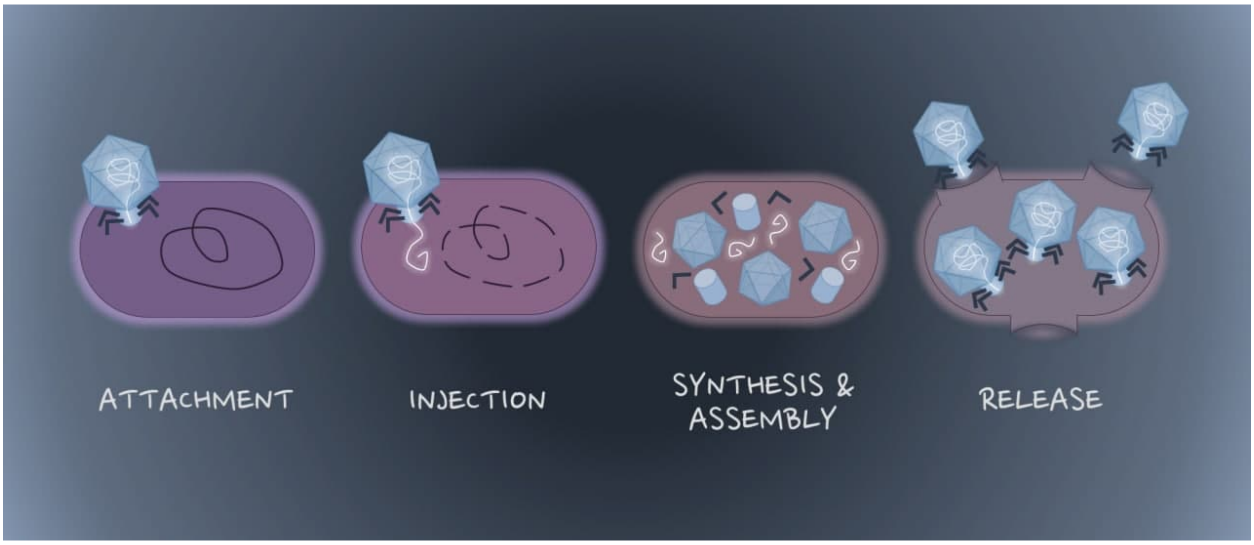

Figure 1: *Life cycle of an average phage.* (1) The phage first lands on the bacterium. (2) Then it injects its DNA into bacteria. (3) The DNA is copied and used to make capsids for new phages. (4) The new phages gather and explode the bacterium, killing it in the process. [16]

Bacteriophages are viruses that infect and replicate inside bacteria and archaea. Bacteriophages usually consist of a protein shell and genetic material. They can have single-stranded or double-stranded nucleic acid (DNA or, less commonly, RNA). Phages reproduce inside the bacterium after inducing their genome into its cytoplasm. [13]

More than 90% of bacteriophages have large genomes in the form of double-stranded DNA. Phages belong to the three main families of bacteriophages: Myoviridae (with long, stiff, contractile tails), Siphoviridae (with long, flexible tails) and Podoviridae (with short tails). [14]

For a bacterium, the outcome of infection with a bacteriophage can be different. Some bacteriophages cause lysis, in which case the cell dies in a very short amount of time. As a rule, this leads to the birth of hundreds of new viruses within a few minutes or hours. This process can be repeated as long as bacteria are present in sufficient numbers to support replication. Many bacteriophages behave this way (they are called virulent) and are not capable of causing any other kind of infection. There are also temperate bacteriophages that infect cells and then become dormant in a latent state: replicate along with the host chromosome and are subsequently passed on to each daughter cell after division. However, these dormant phages can be activated by several causes, such as DNA damage. For some bacteriophages, host' chromosomal DNA can be packaged into bacteriophage particles during bacteriophage' replication instead of the bacteriophage' genome. This can lead to high levels of horizontal gene transfer in the bacterial population. [15]

### 2.1.1 Use of bacteriophages

The use of bacteriophages as specific antimicrobial agents is widely documented in the literature. They can even be used as a controlled therapeutic agent.

As already mentioned, in most cases, bacteriophages prevent bacteria from multiplying. Instead, new phages are born. At a time when, for whatever reason, chemical antibiotics may not work, phages can be a good replacement for them [14]. Benefits include reduced side effects and reduced risk of developing bacterial resistance. Disadvantages include the difficulty of finding an effective phage for a particular infection. [16]

Phages are not used only in medicine. They can be used in the food industry, they are used as a counter to biological weapons and toxins, in diagnostics (for example, to search for staphylococcus in the blood), in the dairy industry. Bacteriophages are also widely used in various biological research. [13]

### 2.1.2 Motifs

A DNA, RNA, or protein sequence motif is a short pattern that persists through evolution. A motif in DNA can be a protein binding site; in proteins, a motif can be the active site of an enzyme or a structural unit required for proper protein folding. Motifs in sequences are one of the basic functional units of molecular evolution.

Many phage genomes encode small proteins that specifically alter the bacterial host's RNA polymerase (RNAP), inhibiting bacterial DNA transcription and promoting regulated transcription of phage DNA.

Transcription of E. coli begins with the binding of the $\sigma$-factor to the catalytic "core" of RNAP. The initial $\sigma$-factor, $\sigma^{70}$, binds specifically to individual nucleotide sequences (motifs) in the promoter. Phages have evolved mechanisms that modify $\sigma^{70}$ and redirect bacterial RNAP to phage' gene transcriptions.

# 3 Goal and tasks

Phages have been shown to have a wide variety of applications. Understanding the structure of phages and how they work can simplify and/or improve the way they are used. In this work, we have tried to improve the understanding of the interaction between bacteriophages and bacteria. We wanted to find common motifs between the phages and the bacteria they attack in the promoter regions of their DNA sequences.

We were also interested in studying the presence of common motifs in promoter sequences between phages. We wanted to check whether the presence of common motifs in the promoter regions of phages is evidence of their kinship. Also, phages with common motifs can have similar bacterial hosts.

# 4 Literature review

## 4.1 Interaction between T4 phage and E.coli

Expression of the T4 genome is a highly regulated process that begins immediately after infection of the host. The main control over this expression occurs at the level of transcription. T4 does not code for its own RNA polymerase (RNAP), but instead codes for a variety of factors that serve to alter host polymerase specificity as infection progresses. Changes in the host polymerase are accompanied by three successive stages of phage transcription: early, intermediate, and late. Early and intermediate RNA (RNA formed during early and intermediate transcription) is detected prereplicatively [26, 27], while late transcription is accompanied by T4 replication. Early T4 transcripts are generated from early promoters (Pe) that are active immediately after infection. Intermediate T4 transcripts are generated approximately 1 minute after infection at $37°C$ and require phage protein synthesis. Intermediate RNA is synthesized in two ways: 1) activation of intermediate promoters (Pm) and 2) expansion of Pe transcripts from early genes to downstream middle genes.

