Unterschrift BetreuerIn

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

DIPLOMARBEIT

Data Analysis of HT-SELEX against Complex Targets

ausgeführt am Institut für Verfahrenstechnik
der Technischen Universität Wien

unter der Anleitung von
Ao. Prof. Mag. Dr.rer.nat Andreas Farnleitner MSc.Tox.
Senior Scientist Dipl.-Ing. Dr.techn. Georg Reischer
Projektass. Dipl.-Ing. Dr.rer.nat. Claudia Kolm

durch

Ulrich Josef Aschl

Getreidemarkt 1A, 1060 Wien

November 22, 2021

Unterschrift StudentIn

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass die vorliegende Arbeit nach den anerkannten Grundsätzen für wissenschaftliche Abhandlungen von mir selbstständig erstellt wurde. Alle verwendeten Hilfsmittel, insbesondere die zugrunde gelegte Literatur, sind in dieser Arbeit genannt und aufgelistet. Die aus den Quellen wörtlich entnommenen Stellen sind als solche kenntlich gemacht.

Das Thema dieser Arbeit wurde von mir bisher weder im In- noch Ausland einer Beurteilerin/einem Beurteiler zur Begutachtung in irgendeiner Form als Prüfungsarbeit vorgelegt. Diese Arbeit stimmt mit der von den BegutachterInnen beurteilten Arbeit überein.

Wien, November 2021

Unterschrift

Kurzfassung

In dieser Arbeit werden aktuelle bioinformatische Tools und Methoden zur Analyse von HT-SELEX Experimenten umgesetzt. SELEX (**S**ystematic **E**volution of **L**igands by **Ex**ponential Enrichment) ist ein *in vitro* Prozess der zur sequentiellen evolutionären Entwicklung von Aptameren genutzt wird. In High-Throughput-SELEX (HT-SELEX) wird SELEX mit Next Generation Sequencing kombiniert, wodurch große Datensätze (10^6-10^8) entstehen welche spezielle rechenintensive Analysemethoden erfordern. Aptamere sind kurze einstränge DNA- oder RNA-Oligonukleotide, welche aufgrund einzigartiger Faltung an spezifische Zielstrukturen binden können. Sie funktionieren ähnlich wie Antikörper und können beispielsweise als Nachweis in der Umweltanalytik dienen.
Für diese Arbeit sollten Datensätze dreier bakterieller Zell-SELEX Experimente mit dem Ziel-Bakterium *Enterococcus faecalis* analysiert werden. Das Ziel dieser Arbeit war es Aptamerkandidaten auszuwählen, welche voraussichtlich gut an die Zielstruktur binden könnten und daher charakterisiert werden sollten. Ebenfalls sollten qPCR-basierte Schmelzkurvenanalysen, die für das Monitoring von bakteriellen Zell-SELEX entwickelt wurden, validiert werden.
Vier bioinformatische Workflows wurden entwickelt. *Selex-ngs-prep* bereitet Rohdaten auf und gibt Information über Datenqualität. *Selex-assess* führt rudimentären SELEX-spezifische Datenanalysen aus und gibt Listen mit den am stärksten replizierten Sequenzen aus. *Selex-blaster* clustert die Daten anhand von Sekundärstrukturen, basierend auf ungebundenen und daher für eine Bindung verfügbaren Strängen, und gibt für jeden Cluster Sequenzen und gehäufte Motive aus. *Selex-kmer* versucht die Bindungsaffinität von Sequenzen anhand der enthaltenen K-mere zu bewerten.
Die Workflow konnten zuverlässig anzeigen ob SELEX Experimente erfolgreich und weitere Datenanalysen indiziert waren. Anhand der Ergebnisse konnten Fehlerquellen gefunden und SELEX- und Sequencing-Experimente optimiert werden. Dabei konnte auch gezeigt werden, dass eine qPCR-basierte Schmelzkurvenanalyse in Bezug auf SELEX zuverlässig Ab- und Anreicherungen von ssDNA anzeigen kann. Es wurden Sequenzen für eine weiterführende Charakterisierung anhand von Anreicherung, K-mer Bewertung und Clustering ermittelt. Aptamer EF05-508, der sich unter den ausgewählten Sequenzen befand, zeigte hohe Spezifität und Affinität für *E. faecalis*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Abstract

In this thesis current data analysis tools and methods for analyzing HT-SELEX experiments were employed. SELEX (**S**ystematic **E**volution of **L**igands by **Ex**ponential Enrichment) is an *in vitro* process that is used to develop aptamers in a sequential, evolution-like fashion. In High-throughput-SELEX (HT-SELEX), SELEX is combined with next generation sequencing, resulting in large data sets (10^6-10^8) that require specialized computational approaches for data analysis. Aptamers are short single-stranded DNA or RNA oligonucleotides, folding into unique structures and binding to a specific target. They work in a similar fashion as antibodies, and can be used i.e. to detect targets in environmental analysis. For this work data sets generated in three bacterial cell-SELEX experiments targeted at the bacterium *Enterococcus faecalis* were to be analysed. The aim was to prepare the data sets generated by sequencing and choose aptamer candidates for further characterization. Also, qPCR-based remelting curve analyses methods developed for monitoring the bacterial whole cell-SELEX process needed to be validated.
Four bioinformatic pipelines were developed to perform the analyses. *Selex-ngs-prep* performs data preprocessing and NGS quality analysis. *Selex-assess* was developed for rudimentary SELEX-specific data analysis tasks and returns lists of the most abundant sequences. In *Selex-blaster* an attempt was made to perform clustering based on unbound subsequences (looping regions), which are thought to be the target-specific parts of aptamers, and provide sequences and enriched motifs for every cluster. In *Selex-kmer* an attempt was made to predict binding affinity based on k-mer enrichment.
The pipelines were used to show whether SELEX experiments were successful and thus more thorough data analysis was indicated. They have proven helpful for determining error sources and consequently in optimizing SELEX and sequencing experiments. Moreover, NGS-based data analyses confirmed that qPCR-based remelting curve analyses of qPCR products during SELEX reliably indicate changes in ssDNA sequence diversity. Aptamer candidates were provided for further characterization using replication counts, k-mer-based scoring and clusterings. Amongst the aptamer candidates identified, aptamer EF05-508 was found to provide high binding and specificity against *E. faecalis*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Acknowledgements

First of all, I would like to thank **Prof. Andreas Farnleitner** for introducing me to the field of bioinformatics. Even though I was late for the very first lecture, I am really happy to have found the way to your class back then.
Georg Reischer, thank you for having me in your group and introducing me to the field of bioinformatics as well. Thanks for being available when I needed you and providing support, be it in the form of advice or by organizing access to the VSC (Vienna Scientific Cluster). On this note I should also send a thank you to the people at **VSC** for letting me prototype on their cluster computers.
I would like to express my greatest thanks to my supervisor **Dr. Claudia Kolm**. Even though I was still a beginner bioinformatician you were eager to have me in the group. You always took the time to listen to me when I was stuck, and pointed me in the right direction by providing guidance and suitable scientific articles. When I came up with odd new ideas on how to analyse our data, you listened and engaged with my ideas, and I will always be grateful for that! The respect, patience and acknowledgement you showed towards me and my work really means a lot.
I am very happy to have met **Isabella Cervenka**, who has become a dear friend of mine. I am amazed by your ambition and eagerness, as well as your capability of working through my three hours long crash course on Linux.
I'd also like to acknowledge the help of **Dr. Gabriel Vignolle**, who had an open ear for any bioinformatics related questions and checked in on my progress from time to time.
I must also give a shout-out to my brother **Bernhard**. You helped me with some really hard to crack nuts when I was experimenting with different analysis methods, and provided some interesting insights I couldn't come up with. Thank you for all your support and being there for me.
A big thank you goes out to my sister **Hedwig** and my friends **Peter**, **Lara** and **Selina** for always having my back when the diploma thesis looked too big to handle.
Last but not least I would like to thank **my parents**. Thank you for listening when I needed you. Without your support over the last few years I could have never made it. Your support meant so much to me, I will always be thankful for that.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Contents