### 4.1.1 E. coli transcription start

E. coli RNA holoenzyme, like all bacterial RNA, consists of a core of subunits ($\beta, \beta', \alpha_1, \alpha_2$ and $\omega$). It contains an active site for RNA synthesis, and a specificity factor $\sigma$, which recognizes promoters in DNA and is set to the starting site for transcription. Primary $\sigma$, $\sigma^{70}$ in E. coli, used during exponential growth; alternative factors $\sigma$ direct the transcription of genes required in different growth conditions or during times of stress. [17]

To start transcription, RNAP parts must first recognize and bind to double-stranded (ds) DNA recognition elements present in the promoter [18]. Each of the C-terminal domains of $\alpha$-subunits ($\alpha$-CTDS) can interact with upstream elements, A/T rich sequences present between positions -40 and -60. The $\sigma^{70}$ parts present in RNAP can interact with three different dsDNA elements: element -35, sequence -15TGn-13 (TGn), and positions -12/-11 of element -10. As a rule, the promoter must contain only two of the three $\sigma^{70}$-dependent elements for activity; thus E. coli promoters can be loosely classified as -35/-10 (main class), TGn/-10 (also called extended -10) or -35/TGn. [18]

### 4.1.2 T4 phage transcription steps

T4 infects E. coli only during exponential growth. Transcription of early T4 genes begins immediately after infection. Thus, for effective infection, the phage must rapidly redirect $\sigma^{70}$-associated RNAP, which is actively involved in transcription of the host' genome, to early T4 promoters. Such immediate redirection often succeeds in part because most early T4 promoters contain the same $\sigma^{70}$-RNAP recognition elements (-35, TGn and -10 elements) and $\alpha$-CTD UP (lists of known sequences of early T4 promoters in [19]). However, sequence alignment of the early T4 promoters showed additional regions of overlap, and it is speculated that they contain other sequences that may op-

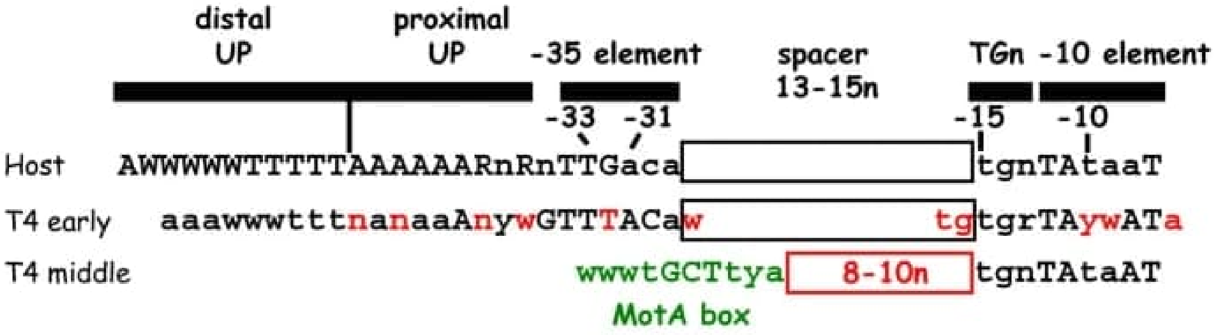

Figure 2: *Sequence comparison of E. coli host, early and intermediate T4 promoters.* Top shows sequences and positions of host promoter recognition elements for $\sigma^{70}$-RNAP (UP, -35, TGn, -10)[18]. Below, similar sequences found in early [19] and intermediate [20] T4 promoters, similarities in black and differences in red. W = A or T; R = A or G; Y = C or T, n = any nucleotide; a capital letter represents a more highly conservative base. [1]

timize host RNAP interaction with T4 promoter elements. Therefore, unlike most host promoters, which are classified as -35/-10, TGn/-10, or -35/TGn, early T4 promoters can be described as "super" UP/-35/TGn/-10 promoters [1]. Indeed, most early T4 promoters compete very well with host promoters for available RNAP [21].

## 4.2 Interaction between cyanophage and cyanobacteria

In another article, authors were interested in climate change. The study of this issue requires a detailed knowledge of the biological transformation of carbon on Earth. It became clear that the ocean is an important sink of atmospheric carbon dioxide. $CO_2$ uptake in the open ocean is dominated by two organisms: Prochlorococcus and Synechococcus [22, 23]. These organisms have become a model for studying the flow of carbon from $CO_2$

into microbial life [24]. However, the biological losses (antagonistic interactions, grazing, viral lysis) of Prochlorococcus and Synechoccus have been little studied. New genes are being discovered in viruses that act to maintain photosynthesis during infection [25] so that despite the eventual loss of fixed carbon in dissolved organic matter through lysis, $CO_2$ fixation can be maintained temporarily for relatively long latent periods virus. It has been shown that the cyanophage actually turns off the photosynthesis of cyanobacteria in the early stages of infection, but at the same time maintains the photosynthetic reactions itself. [28]

### 4.2.1 Effect of Light Intensity on Phage Transcription

The authors found that the rate of phage transcription increased under strong light, while DNA replication was the same as under low light. They also found that the expression rate of only one cyanophage gene was proportional to the light intensity. It was a photosynthetic AMG psbA encoding a D1 polypeptide located in the core of the PSII reaction center. Synechococcus also code for this protein, but they did not show any increase in transcription rate. [2]

Cyanobacteria, including marine Synechococcus, exhibit light-dependent transcription of psbAs [29, 30]. In a model cyanobacterium, the light-dependent change in psbA transcripts is affected by several factors, including alternative $\sigma$-factors [37, 38]. Moreover, D1 degradation products directly bind to upstream regions of psbA sequences, so that light-induced damage can positively influence psbA transcription. The authors were unable to find conserved motifs in the upstream regions in the S-PM2d psbA sequences that would be in common with Synechococcus.

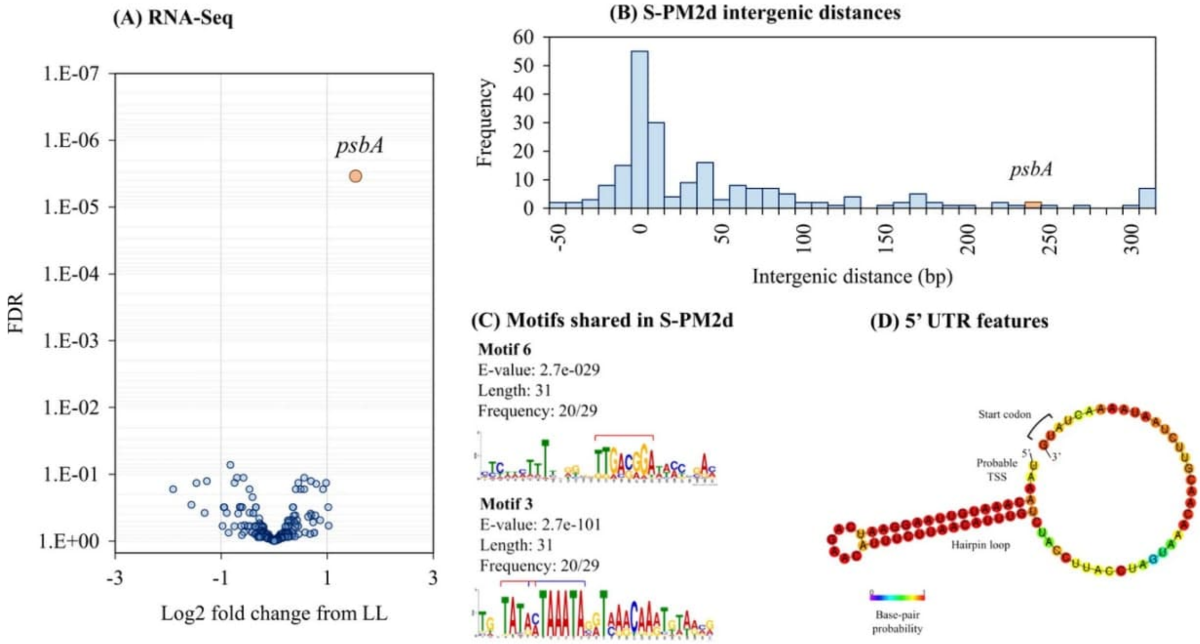

Figure 3: *Influence of light intensity on global gene expression.* **a** Volcano-plot with logarithmic scale showing relative transcriptome change (relative to low light exposure). The y-axis shows the adjusted p-value of false detection calculated using edgeR. The orange circle shows statistically significant differential gene expression, according to edgeR. **b** Histogram of the lengths of upstream intergenic sequences in the genome of the cyanophage S-PM2d. The cell containing psbA is shown in orange. **c** Upstream DNA sequence motifs found in cyanophage PSBAs. The red bars indicate the -35 and -10 elements of the $\sigma^{70}$ binding sites and the blue bar shows the binding site for $\sigma$ Gp55. **d** Predicted S-PM2d psbA 5'-UTR hairpin loop folding structure [2]

The intergenic upstream regulatory sequence in psbA is uncharacteristically long in S-PM2d (232 base pairs, compared to a genome-wide median of 6). This is also true for other cyanophages, where the upstream regulatory regions of psbA range from 125-453 bp. Also, with a high degree of reliability, 6 common motifs were found between the considered phages in these sequences [2]. Of these, two were present in S-PM2d. These two motifs are -35 (motif 6) and -10 (motif 3) (Fig. 3) elements of $\sigma^{70}$ transcription factor binding sites typical of early genes of T4-like phages [39]. In addition, it was found that, like in T4, motif 3 contains a site for the late $\sigma$-factor Gp55. Thus, S-PM2d has the motifs necessary for the coordinated expression of psbA at the initial and late stages of infection.

## 4.3 GeneMarkS-2

To search for prokaryotic genes, there are several tools (eg GeneMarkS, Glimmer3, Prodigal) that are known for fairly high accuracy in predicting protein-coding ORFs. On average, these tools are able to find over 97% of the genes in the validated test set in terms of correctly predicting gene ends 3' [41]. In addition, the accuracy of determining gene starts is on average 90% [41]. But the bulk of the genes that were not found belonged mainly to the atypical category, i.e. genes with patterns that do not match the species-specific trained model trained on the main part of the genome [42]. However, the authors have described a method by which higher accuracy in gene detection can be obtained. [3]

GeneMarkS-2 uses a rather complex gene model (Fig. 4). Most protein-coding regions in prokaryotic genomes have species-specific patterns of several

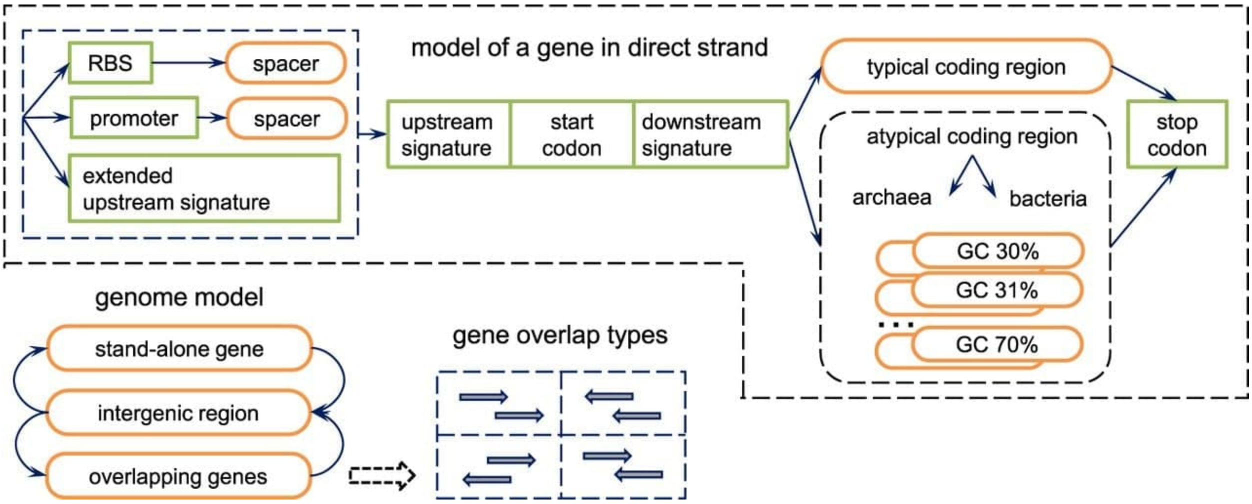

Figure 4: *Gene model in GeneMarkS-2.* Ground state diagram of a generalized hidden Markov model (GHHM) of a prokaryotic genomic sequence. The states shown in the top panel were used to model the gene in the forward direction. Genes in the reverse direction were modeled with an identical set of states (with reverse transition directions). Different states modeling genes in forward and reverse chains were linked through the state of the intergenic region or by overlapping regions in opposite chains. [3]

oligonucleotides (e.g., codons) [43]. GeneMarkS-2 studies these patterns and evaluates the parameters of a typical model of protein-coding regions, the three-period Markov chain [44], iteratively self-learning on the entire genome.