1. Introduction **1**
- 1.1. Aptamers . . . 1
- 1.2. Systematic Evolution of Ligands by Exponential Enrichment . . . 4
- 1.3. Sequencing and Aptamer Identification . . . 6
  - 1.3.1. Sequencing Approaches . . . 6
  - 1.3.2. SELEX Bioinformatic . . . 8
- 1.4. Overview of Current SELEX Data Analysis Tools . . . 11
  - 1.4.1. Graphical User Interface Tools . . . 11
  - 1.4.2. Command Line Interface Tools . . . 14
  - 1.4.3. Libraries . . . 16
- 1.5. Identification of Aptamers against Bacteria . . . 17

2. Aim of Thesis **19**

3. Methods **20**
- 3.1. Pipeline Development . . . 20
- 3.2. Preprocessing and Quality Assessment: *selex-ngs-prep* . . . 20
- 3.3. SELEX Assessment: *selex-assess* . . . 21
- 3.4. K-mer-based Aptamer Scoring: selex-kmer . . . 23
- 3.5. Secondary Structure-Based Motif Detection: selex-blaster . . . 25

4. Results and Discussion **28**
- 4.1. Bioinformatic Analysis Pipelines . . . 28
  - 4.1.1. Preprocessing and Quality Assessment: *selex-ngs-prep* . . . 28
  - 4.1.2. SELEX Assessment: *selex-assess* . . . 31
  - 4.1.3. K-mer-based Aptamer Scoring: selex-kmer . . . 34
  - 4.1.4. Secondary Structure-Based Motif Detection: selex-blaster . . . 36
  - 4.1.5. Discussion of Workflow Development . . . 39
- 4.2. Analysis of whole-cell HT-SELEX Experiments . . . 40
  - 4.2.1. SELEX EF01 . . . 40
    - 4.2.1.1. Results from workflow *selex-ngs-prep* . . . 41
    - 4.2.1.2. Results from workflow *selex-assess* . . . 44
  - 4.2.2. SELEX EF05 . . . 47
    - 4.2.2.1. Results from workflow *selex-ngs-prep* . . . 47

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2.2.2. Results from workflow *selex-assess* 51
4.2.2.3. Results from workflow *selex-blaster* 56
4.2.2.4. Results from workflow *selex-kmer* 56
4.2.3. SELEX EF07 60
4.2.3.1. Results from workflow *selex-ngs-prep* 60
4.2.3.2. Results from workflow *selex-assess* 65
4.2.3.3. Results from workflow *selex-blaster* 68
4.2.3.4. Results from workflow *selex-kmer* 69

5. Conclusion and Outlook **73**

Bibliography **i**

A. Appendix **vii**

List of Figures

1. Example aptamer L454; aptamer from Allnutt et al., 2018[1], Levay et al., 2015[2] 2
2. An example of shape complementarity; adapted from Kinghorn et al., 2017[6] 2
3. SELEX aptamer targets; scientific advances in SELEX; adapted from Dunn et al., 2017[9] 3
4. Scheme of the SELEX process as an UML diagram 6
5. Visualization of shared motifs 7
6. Starting library analysis; adapted from Blind & Blank, 2014[24] . . 8
7. Aptamer family population tracing; adapted from Schütze et al., 2011[15] 9

8. Schemes of the actions performed in *selex-assess* 24
9. Scheme of secondary structure-based pairwise alignment step in *selex-blaster* 27

10. Scheme of the *selex-ngs-prep* workflow 30
11. Scheme of the *selex-assess* workflow 32
12. Scheme of the *selex-kmer* workflow 35
13. Scheme of the *selex-blaster* workflow 38
14. EF01: Quality profiles of raw and preprocessed reads 42

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

15. EF01: Preprocessing plots . . . 43
16. EF01: Visualization of enrichment dynamics . . . 45
17. EF01: Nucleotide dynamics plots . . . 46
18. EF05: Quality profiles of raw and preprocessed reads . . . 48
19. EF05: Preprocessing plots . . . 49
20. EF05: Visualization of enrichment dynamics . . . 52
21. EF05: Nucleotide dynamics plots . . . 53
22. EF07: Quality profiles of raw and preprocessed reads . . . 62
23. EF07: Preprocessing plots . . . 63
24. EF07: Motif logo of concatemer (R9) . . . 64
25. EF07: Visualization of enrichment dynamics . . . 66
26. EF07: Nucleotide dynamics plots . . . 67

List of Tables

1. Examples of bioinformatic tools used in SELEX data analysis . . . 12

2. EF01: Sequenced reads for R0-R9 . . . 44
3. EF01: Discarded reads (in %) . . . 44
4. EF05: Sequenced reads for R02-R11 . . . 50
5. EF05: Discarded reads (in %) . . . 50
6. EF05: Top 25 Reads from last enrichment round . . . 54
7. EF05: Chracterized Sequences . . . 55
8. EF05: Found motifs in clustered data (1/2) . . . 57
9. EF05: Found motifs in clustered data (2/2) . . . 58
10. EF05: Maximum shifting scores from last enrichment round . . . 59
11. EF05: Motifs of top 1000 sequences by max shifting score . . . 59
12. EF07: Targets used in the Toggle-SELEX approach . . . 60
13. EF07: Sequenced reads for R02-R09 . . . 61
14. EF07: Discarded reads (in %) . . . 64
15. EF07: Increase of concatemers . . . 65
16. EF07: Top 25 Reads from last enrichment round . . . 68
17. EF07: Found motifs in clustered data (1/2) . . . 70
18. EF07: Found motifs in clustered data (2/2) . . . 71
19. EF07: Maximum shifting scores for target *E. faecalis* . . . 72

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

20. EF07: Motifs of top 1000 sequences by max shifting score for *E. faecalis* 72

21. EF01: Reads after preprocessing xi
22. EF01: Discarded reads xi
23. EF01: Reads after preprocessing (in %) xii
24. EF01: Top 25 reads from last enrichment round xii
25. EF05: Reads after preprocessing xiii
26. EF05: Discarded reads xiii
27. EF05: Reads after preprocessing (in %) xiv
28. EF07: Reads after preprocessing xiv
29. EF07: Discarded reads xv
30. EF07: Reads after preprocessing (in %) xv
31. EF07: Maximum shifting scores for target *E. faecium* xvi
32. EF07: Maximum shifting scores for target *E. durans* xvi
33. EF07: Maximum shifting scores for target *E. hirae* xvii
34. EF07: Motifs of top 1000 sequences by max shifting score for *E. faecium* xvii
35. EF07: Motifs of top 1000 sequences by max shifting score for *E. durans* xvii
36. EF07: Motifs of top 1000 sequences by max shifting score for *E. hirae*xvii

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1. Introduction

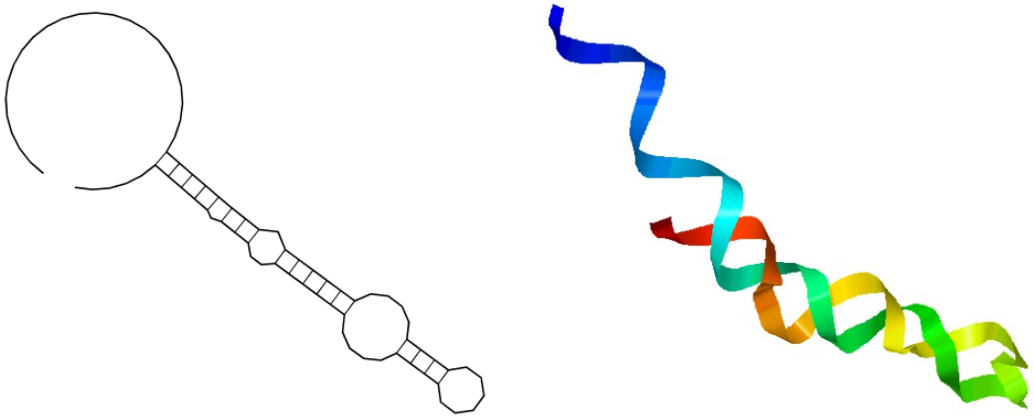
1.1. Aptamers

General Background Aptamers are synthetic oligonucleotides (single-stranded DNA or RNA molecules), which are generated via an *in vitro* selection process called SELEX (Systematic Evolution of Ligands by Exponential Enrichment)[7]. Aptamers fold into three-dimensional structures minimizing free energy, depending on the arrangement of nucleic acids and properties such as temperature or chemical composition of the buffer or liquid it is in. They bind to a specific target with high affinity by non-covalent interactions, similar to antibodies made from amino acids, by fitting perfectly into an area on the surface of the target and minimizing free energy even further. Exemplary secondary and tertiary structure predictions of an aptamer are shown in Figure 1. An example for shape complementarity of aptamer-target complexes can be seen in Figure 2

RNA aptamers can also be encountered in nature where they constitute the ligand binding parts of riboswitches and ribozymes. Riboswitches are RNA molecules which inhibit or promote the synthesis of proteins by binding specifically to certain nucleotide strands in the genome. Ribozymes are RNA molecules that are similar in function to protein enzymes and act as catalysts for certain chemical reactions[8].