The unsupervised learning algorithm performs several two-stage iterations (Fig. 5). Each iteration results in (1) segmentation of the genome into coding proteins (CDS) and non-coding regions (gene prediction) and (2) re-estimation of parameters of the model.

*On the first iteration* the Viterbi algorithm calculates the maximum probable sequence of hidden states (Fig. 5) along the genome. After the first run of the Viterbi algorithm, all regions of the genome labeled as "protein-encoding" are assembled into a training set to evaluate the parameters of a "typical" (for the given genome) model. Similarly, regions labeled as "non-coding" are needed to estimate the parameters of the non-coding model,

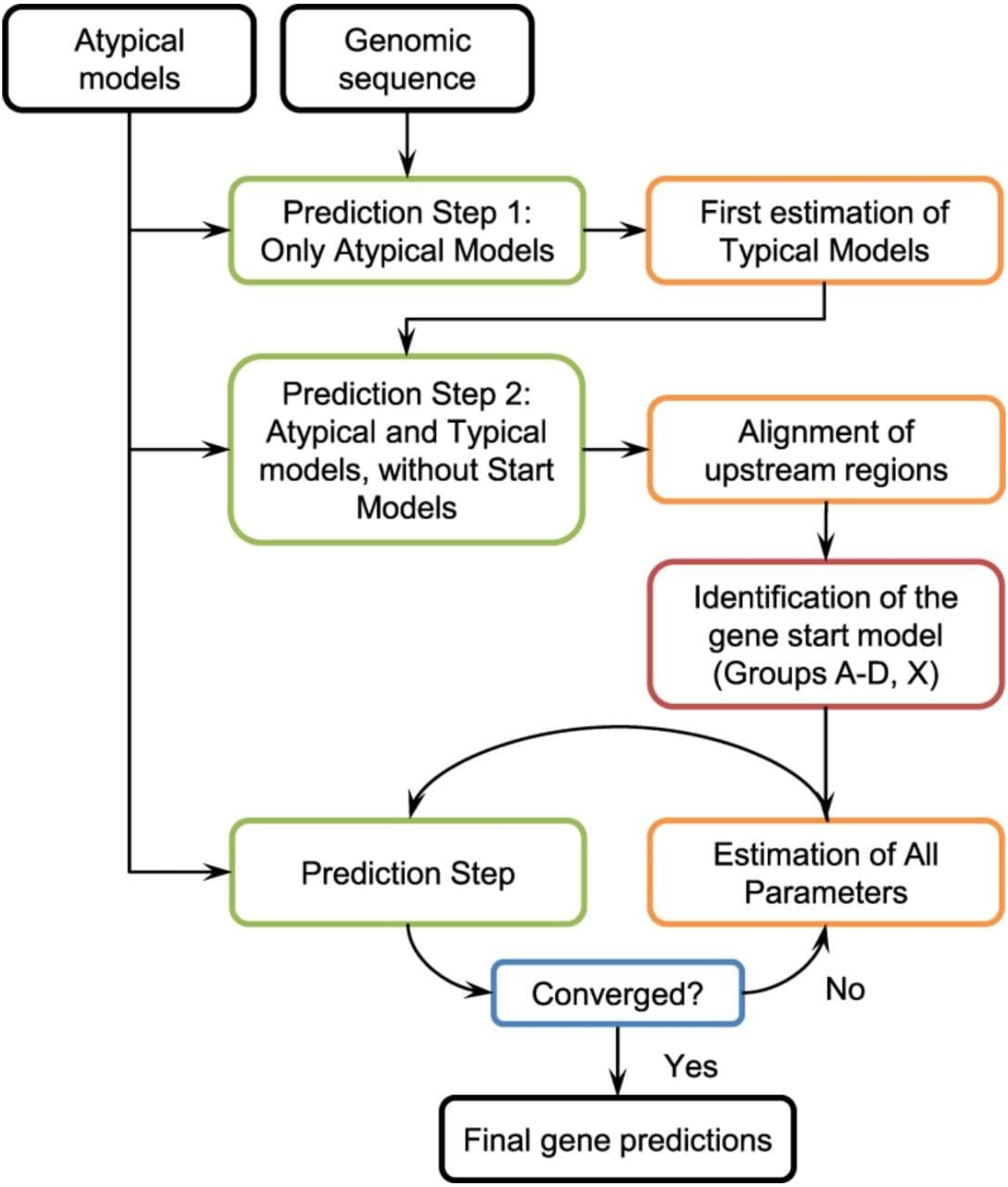
Atypical models
Genomic sequence
Prediction Step 1: Only Atypical Models
First estimation of Typical Models
Prediction Step 2: Atypical and Typical models, without Start Models
Alignment of upstream regions
Identification of the gene start model (Groups A-D, X)
Prediction Step
Estimation of All Parameters
Converged?
No
Yes
Final gene predictions

Figure 5: *Unsupervised Learning Chart for GMS2* [3]

which is constructed as a homogeneous second-order Markov chain.

*On the second iteration* the model selects sequences that are located around gene starts and infers patterns that encode regulation of transcription and/or translation. *Third iteration and beyond* GeneMarkS-2 continues the prediction/evaluation steps until the convergence condition is met ( 99% identity in the gene starts between successive iterations).