Targets and Applications As aptamers are able to reliably bind to their target, they have become interesting affinity reagents for diagnostic and therapeutic applications. Within the last decades, aptamers have been developed to recognize small molecules such as metabolites, drugs, and environmental toxins[9] and numerous different proteins, such as thrombin, human interleukin (IL)-10 receptor and human 4-1BB receptor (Figure 1), cholera toxin (*Vibrio cholerae*), amyloid fibrils, prostate-specific antigen or hemoglobin[10, 2, 11, 12, 13]. More recently, aptamers have also been developed against *complex targets*, such as bacterial or mammalian cells, which offer multiple different binding sites on their cell surface[14, 15]. An overview of aptamer targets and nucleic acid-backbone structures as their applications is shown in Figure 3.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Secondary structure **(b)** Tertiary structure

Fig. 1: RNA-Aptamer L454 *'p5-GGAAUCUCGCGCUCGUUGGUACCCUU-AAAAUAAAGGCAUA-p3'*, developed using HT-SELEX against two targets: human interleukin (IL)-10 receptor and human 4-1BB receptor proteins[1, 2]. Figure 1a: L454 secondary structure prediction using RNAfold[3] with standard parameters. Figure 1b: L454 tertiary structure prediction using the folding simulation tool RNAComposer[4, 5].

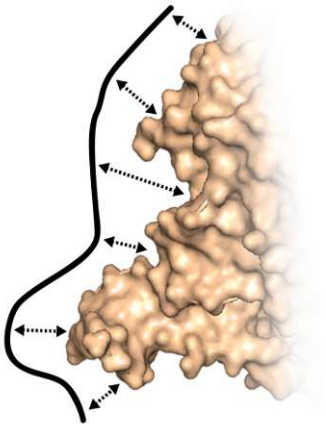
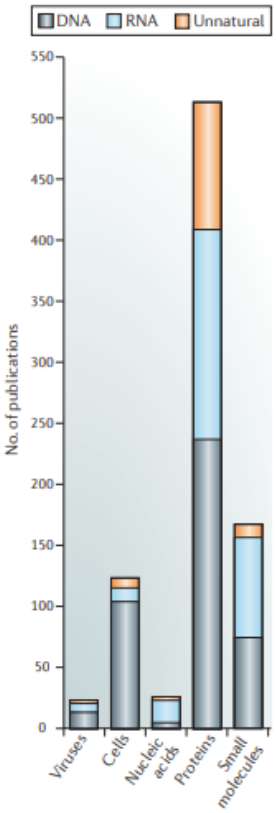
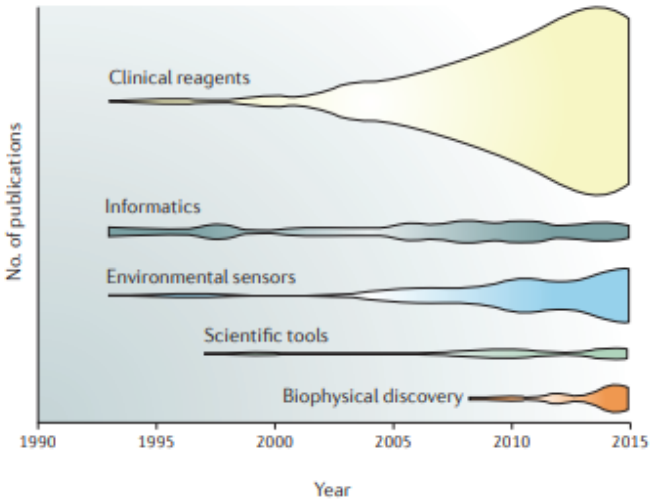


Fig. 2: Simplified example showing how aptamers are recognizing their target. They bind to their target by minimizing the exposed surface area. This is highly dependent on the nucleotide sequence. Figure adapted from Kinghorn et al., 2017[6].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Aptamer target classes and nucleic acid-backbone structures of in vitro selected aptamers ("unnatural" refers to the use of modified nucleotides).



(b) Aptamer applications categorized into biophysical discovery (e.g. structural and thermodynamic analyses); clinical reagents (therapeutics, drug conjugates, diagnostic agents and clinically tailored biosensors); informatics (*in silico* modeling and selections, machine learning and software development); scientific tools (e.g. chromatography, non-clinical sensors, gene regulation and nanotechnology) and environmental sensors (e.g. food and water sample analysis).

Fig. 3: Figures adapted from Dunn et al., 2017[9].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Aptamer Properties Aptamers can be produced relatively easy and cheap, compared to antibodies, which are usually produced *in vivo*, raising ethical questions. Unlike antibodies, aptamers are produced completely synthetically by chemical synthesis. They can be tagged with fluorophors or different chemical groups for detection and immobilization. Moreover, aptamers are resistant to heat denaturation and recover and take their designated structure when the optimal temperature is reached again.

A major disadvantage of using aptamers for detection or therapeutics is their vulnerability to nucleases, which are enzymes produced in organisms to break down nucleotide strands[13].

1.2. Systematic Evolution of Ligands by Exponential Enrichment

Systematic evolution of ligands by exponential enrichment (SELEX) is an *in vitro* selection process, which was simultaneously developed by Tuerk and Gold[7] and Szostak and Ellington[16] in 1990. It is an iterative evolutionary-like process, which was initially designed to enrich RNA aptamers selected against specific targets. Since then, the process has been used to generate both RNA and DNA aptamers.

SELEX Process The SELEX process is sketched in Algorithm 1 and Figure 4. First, a random pool of single-strand DNA or RNA is designed and chemically synthesized. This aptamer pool, called the starting library, consists of up to 10^{15} strands, all with distinct sequences. During SELEX the target is iteratively exposed to the aptamer pool in order to select and enrich sequences with affinity for the target, resulting in target-aptamer complexes. In each SELEX round, bound aptamers are partitioned from non-binding sequences after incubation with the target, then eluted from the target and enriched by PCR amplification for the next round of SELEX. In order to select aptamers with desired binding properties (high affinity and specificity for a target of interest), a selection pressure needs to be applied during SELEX by e.g., increasing the number of washes, decreasing the incubation time or introducing counter-selection rounds[17]. In this way, aptamer sequences need to compete for an epitope ("survival of the fittest"). Over the course of the SELEX experiment, the aptamer pool is expected to decrease in heterogeneity, while more and more aptamer candidates with affinity/specificity to the target are enriched[7].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Algorithm 1: SELEX procedure sketched as pseudocode

Input : a_0 ... randomly generated nucleotide library
T ... set of target structures
r_{max} ... number of SELEX rounds
s ... incubation time

Output: a ... library and r_{max} sets of enriched aptamers

for i *ranging from* 1 *to* r_{max} **do**

```
    # Take fresh target from T
    t = T.get_fresh_target();
    # Expose target to aptamer set from previous SELEX round
    # (or library) and wait s minutes.
    t = t.expose_to(a_{i-1});
    t = t.incubate(s);
    # Wash off non-binding DNA from exposed target.
    t = t.wash_unbound_DNA();
    # Recover bound aptamers from target.
    a_i = t.elute_aptamers();
    # Amplify a_i to the original size of library a_0.
    a_i = pcr(a_i, size(a_0));
# Return all enriched aptamer pools for sequencing
return a;
```

Toggle-SELEX Toggle-SELEX is a variation of the standard SELEX protocol. In Toggle-SELEX different targets of interest are used and alternated in consecutive rounds of a SELEX experiment in order to select and identify cross-reactive aptamers[18].

Starting library and aptamer sequence design Aptamers developed in SELEX usually consist of a forward primer, a random region and a reverse primer. The forward primer at the 5'-end, and the reverse primer at the 3'-end are fixed regions needed for PCR amplification to step up the number of molecules during the *in vitro* selection process. The random region ranges from 20 to 60 nucleotides in length, containing all four nucleotides (A, C, G and T). Aptamers binding to same target structure may show similarities in their structure and/or nucleotide composition. Apparently, the binding potential of aptamers stems from unbound looping regions, which are available to interact with other molecules[19]. A motif is usually a short (4 to 12 nucleotides) subsequence, which is enriched and therefore observed in multiple aptamer sequences[20, 21]. Figure 5 shows a set of motifs.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

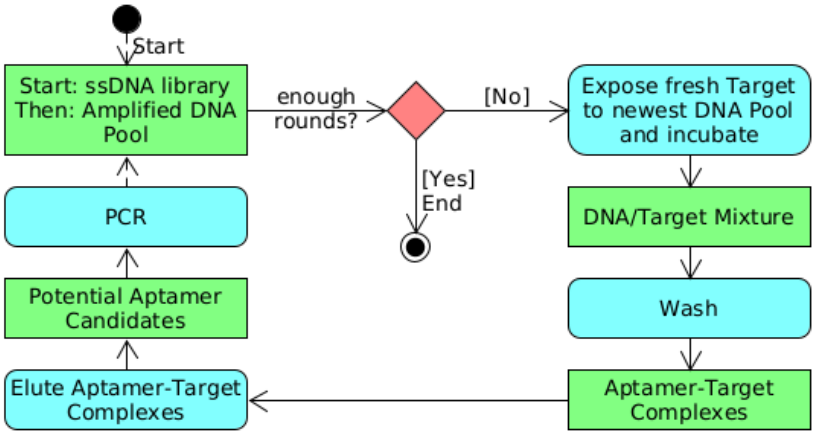


Fig. 4: Scheme of the SELEX process as an UML diagram.

1.3. Sequencing and Aptamer Identification

1.3.1. Sequencing Approaches

Sanger sequencing has been the traditional sequencing technique since the SELEX procedure was developed. Even though newer more powerful sequencing technologies have emerged in the last two decades, it is still widely used for sequencing tasks, including SELEX. After the final round of SELEX, when enrichment in the SELEX pool is observed via binding assays, the aptamer pool is cloned and a few aptamers, usually the most frequent 30 to 100 clones, are picked and sequenced[15, 22].
Sanger sequencing is still a common approach as it is easy to perform and relatively cheap. It is available at most facilities and is a well-known procedure, beating most other techniques regarding error rates as it is not as susceptible to inserts, deletes or mutations. Unfortunately, its low resolution makes it impossible to gain insight on the dynamics of the SELEX process and to identify rare but high affinity binders. Moreover, it is a time-consuming approach and studies have shown that the most frequent sequences found are not necessarily the best binders[23].
Starting in the last decade, next generation sequencing (NGS) was applied to SELEX aptamer pools. Depending on the used technique, sequence datasets comprise between 10^6 to 10^7 reads. Reads sequenced in NGS are usually of lengths up to 300 nucleotides, fitting well for SELEX. SELEX employing NGS is commonly referred to as HT-SELEX, short for high-throughput-SELEX[15].
Sequencing using NGS reveals only a very small fraction of the original library size of 10^{15} different nucleotide strands, but still imposes major computational

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

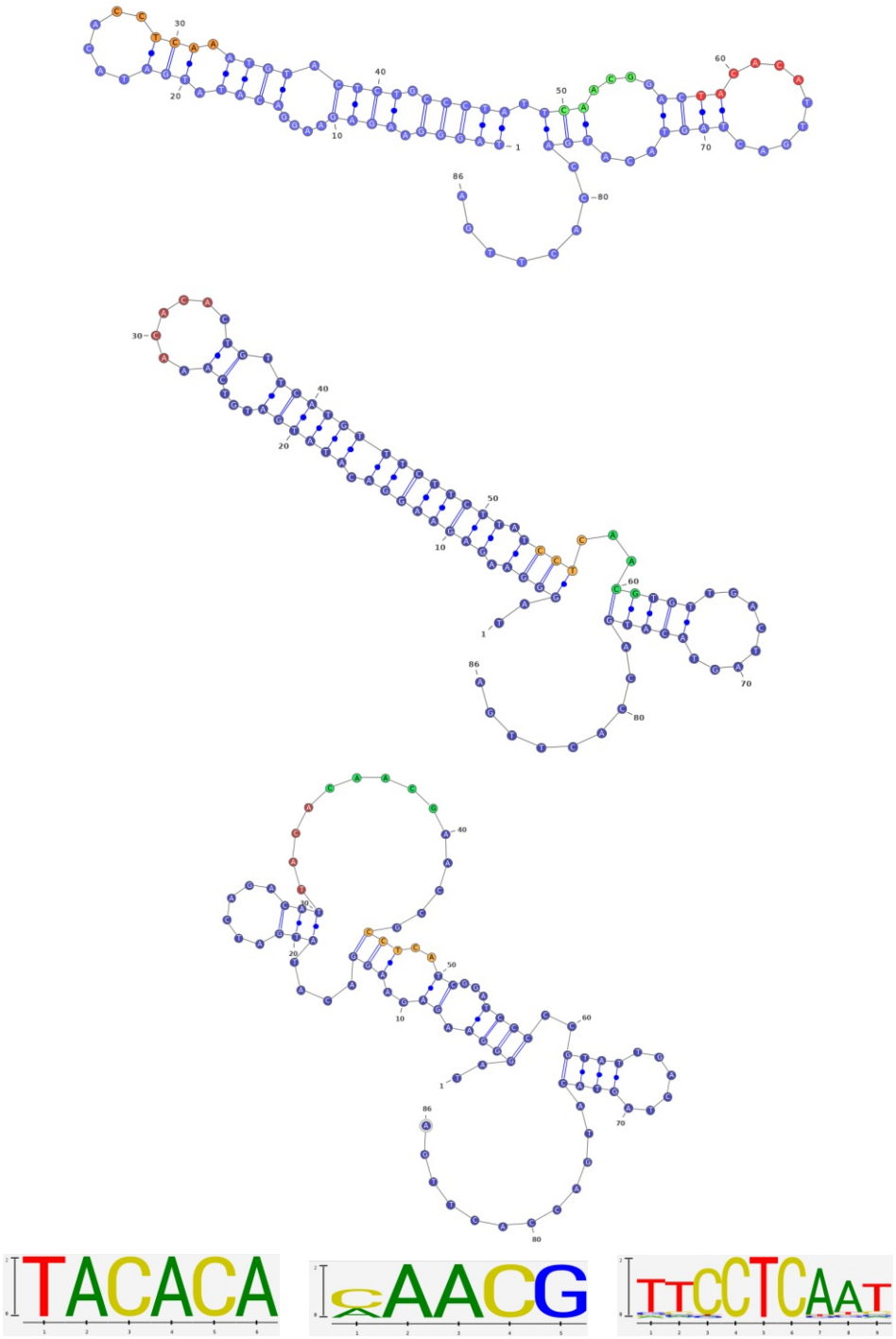


Fig. 5: Figure a) shows a sequence and the two analog sequences b) and c). Discovered motifs TACACA, CAACG and CCTCAA are marked in color. Figures d) to f) show the motif logos. Sequences stem from SELEX experiment EF07. Motif detection, motif logos and figures were done using AptaTrace[20].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

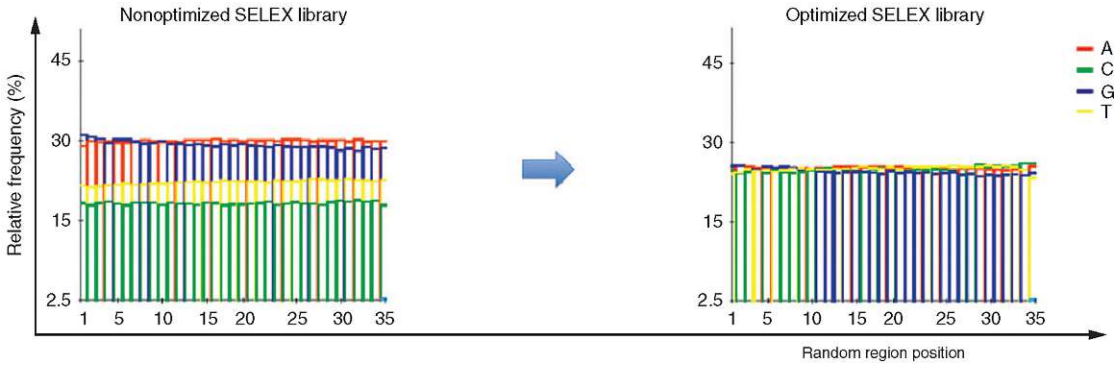


Fig. 6: Starting library analysis plot. Ideally nucleotides are distributed evenly at every position for a successful SELEX experiment. Figure adapted from Blind & Blank, 2014[24].

requirements. Employing NGS for SELEX experiments has lead to many new possibilities for aptamer development.