It was important to us that the observed error rate of GeneMarkS-2 was 4.4%, followed by Prodigal at 6.1%, GeneMarkS at 10.2% and finally Glimmer3 at 13.2%. Thus, GeneMarkS-2 made the highest number of correct predictions among the four gene finders. [3]

## 4.4 The MEME-Suite

The MEME-Suite is a programmable set of tools that analyze sequence' motifs. The core of the set is the `MEME` motif detection algorithm, which finds motifs in misaligned DNA, RNA or protein sequences [40]. The motif finder looks for *new* motifs in the provided sequences. There are many tools designed to look for motifs in DNA alone; The MEME-Suite can search for motifs and scan DNA, RNA and protein sequences. [4]

As described above, the `MEME` algorithm detects one or more motifs in a set of DNA, RNA or protein sequences using the expectation maximization method to fit a two-component final model to the set of sequences. By fitting the model to the data, a single motif is found, then the model erases the occurrences of the maximum probable motif thus found, and the process is repeated. Thus, many motifs are sought. The algorithm has two necessary parameters - the minimum and maximum width of the desired motifs. It

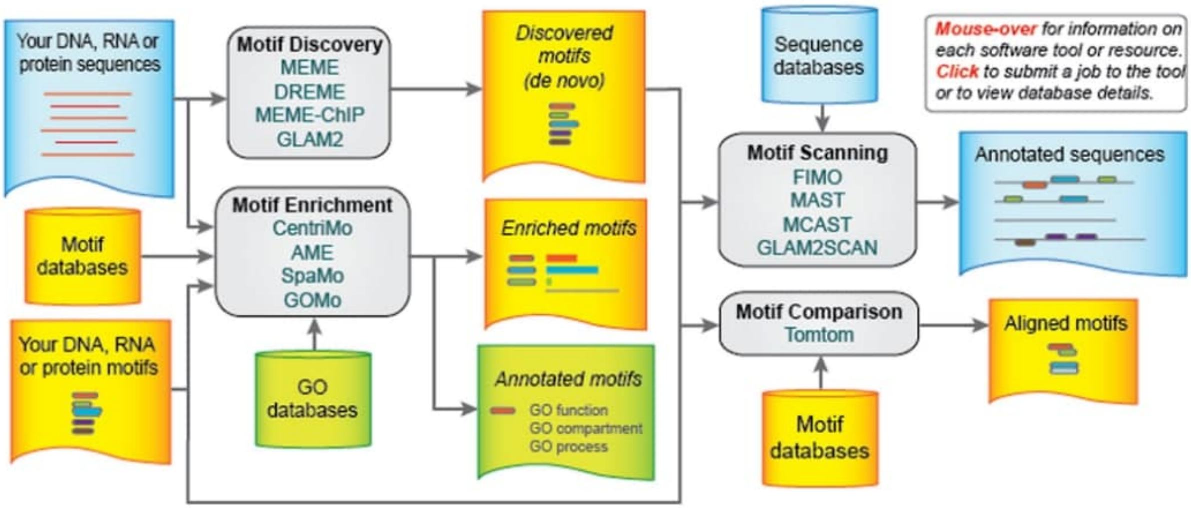

Figure 6: *What The MEME-Suite can do.* We need to find new motifs in sequences - this is indicated at the top in the middle. [4]

returns a profile of each motif and a threshold, which together can be used as a Bayesian optimal classifier to find occurrences of the motif in other databases. The algorithm estimates how many times each motif occurs in each sequence in the dataset and outputs the frequency of occurrence of that motif. [12]

# 5 Methods

## 5.1 Software

- Virus-Host DB - A database that contains information about the relationships between viruses and their hosts, represented as pairs of NCBI taxonomy identifiers for viruses and their hosts. [31]

- GeneMarkS-2 - software for searching for genes in prokaryotic genomes. [3]

- NCBI Entrez Programming Utilities - software, a set of eight server programs that provide a stable interface to the Entrez query and database system at NCBI [36]

- Biopython - a set of biological computing tools written in Python. [32]

- The MEME Suite - software for searching for new, unrevealed motifs (fixed length repeating patterns) in sequences. MEME breaks patterns of variable length into two or more separate motifs. [4]

- Python 3.8.5 - environment for running *.py scripts [33]

- Jupyter Notebook - a web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text. [34]

- https://github.com/poolsar42/phages-and-hosts - GitHub repository that contains all the underlying scripts. [35]

- Fasta_merge.ipynb - a script for combining the regulatory regions of the host and its phages into one file.

- Hosts_upstream_regions.ipynb - script for highlighting regulatory areas.

- Proteins_formation.ipynb - script for collecting protein sequences into one file.

- Searching_host_genomes.ipynb - script for unloading host genomes necessary for work.

- Unpacking.ipynb - script for unpacking downloaded genomes.

- VH_tsv_writer.ipynb - a script that writes the host ID and the ID of the phages adjacent to it to the file.

- Virus_genes.ipynb - script for unloading phage genomes.

- missing_phage_genomes.ipynb - script for searching for phages for which GMS2 does not find genes.

- proteins_convertation.ipynb - script for converting nucleotide sequences to protein ones.

- regex_promoter.ipynb - script for the primary analysis of the data obtained above

- run_gms2.py - script for searching genes in genomes using GeneMarkS-2

- tsv_write.ipynb - script for writing out information about the connection between hosts and phages in one tsv file

## 5.2 Getting common motifs

### 5.2.1 Unloading genomes

- The virushostdb.tsv file was uploaded from the Virus-Hist DB FTP server - it contains information about the connections between phages and their hosts.
- Using tsv_write.ipynb, we wrote out in a convenient format all the relationships between phages and hosts.
- Using Virus_genes.ipynb, phage genomes were uploaded.
- Launched Searching_host_genomes.ipynb - to extract host genomes.
- And unpacked with Unpacking.ipynb.

We have collected the genomes of all phages that are in the Virus-Host DB; for each phage, the hosts to which it can attach were also indicated. Their RefSeq ID. We downloaded everything from the FTP server. The host genomes were then downloaded from the RefSeq NCBI database.

### 5.2.2 Obtaining upstream regulatory regions

- Run the `run_gms2.py` script to search for genes in all available genomes
- And then we ran `Hosts_upstream_regions.ipynb` to isolate regions of 50 nucleotides in length before the transcription start point for each found gene

We found genes using GeneMarkS-2. GeneMarkS-2 also gave the position of each gene in the genome. Each gene was counted for its translated protein, also using GeneMarkS-2. After that, upstream regulatory regions, 50 nucleotides long, were found for each gene. This is our estimate of the distance at which there should be common motifs other than -35 and -10 elements.

### 5.2.3 Search for motifs

- The program that helped prepare the data files: `Fasta_merge.ipynb`. It creates a file that stores the parent regions of the bacterium and all adjacent phages. This file format is suitable for running MEME Suite.
- Launched MEME Suite with the resulting files.

# 6 Results and discussion

## 6.1 The result of the MEME-Suite

461 files were submitted for entry into the MEME-Suite. Each file contained the promoter regions of the host bacterium and the promoter regions of all known [31] phages attacking this bacterium. A total of 1946 motifs were found in 426 source files.

For each motif, the following files were also obtained:

- Motif logo. A multiple sequence logo is a graphical representation of the information content stored in a Multiple Sequence Alignment (MSA) system and provides a compact and intuitive representation of position-specific nucleotide composition of binding motifs, active sites, etc. in biological sequences.