1.3.2. SELEX Bioinformatic

The extensive number of sequences gathered during every SELEX round can be used for all kinds of novel analyses such as i) starting library quality analysis, ii) single-round aptamer detection, iii) aptamer population tracing, iv) motif detection and clustering or v) mutational dynamics tracking[24, 25, 15, 26].

Starting Library Quality Analysis Typically, the random region of the starting ssDNA/RNA library is completely random, i.e. equal distribution of A, C, G and T (25% each), and does not show bias towards certain nucleotides, k-mers or sequences. As the selection heavily depends on the randomness of the library, the quality of the library should ideally be analysed prior to its use for SELEX experiments. Blind and Blank used a line diagram to show positional distribution of nucleotides, seen in Figure 6. To visualize k-mer distribution histograms can be used[24].

Single-Round SELEX Aptamer Detection Finding high-affinity aptamers early on is key in SELEX for several reasons. The more cycles are performed, the higher the overall costs of aptamer development, as SELEX is a labor-intensive and expensive process. Also, the more SELEX rounds are performed, the higher the chance that well replicating sequences outperform good binders[24]. Hoon et al., 2011, have shown that a single-round of SELEX can produce well binding aptamers for antithrombin. Usually one single step is not enough to show

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

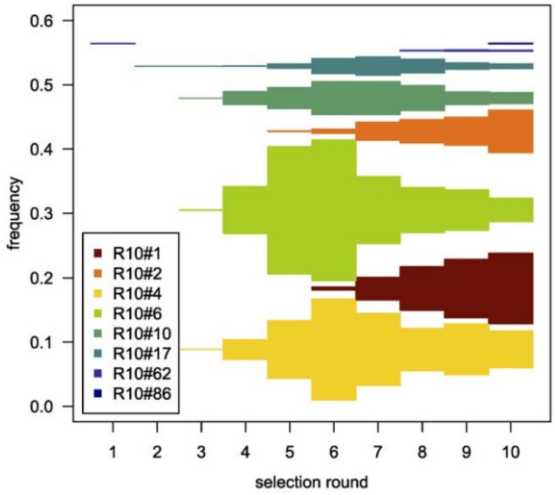


Fig. 7: Visualization of aptamer population tracing. They have taken the 100 highest frequency sequences per SELEX round and visualized their relative proportions. Figure adapted from Schütze et al., 2011[15].

significant enrichment on the full-length sequences, so they focused on the detection of enriched motifs[26]. Enriched motifs can thus be used as guide to find high-affinity aptamers, as long as starting library randomness is established[24].

Aptamer Population Tracing As the frequency and abundance of aptamers are expected to change over the course of many SELEX rounds, it may be helpful to track population dynamics. Aptamers are expected to get enriched due to their high-affinity for the target. However, especially in later rounds, PCR bias increases, and certain aptamer sequences may outperform others simply because they are easy to amplify but not necessarily good binders. Vice versa, aptamer sequences that are enriched later in the SELEX due to the increasing selection pressure may not be easily identified due to their relatively low abundance. Schütze et al., 2011, have thus visualized changing population sizes in a small scale, using only the top 100 aptamers per SELEX round, as seen in Figure 7.

Motif Detection and Clustering Clustering is the process of assigning sequences to sequence families by similarity. Having assigned families offers the advantage of being able to move from the single-aptamer level to whole families at once. Families of interest can be dissected in detail and be analysed without the noise of unrelated sequences.
An abundance of general clustering techniques is available, and many SELEX specific adaptions have been made. Computing similarities for every sequence pair is infeasible due to the vast amount of data per experiment and the resulting exponential time requirement needed for computation. Therefore, often times heuristics are applied to estimate distances. There are mainly two approaches for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

clustering HT-SELEX data. The first approach is based on clustering sequences one-by-one, i.e. a sequence is added to a cluster if the similarity is high enough to an existing cluster, otherwise a new cluster is created. The other approach is based on computing an extensive similarity graph beforehand and then clustering the graph as a whole by applying existing graph clustering techniques.

In SELEX, different types of data can be used for clustering. Most approaches use the whole random region, variable length subsequences, subsets of nucleotides or k-mers[27, 28, 29, 1]. Clustering can also be done on secondary structure information, computed using folding prediction software[30, 20, 31, 21]. Time-series dependent scoring of changes of the average secondary structures of aptamer families have proven useful for finding well suited binders as seen in AptaTrace[20]. Pei et al., 2017, have tried applying a kind of tertiary structure prediction to analyse binding potential of aptamers[32].

Mutational Dynamics Tracking Quang et al., 2018[25], have shown that tracking the mutational dynamics of aptamer families may lead to the discovery of better binders. For their experiments they performed cell-SELEX using MFC-7, a known mammalian breast cancer cell. They found that mutations observed in the data are mainly caused by the amplification step in every SELEX round, and that the choice of polymerase used heavily affects the overall mutational landscape. In contrast, sequencing errors had a comparatively negligible impact. Based upon these analyses, they inferred that errors during PCR led to the emergence of new better suited aptamers, but also that visualizing aptamer families using empirical genealogical evolutionary trees together with enrichment data can be a viable tool in tracking variant emergence and their fitness.

Another experiment performed by them built upon an identified aptamer family targeting a specific protein on the cell surface. They wanted to know whether they could find better aptamer variants by performing a doped cell-SELEX against the same target. To this end, a starting library was used that consisted of the identified aptamer and mutations of the previous step only. Four rounds of SELEX were done, but none of the newly identified variants outperformed the original aptamer. Nonetheless, NGS data revealed which parts of the original aptamer were crucial for binding[25].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.4. Overview of Current SELEX Data Analysis Tools

In the last decade many SELEX specific analysis tools have been developed. Most of these tools adapt conventional clustering techniques and apply them to data derived from sequenced SELEX nucleotide pools. These tools usually take innovative approaches on how to associate aptamers with each other, how clustering is done in general and how scoring takes place.
Finding the best tool for a given task is not trivial, especially for researchers new to the field of SELEX bioinformatics. Programs differ on multiple aspects.
Aspects to consider:

- Code quality and robustness
- Ongoing code maintenance
- Documentation
- Assumptions on experiment parameters like target, properties and process
- Algorithmic choices and performance
- Operating system dependence
- Format of results
- GUI vs CLI vs Libraries

In general, tools with command line interface (CLI) require the user to have a basic understanding of how to interact with a computer on the command line. Many tools also include a graphical user interface (GUI) which lowers the bar for inexperienced users. Whether to opt for CLI or GUI tools depends on the use case. GUI tools can be useful for users unfamiliar with the command line, one-time-only uses and interactive visualization. CLI tools are easily integrable in automated bioinformatic pipelines. CLI tools can also be used on remote cluster computers, which usually offer more computing resources and are accessible via command line using the SSH-protocol.
Regardless of the interface choice, researchers using these tools still have to understand the underlying principles, for proper result interpretation. On the following pages, programs used in SELEX bioinformatics are outlined. An overview of bioinformatic tools for SELEX data analysis is given in Table 1; adapted from Komarova et al., 2020[33].