- Position Frequency Matrix (PFM). This matrix is calculated as the ratio of the number of occurrences of each nucleotide in each position and the total number of motifs.

- Position-Specific Scoring Matrices (PSSM) - needed for use by database lookup programs such as MAST. This matrix is a matrix of logits, which are considered to be a hundred times the base 2 logarithm of the ratio $p$ at each motif position and something else [5], where $p$ is the probability of a particular letter at that position motive. In this work, we did not use these matrices.

Collectively, this data turned out to be large enough to include in the application. As an example, I will attach all three files for a separate array. They are available at the link: `https://github.com/poolsar42/BachelorThesis/tree/main/results_example`

## 6.2 Analysis of the results

For each motive, we found its localization and added a separate “location” column to the results table. This column contains the values “BOTH“ - if the motif was found in both the bacterium and the phage, “PHAGE” - if the motif was found only in the phage, or “HOST” if the motif was found only in the bacterium. To our surprise, some of the motifs were not found anywhere - for those we wrote “NOT FOUND” in this column. We explain this by the fact that the resulting ”averaged” motif has a different structure from any other, over which it was averaged.

91 motifs were common only in phages that infect one bacterium, but were not found in the bacteria themselves, these motifs were found in 53 source files. 858 motifs were found within the bacterial promoter regions themselves, but they were not found in the phages that infect these bacteria. They were found in 329 source files. And finally, 727 motifs were common to bacteria and some of the phages that infect them, they were found in 195 source files. The remaining 270 found motifs were not found in any promoter regions in their files.

Also, we know that the -35 and -10 promoter elements can often be repeated between the bacterium and the host. We were not interested in them. We found possible sequences for -35 and -10 plots [6, 7, 8] and removed them from our dataset using regular expressions [9]. The remaining table

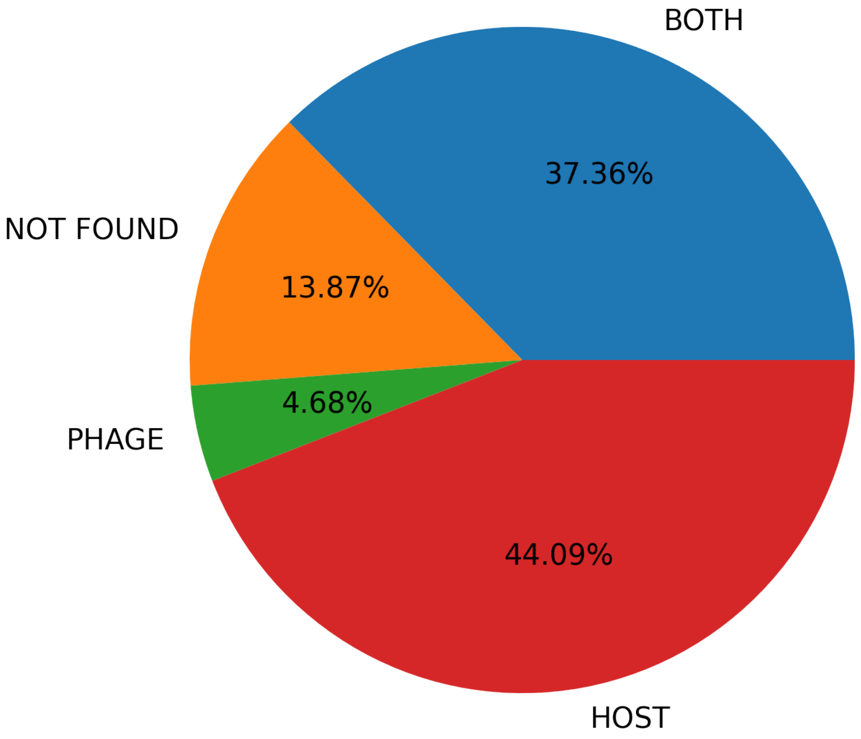

Figure 7: *The result of searching for motifs in files without deleting -35 and -10 elements.* As a percentage of 1946 motives.

contained 1344 motifs from 362 source files. Of them:

- 63 motifs were common only among phages in 39 files.
- 624 motifs were found only among the bacterial promoter regions in 261 files.
- 533 motifs were shared between hosts and their phages in 153 files.
- There were also 124 mystical motifs that did not belong to either phages or bacteria in 104 files.

We were interested in the position of motifs relative to the start of transcription. As already mentioned, we isolated 50 nucleotides - before the start of transcription. Again, using regular expressions, we found the

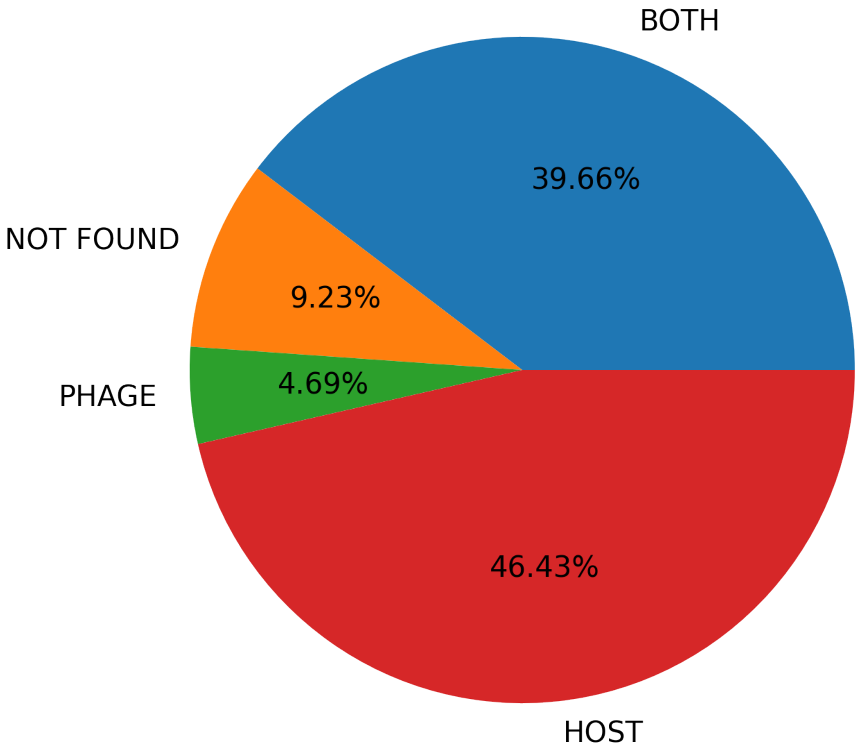

Figure 8: *The result of searching for motifs in files after deleting -35 and -10 elements.* As a percentage of 1344 motives.