1.4.1. Graphical User Interface Tools

AptaSUITE AptaSuite is a fully featured SELEX data analysis toolkit, in which AptaPLEX, AptaCLUSTER, AptaTRACE and AptaSIM are combined.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Software	Platform					Interface		SELEX specific	NGS
	Unix	DOS	Web	Galaxy	Lib	Graphical	Command Line		
FASTAptamer	X	X		X			X	X	
MEME	X	X	X	X		X	X		
STREME	X	X	X			X	X		X
MEMERIS	X						X	X	
MPBind	X						X	X	
AptaMotif	X						X	X	
APTANI	X	X					X	X	X
APTANI²	X	X				X	X	X	X
RaptRanker	X	X					X	X	X
AptaSUITE	X	X				X	X	X	X
AptaCluster	X	X					X	X	X
AptaTrace	X	X					X	X	X
RNAmotifAnalysis	X						X	X	X
AptCompare	X	X				X		X	X
SMART-Aptamer	X						X	X	X
Unoise	X	X					X		X
Uclust	X	X					X		X
DADA2					R				X
NCM					Clojure			X	X

Tbl. 1.: Examples of bioinformatic tools used in SELEX data analysis

The components of AptaSUITE are all specifically developed for the analysis of HT-SELEX data and are highly regarded in the field. AptaSuite is written in Java[34].

- AptaPLEX: Used to demultiplex and prepare sequencing files.
- AptaSIM: Used to simulate SELEX experiments for validation and benchmarking of analysis tools.
- AptaCLUSTER: HT-SELEX data analysis tool based on locality sensitive hashing.
- AptaTRACE: HT-SELEX data analysis tool based on secondary structure dynamics.

AptaCLUSTER AptaCLUSTER is included in AptaSUITE and is used for clustering aptamers based on their sequence.
For distance estimation of two aptamers locality sensitive hashing is used. The hash function produces the same result when two highly similar inputs are used, also referred to as collision. It is called multiple times on differently sampled data points of an aptamer. The resulting set of hash values is the hash ensemble of the aptamer. LSH ensembles are used as a fast to compute estimation on the upper bound of the distance between two aptamers. As seed for the first cluster

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the most abundant aptamer is used. Any aptamer having the same hash in their ensemble as the seed is considered a potential cluster candidate and added if their true distance is below a certain threshold[27].

AptaTRACE AptaTRACE is also included in AptaSUITE and is used for clustering aptamers based on secondary structure. For every SELEX round, and for every aptamer a secondary structure is predicted. For every k-mer a representation of the secondary structures associated with it is calculated per SELEX round. Linking the secondary structure representations of a k-mer over all SELEX rounds is called K-context. K-contexts can be utilized to determine the overall change of secondary structures associated with a k-mer during the SELEX experiment and also to exactly trace the dynamics of change in secondary structure. The resulting K-contexts are tested for significance against low-affinity k-mers and the k-mers of the initial library. The highest scoring k-mers are chosen for seeding and aptamers are added depending on k-mer alignment and structure[20].

AptCompare AptCompare is, in contrast to the other tools here, a combination of multiple SELEX-specific tools and methods. Included are the following aptamer analysis and motif discovery methods: sequence frequency analysis, FASTAptamer, MPBind, AptaCLUSTER, APTANI, RNAmotifAnalysis. AptCompare runs the tools sequentially in a single-threaded way and writes their results to the disk. In the end all results are combined in a table[35].

MEME-Suite The MEME-Suite is a toolbox developed for the task of motif discovery in sequence data sets and offers many tools for motif data analysis. The most prominent tool MEME is based on the Expectation Maximization-algorithm (short EM-algorithm) and has found application in SELEX bioinformatics. Due to the computational complexity of the EM-algorithm, it can only be used for smaller data sets (10^1 to 10^3) in feasible time.[36, 28].
In contrast, the tool STREME (Sensitive, thorough, rapid, enriched motif elicitation), replacing the tool DREME, is not based on the EM-algorithm. It uses a set of statistical test and iterates them until enough motifs have been found. STREME does not cover the whole motif search space, and starts with the most prominent ones based on k-mer frequencies. It is not as accurate as MEME, but much faster due to linear time scaling, and thus is not as susceptible to performance issues regarding data set size. A control file can be provided to perform differential motif search[37].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

As the MEME-Suite programs do not employ secondary structure information by default, MEME was adapted to also include secondary structure in a tool called MEMERIS (see 1.4.2)[30].

APTANI² APTANI and its successor APTANI² are based on AptaMotif (see 1.4.2). AptaMotif is due to extensive secondary structure calculation and multiple sequence alignments (MSA), which are both computationally expensive, not applicable to HT-SELEX data. APTANI takes multiple optimization steps on the AptaMotif approach, by limiting the covered search space. APTANI first runs a frequency filter on the data set, so only highly enriched aptamers are used for secondary structure calculation. Secondary structures are calculated within a defined energy range around the MFE structure. Further optimizations in the picking of structures for the MSA have been taken.
APTANI² then extends APTANI by adding a ranking scheme based on a combination of frequency and structural stability[38, 31].

1.4.2. Command Line Interface Tools

FASTAptamer FASTAptamer is one of the most prominent tools used for HT-SELEX data analysis. It is often used for dereplication of aptamers and enrichment calculation. FASTAptamer includes a clustering tool which is based on threshold-based clustering. The clustering is based on Levenshtein distance and adds aptamers to a cluster if their distances are below a certain threshold. For fast computation the user should define a minimum frequency at which aptamers are considered, as Levenshtein distance calculation is computationally expensive[29].

AptaMotif AptaMotif was developed by Hoinka et al., 2012, and inspired many advances of other SELEX data analysis programs. AptaMotif is designed for secondary-structure based aptamer clustering. It is not part of the AptaSuite. Clustering seeds are selected by performing multiple sequence alignments (MSA) of aptamer secondary structures. First folding prediction is done for all aptamers. For every aptamer MEA, MFE and a set of suboptimal structures are calculated, called the structure ensemble. The structures are annotated and put in a database.
A number of random aptamers is sampled from the database. For every sampled aptamer one structure is drawn from their structure ensemble. MSA is performed on the structures. This is done multiple times.
The MSAs are scored and used as clustering seeds. Then aptamers are added to these seeds when the difference is small enough[21].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

MPBind MPBind was developed in 2014 by Jiang et al. It scores aptamers based on k-mer enrichment. MPBind can be used to compare two data sets for motif enrichment in aptamers. The algorithms used in MPBind would allow for fast computation and scoring, however, the implementation of MPBind is not optimized for large data sets.
In both data sets the aptamers are dereplicated and the k-mers are counted. Fisher's exact tests and one-sided Spearman correlation tests are used to determine enrichment significance values. Based on these statistical tests, a combined enrichment score is calculated for every k-mer. For every aptamer an averaged score is then calculated, using these combined k-mer scores, depending on the included k-mers. Aptamers which contain k-mers that have been enriched from one round to another will presumably have a higher score than others[39].

SMART-Aptamer SMART-Aptamer extends the approach used in MPBind and can be used for HT-SELEX data sets. It scores aptamers based on a combination of three scores. The output is a table with aptamers for which the two maximum scores are summed and used as ranking criteria.
One score is a k-mer enrichment score similar to the one used in MPBind. The second score is a family size score, based on the size of the cluster the aptamer is in. Clustering is based on graph clustering using MCL, run on a similarity graph created from BLAST[40]. The third score considers G-Quadruplex structures and overall secondary structure stability[41].

MEMERIS MEMERIS is an extension of the original MEME algorithm and implements secondary structure prediction. It computes MFE structures beforehand and only considers unpaired subsequences (looping regions) as starting positions for the algorithm.
However, just like the original MEME algorithm, it can only handle smaller data sets due to its computational complexity[30].

Unoise Unoise is a denoiser made for amplicon sequencing. It essentially is an error-correction tool for data sets acquired from Illumina sequencers and has been used to reliably dereplicate highly-similar sequences[42]. It is included in the bioinformatic toolkit Usearch.
Allnutt et al. have successfully used Unoise to perform data analysis on SELEX data sets[1]. An interesting benchmark comparing the denoisers has been done by Nearing et al., 2018[43].

Uclust Uclust is a clustering tool made for amplicon sequencing tasks is also included in Usearch[44]. Allnutt et al. have shown that Unoise and Uclust (and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

DADA2 as well) performed very well compared to FASTAptamer and AptaTrace, and with some benchmarking data sets even outperformed them[1].

RNAmotifAnalysis Ditzler et al., have developed an analysis program (and Perl library) that can be used for clustering of aptamers from HT-SELEX experiments based on secondary structure. Clustering in RNAmotifAnalysis is done in a multi-step repetitive fashion.
First, the sequences are clustered in an iterative threshold-based clustering process. The clusters are aligned in an MSA process using MAFFT. These alignments are used as input for RNAalifold from the Vienna RNA package for prediction of base pairing probabilities and to score how well these sequence alignments align structurally. Then covariance models (CMs) are created for every cluster. A subset of the CMs is then refined in an iterative search and refinement process against the aptamer population. When the change in the CMs is only marginal a new subset of CMs is used for refinement. In the overall analysis process some constraints, as limitations on the maximum amount of sequences or clusters, have been taken to ensure fast computation[45].

RaptRanker RaptRanker is an HT-SELEX analysis tool based on threshold-based clustering. Contrasting other methods, RaptRanker clusters subsequences based on their nucleotide composition and the secondary structure they take in their full-length aptamers. Subsequences and their secondary structures are stored as profiles, which are then put in relation using a fast multidimensional sorting algorithm. The result is an undirected (disconnected) graph. Clustering is done, by finding all minimum spanning trees (MSTs) in the graph. Aptamers are scored based on the average enrichment of the motif clusters included in them, similar to the approach used in MPBind[46].

1.4.3. Libraries

Some SELEX analysis tools have been released as libraries instead of stand-alone programs.

DADA2 DADA2 is a library developed for amplicon sequencing tasks. It performs error-correction and amplicon inference. Using the sequenced data sets DADA2 learns an error model of the sequencer and the sequencing run. The error model is used to perform error-correction and infer the real sequences. DADA2 finds ASVs, short for Amplicon Sequencing Variants which are representative sequences for a group[47]. It is a package developed for R.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

HTS-Exploration using NCM Pei et al., considered nucleotide cyclic motifs (NCMs) in their analysis of HT-SELEX experiments. NCMs have been found to complement tertiary structure prediction in RNA. In their approach a model is built based on enrichment/depletion of NCMs. This model can then be used to find promising aptamers expected to bind with high affinity. In benchmarks against AptaTrace and RCK their method has performed considerably well. Correlations between NCM enrichment/depletion and binding affinities have been shown.
They provide a library implemented in Clojure[32].

1.5. Identification of Aptamers against Bacteria

In the last decade, next generation sequencing has changed SELEX substantially, enabling the detection of potentially better binding aptamers in the low numbers, compared to high-abundance sequences. Even though NGS has arrived in the field, Sanger sequencing is still the dominant sequencing technique for bacteria-targeted SELEX experiments[48, 49, 50, 18, 51, 52, 53, 54]. Only few groups have performed bacteria-targeted SELEX with NGS[55, 56, 57].
In whole-cell SELEX, enriched sequences may bind to unspecific epitopes of the targeted cell, leading to aptamers which may be sensible to the targeted cells, but not specific to them.
Therefore, some groups have treated the task using a differential approach to increase confidence in expected sequence specificity. Meyer S. et al., 2013[58], were the first to use a differential approach, using engineered mammalian cells over-expressing a known membrane protein, as target and unmodified cells for counter-selection. They sequenced the selection and counter-selection of the last SELEX round using NGS, and chose sequences based on their comparative abundance ratios. Pleiko et al., 2019[59], have gone a similar way, using cancerous renal cells as selection targets, and healthy renal cells for counter-selection. They treated their SELEX experiment as an RNA-seq problem, using conventional RNA-seq tools for aptamer identification.
Beside these two, other groups have used NGS for whole-cell SELEX as well[60, 57, 61, 55], most of them selecting sequences based on enrichment, abundance, secondary structure, or on clusterings.
While many whole-cell SELEX protocols have been developed[62], the development of cell-SELEX specific analysis methods and software tools is still lagging behind. Analysis tools created for SELEX usually cover a set of different use cases, i.e. identification of well-replicating aptamers or clustering of aptamer families, and differ on computational complexity, code quality and applied analysis methods. Most tools have been developed for conventional single-target

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

SELEX, and therefore not all of them are suitable for every type of SELEX experiment. Additionally, the way experiments are analysed by these tools vary greatly, so interpretation and validation of results is difficult. Tools are usually validated by correlating results of sequence ranking and experimentally determined binding affinities[15]. As the SELEX procedure is not standardized, choosing the right tools and methods strongly depends on experiment structure, dataset sizes and the research questions.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

2. Aim of Thesis

The aim of this thesis was to analyse NGS data sets derived from in-house performed whole-cell SELEX experiments against bacterial cells. The goal was to set up bioinformatic pipelines that enable to:

1. Pre-process and prepare raw NGS data sets for analysis
2. Analyse the quality of the starting ssDNA library
3. Assess the SELEX procedure and enriched ssDNA pools
4. Identify and extract aptamer candidate sequences for downstream experimental screening and testing.

NGS data sets to be analysed originated from three different whole-cell SELEX experiments:

- SELEX-EF01: whole-cell SELEX over 9 consecutive rounds; target: *Enterococcus faecalis*
- SELEX-EF05: whole-cell SELEX over 11 consecutive rounds; target: *E. faecalis*
- SELEX-EF07: whole-cell Toggle-SELEX over 9 consecutive rounds; toggle-target-order: *E. faecalis*, *E. faecium*, *E. durans*, *E. hirae*

In addition, attempts were made to perform secondary-structure-based clustering and motif detection and k-mer-based sequence ranking.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3. Methods

In this project, ssDNA pools from multiple SELEX experiments against target bacteria (*Enterococcus spp.*) were pooled and sequenced with an Illumina MiSeq platform. Demultiplexing was done automatically by the Illumina MiSeq sequencer.

3.1. Pipeline Development

The bioinformatic pipelines were developed using the workflow manager Nextflow 20.10.0[63].
Custom data manipulation and analysis tasks were done using a set of different scripting languages and tools. Required libraries and software were made available to the workflows using the conda channels bioconda, conda-forge and the default channel.
Python 3.9 was used with the libraries scipy[64], biopython[65], pandas[66] and networkx[67].
R 3.6[68] was used with the libraries dplyr[69] and tidyr[70] for data wrangling, ggplot2[71] for visualization, xlsx[72] for creating Excel tables and RMarkdown[73] for creating HTML-files. If not otherwise specified, ggplot2 was used for all visualization tasks.
Common tools available in the POSIX environment were used for low-level tasks, such as converting from FASTQ to FASTA or extracting data from tables.

3.2. Preprocessing and Quality Assessment: *selex-ngs-prep*

The workflow *selex-ngs-prep* performs data preparation and quality assessment of raw FASTQ-files. It trims off flanking regions (forward and reverse primers) and discards sequences missing primers (artifacts). It then filters the sequences by quality, merges paired-end reads and discards sequences which are out of the designated length range. The workflow returns preprocessed files in FASTA and FASTQ format. Info on the sequencing quality is given for the input and output files in the form of plots and tables.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Data Preparation Input sequences are expected to consist of a forward primer, random region, and reverse primer.
The first step of data preparation is the trimming of forward and reverse primer. *Cutadapt* 3.3[74] is used for this task. Forward and reverse primers are provided in an anchored, linked form ('`^PRIMER1...PRIMER2`'), enabling global alignment for the forward primer and semi-global alignment for the reverse primer. Likewise, for the reverse complement file, primers are provided in reverse complemented form, also anchored and linked. Cutadapt is configured to detect primers in an error-tolerant way allowing mismatches. Quality filtering is done using the tool *fastp* v0.20.1[75]. Fastp is configured to keep sequence pairs in which both reads have a high enough average phred quality. It is used as well for paired-end read merging. It is configured to do error-correction and allows a settable number of mismatches. A custom-made Python script is used to check every sequence for the correct length, discarding any sequence out of the defined bounds.
The resulting FASTQ-files are then converted to FASTA using the POSIX tool 'sed'. FASTA-files produced in the workflow consist of two lines per sequence. The number of remaining sequences after every step is gathered in tabular csv-files, using the POSIX tool 'wc' for all FASTA- and FASTQ- files. Line counts are divided by 2 (4 for FASTQ respectively) using bash math mode. Two plots are created with a custom R-script, visualizing loss during preprocessing with stacked bar charts, scaled to 100% and unscaled in absolute numbers.