position of the motif in each 50-nucleotide region in which it occurs. For each motif, we calculated the average value for its left end relative to the start of transcription, for its right end, and the standard error for both ends. All this was recorded in the results table. The table turned out to be large enough to be included in the application. It is available at the link: `https://raw.githubusercontent.com/poolsar42/BachelorThesis/main/results.tsv`

Also, as I wrote above, we were surprised when we found that some motifs were not found in any of the presented upstream regulatory regions. Also, in many motifs, you can notice a large number of letters W, N, Y, and some more that are not part of the usual A, T, G and C for nucleotides. The fact is that different nucleotides can be located at the same place in the motif in different sequences. And for such dual places, MEME-Suite uses a special

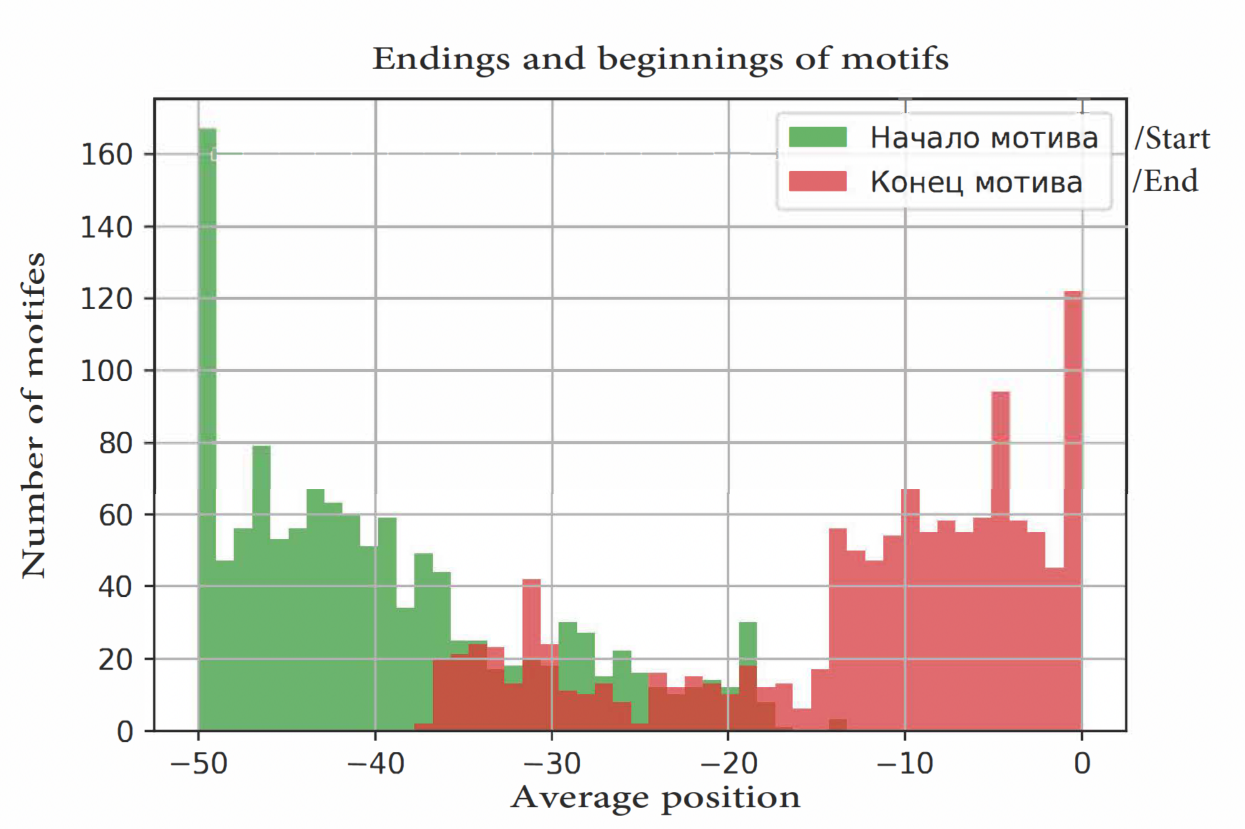

Figure 9: *Averaged for each motif, the positions of its beginning (green histogram) and end (red) relative to the transcription start point.*

alphabet, in which one letter replaces all possible ones at a given position [10].

But when we started to analyze these sequences, we noticed that some of the regular expressions and the dual letters that should match them do not match those given in the alphabet [10].

In our search for motifs, the minimum width of the searched motifs was 10 nucleotides. The width of several known motifs (-35, -10 etc.) ranges from 6 to 10 nucleotides. In addition, an intergenic distance of 50 nucleotides is not always sufficient to search for common motifs. The transcription start point was also not always correct. Taking this into account, in 42% of cases, common motifs between phages by their hosts were found. In 54% of cases, common motifs were found only between phages. We consider these results, which are primary for this problem, to be quite encouraging. The percentages here are not relative to the total number of motifs, but relative to the total number of files (types of bacteria).

### 6.2.1 Correlation analysis

We also became interested in whether the motives of bacteriophages that attack one bacterium somehow correlate with each other, and also whether there is a correlation between the motives of bacteriophages from the same taxonomic category. To begin with, we wanted to calculate the Pearson correlation between the PFMs of these motives. But PFM is a two-dimensional array; we did not find a suitable function in the known Python libraries for calculating the correlation between 2D arrays. We borrowed a script to precalculate such a correlation `preliminary_correlation.py` using tricks from here [11]. Correlation was calculated for each file with motifs. An example of the found correlation can be found at the link: `https://github.com/poolsar42/BachelorThesis/blob/main/MEME_0_1115_motif.tsv`. The first argument in each position of this matrix is the calculated Pearson correlation coefficient.

## 6.3 Conclusions

- The presence of common motifs in the regulatory elements of bacteriophages and bacteria makes it possible to determine the degree of their relationship.

- The presence of common motifs in regulatory elements between different bacteriophages makes it possible to determine the degree of kinship of these bacteriophages.

# References

[1] Hinton D.M. Transcriptional control in the prereplicative phase of T4 development // Virology journal. 2010. V. 7. P. 289. https://doi.org/10.1186/1743-422X-7-289

[2] Puxty, R.J., Evans, D.J., Millard, A.D. et al. Energy limitation of cyanophage development: implications for marine carbon cycling // ISME J. 2018, 12, 1273–1286. https://doi.org/10.1038/s41396-017-0043-3

[3] Lomsadze, A., Gemayel, K., Tang, S., and Borodovsky, M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes Genome research, 2018. 28(7), 1079–1089. https://doi.org/10.1101/gr.230615.117

[4] Timothy L. Bailey, James Johnson, Charles E. Grant, William S. Noble, The MEME Suite, Nucleic Acids Research, Volume 43, Issue W1, 1 July 2015, Pages W39–W49, https://doi.org/10.1093/nar/gkv416

[5] https://kodomo.fbb.msu.ru/~partyhard/term4/pr9/meme_out_oops/meme.html#pssm1

[6] https://www.addgene.org/mol-bio-reference/promoters/

[7] Chang-Hui Shen // Gene Expression: Transcription of the Genetic Code // Diagnostic Molecular Biology, 2019 https://doi.org/10.1016/B978-0-12-802823-0.00003-1

[8] https://en.wikipedia.org/wiki/Promoter_(genetics)#Bacterial

[9] https://docs.python.org/3/library/re.html#module-re

[10] https://meme-suite.org/meme/doc/alphabet-format.html#standard_DNA

[11] https://stackoverflow.com/questions/30143417/computing-the

-correlation-coefficient-between-two-multi-dimensional-arrays

[12] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36. PMID: 7584402.

[13] https://en.wikipedia.org/wiki/Bacteriophage

[14] Monk, A., Rees, C., Barrow, P., Hagens, S. and Harper, D. (2010), Bacteriophage applications: where are we now? Letters in Applied Microbiology, 51: 363-369. https://doi.org/10.1111/j.1472-765X.2010.02916.x

[15] David R. Harper, Benjamin H Burrowes, Elizabeth M. Kutter (15 August 2014) Bacteriophage: Therapeutic Uses, Letters in Applied Microbiology. https://doi.org/10.1002/9780470015902.a0020000.pub2

[16] https://sitn.hms.harvard.edu/flash/2018/bacteriophage-solution-antibiotics-problem/

[17] Paget, M.S., Helmann, J.D. The $sigma^70$ family of sigma factors. Genome Biol 4, 203 (2003). https://doi.org/10.1186/gb-2003-4-1-203

[18] Hook-Barnard IG, Hinton DM: Transcription Initiation by Mix and Match Elements: Flexibility for Polymerase Binding to Bacterial Promoters. Gene Regulation and Systems Biology 2007. http://la-press.com/article.php?article_id=481:275-293

[19] Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W: Bacteriophage T4 genome. Microbiol Mol Biol Rev 2003, 67: 86-156. https://doi.org/10.1128/MMBR.67.1.86-156.2003

[20] Stoškiene, G., Truncaite, L., Zajančkauskaite, A. and Nivinskas, R. (2007), Middle promoters constitute the most abundant and diverse class

of promoters in bacteriophage T4. Molecular Microbiology, 64: 421-434. https://doi.org/10.1111/j.1365-2958.2007.05659.x

[21] Wilkens K, Ruger W: Characterization of bacteriophage T4 early promoters in vivo with a new promoter probe vector. Plasmid 1996, 35: 108-120. https://doi.org/10.1006/plas.1996.0013

[22] Zwirglmaier K, Jardillier L, Ostrowski M, Mazard S, Garczarek L, Vaulot D, et al. Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. Environ Microbiol. 2008;10:147–161.

[23] Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WKW, Platt T, et al. Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. Science. 2006;312:918–921.

[24] Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, et al. Ecological genomics of marine picocyanobacteria. Microbiol Mol Biol Rev. 2009;73:249–299.

[25] Mann NH, Cook A, Millard AD, Bailey S, Clokie M. Bacterial photosynthesis genes in a virus. Nature. 2003;424:741–742.

[26] Stitt B, Hinton DM: Regulation of middle-mode transcription.In Molecular biology of bacteriophage T4.Edited by: Karam JD, Drake J, Kreuzer KN, Mosig G, Hall D, Eiserling F, Black L, Spicer E, Kutter E, Carlson K, Miller ES.Washington, D.C.: American Society for Microbiology; 1994:142-160

[27] Brody E, Rabussay D, Hall D:Regulation of transcription of prereplicative genes. In Bacteriophage T4.Edited by: Mathews CK, Kutter EM, Mosig G,Berget PB. Washington, D. C.: American Society for Microbiology;1983:174-183

[28] Puxty RJ, Millard AD, Evans DJ, Scanlan DJ. Viruses inhibit $CO^2$

fixation in the most abundant phototrophs on Earth. Curr Biol. 2016;26:1585–1589.

[29] Garczarek L, Dufresne A, Blot N, Cockshutt AM, Peyrat A, Campbell DA, et al. Function and evolution of the psbA gene family in marine Synechococcus: Synechococcus sp. WH7803 as a case study. ISME J. 2008;2:937–953.

[30] Mulo P, Sakurai I, Aro EM. Strategies for psbA gene expression in cyanobacteria, green algae and higher plants: From transcription to PSII repair. Biochim Biophys Acta Bioenerg. 2012;1817:247–257.

[31] `https://www.genome.jp/virushostdb/`

[32] `https://biopython.org/`

[33] `https://docs.python.org/3/`

[34] `https://jupyter.org/`

[35] `https://github.com/poolsar42/phages-and-hosts`

[36] `https://www.ncbi.nlm.nih.gov/books/NBK25501/`

[37] Imamura S, Yoshihara S, Nakano S, Shiozaki N, Yamada A, Tanaka K, et al. Purification, characterization, and gene expression of all sigma factors of RNA polymerase in a cyanobacterium. J Mol Biol. 2003;325:857–872.

[38] Imamura S, Asayama M, Shirai M. In vitro transcription analysis by reconstituted cyanobacterial RNA polymerase: roles of group 1 and 2 sigma factors and a core subunit, RpoC2. Genes Cells. 2004;9:1175–1187.

[39] Miller ES, Kutter E, Mosiq G, Arisaka F, Kunisawa T, Rüger W. Bacteriophage T4 genome. Microbiol Mol Biol Rev. 2003;67:86–156.

[40] Altman, R, Brutlag, D, Karp, P, Lathrop, R, & Searls, D. Proceedings: Second international conference on intelligent systems for molecular biology. United States.

[41] Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.

2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119

[42] Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Medigue C, Danchin A. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res 23: 3554–3562.

[43] Fickett JW, Tung CS. 1992. Assessment of protein coding measures. Nucleic Acids Res 20: 6441–6450

[44] Borodovsky M, Sprizhitskii Y, Golovanov E, Aleksandrov A. 1986b. Statistical patterns in primary structures of the functional regions of the genome of Escherichia coli. Computer recognition of coding regions. Mol Biol 20