Sequencing Quality Assessment The sequencing quality assessment script, written in R, creates sequencing quality plots using the 'plotQualityProfiles' function provided in *DADA2*[47]. The script works using both, paired-end sequencing files and single-read sequencing files. All forward (respectively reverse) FASTQ-files are summarized and have their overall quality plotted, saved to PNG files and embedded into an HTML-file using RMarkdown[73]. The RMarkdown Render function is called for every single SELEX round separately, embedding their quality profile plots as well and saving them as PNG.

3.3. SELEX Assessment: *selex-assess*

The workflow *selex-assess* is used to analyse sequencing data from bacterial whole-cell SELEX experiments. It dereplicates sequences, returns their counts and scales the number of reads to rpm (reads per million). It ranks and sorts sequences by decreasing abundance. Nucleotide composition analyses are done based on the 40-nucleotide random regions of the ssDNA sequences. Global assessment of sequence diversity, tracing of population dynamics and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

identification of putative aptamer candidates based on sequence abundances is done.

Sequence Dereplication To dereplicate sequences a custom Python script was written. The script takes a FASTA-file for every SELEX round and goes through them sequentially. For every SELEX round a dictionary is created and the sequences are scanned top to bottom. When a new sequence is encountered, it is added to the dictionary and the counter is set to 1. If the sequence is observed again, the counter is increased by 1. The dictionaries of the SELEX rounds are then combined to a DataFrame and written to a tabular csv-file. A FASTA-file containing all dereplicated sequences is created, for which the identifier of every sequence consists of the sequence itself and the sequence counts, separated by the character '-'.

Recovering the Top Sequences A custom R script is used to extract the top n sequences of every SELEX round. The script takes the csv-file created in the dereplication step and uses the library xlsx[72] to create a tabular xlsx-file. The script recovers the top n sequences of every SELEX round, orders them, and fills one sheet per round with sequences, their absolute counts and their counts scaled to rpm (reads-per-million) using the formula $x_{j;RPM} = \frac{x_j}{\sum_{i=1}^{n} x_i} * 10^6$.

Nucleotides Composition Assessment A custom Python script (Figure 8a) is used to count the nucleotides encountered per position in the SELEX rounds. The counting script can be configured to be used with RNA, DNA, proteins or custom alphabets. A dictionary of length 'random region length' is created beforehand with nested dictionaries for the nucleotides set to 0. Every aptamer is scanned from left to right and the corresponding counters are increased by one. By default, the script prints to the console in csv-tabular form. It can also output counts as percentages per position as well as the overall share every nucleotide takes. The script is called once per SELEX round.
For every SELEX round plots are created consisting of vertical stacked bars along the x-axis for every position in the sequences, using a custom R script, Another plot is created consisting of vertical stacked bars along the x-asis for every SELEX round, using another R script. The plots are written to PNG files and embedded in HTML-files using RMarkdown.

Enrichment Analysis A custom R script (Figure 8b) is used to assign sequences into bins. As input the script takes a tabular csv-file from the dereplication step and a logarithmic base as input. The script first reads the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

csv-file and puts it in memory. For every sequence an exponent is found, by logarithmizing the count and rounding to the next lower integer such as: $exp = \lfloor log(count) \rfloor$. The number of sequences per bin and SELEX round are then summed up. This table is written to two tabular csv-files, one containing the summed total count of a bin per SELEX round, and one containing the number of unique reads of a bin per SELEX round.
Another R script then visualizes these tables, plotting one vertical stacked bar per SELEX round, visualizing the share every bin takes per SELEX round.

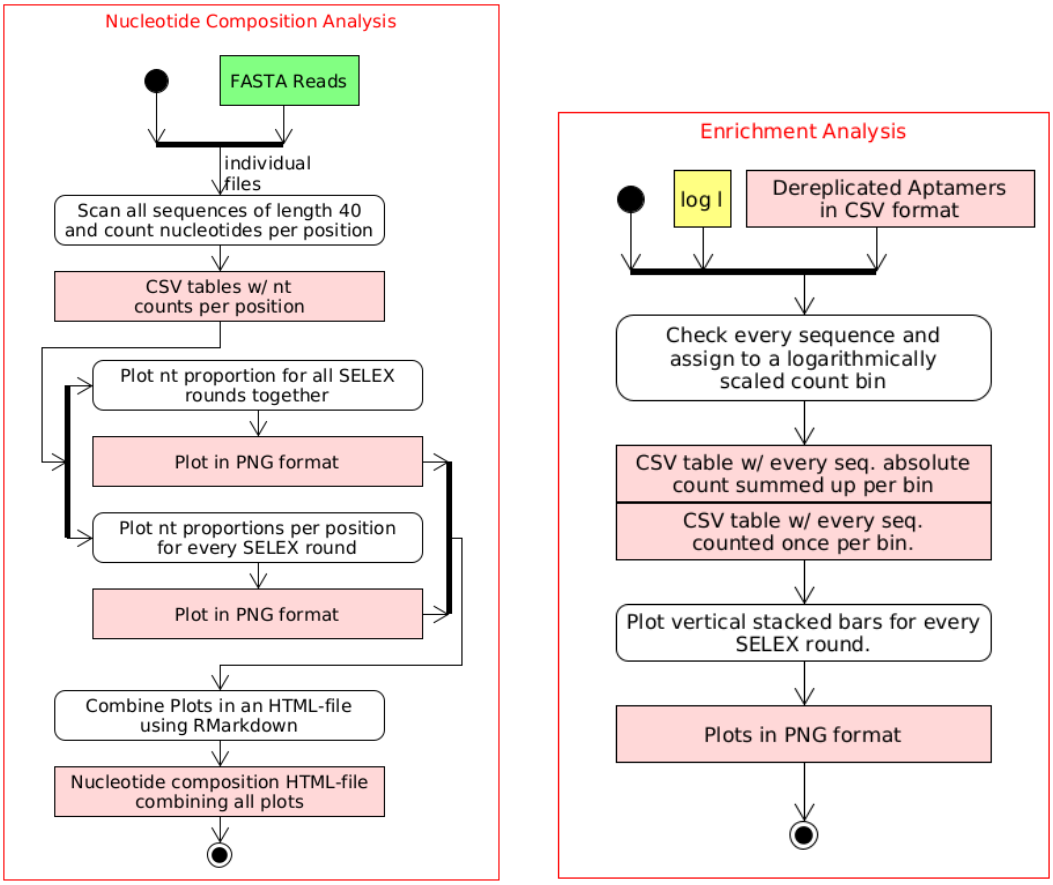
3.4. K-mer-based Aptamer Scoring: selex-kmer

The workflow *selex-kmer* adapts a k-mer-based scoring approach used in MPBind from Jiang et al., 2014[39]. First, the workflow determines the representative sequence for every mutational family. For every SELEX round all k-mers are counted and put into a table. Then one-tailed Fisher Exact tests are done for every combination of SELEX rounds for every k-mer. The results of this step are then used to calculate scores for sequences.

Determining the Representative Sequence for Mutational Families For every mutational family cluster a representative sequence is chosen. *BLAST* 2.12.0[76] is used to find highly-similar sequence clusters. First, a BLAST database is created using the dereplicated sequences. To query the database multiple single-core BLAST searches with 1000 sequences per query are launched in parallel. A cut-off of 0.05 is set for the E-value. The results of the BLAST searches are then combined using the POSIX tool 'cat' and parsed into a graph using a custom-made Python script, using the *networkx*[67] library. The graph consists of a number of disconnected sub-graphs, called components, containing sequences which are similar to each other. The script uses the *networkx* library to find all disconnected components and chooses the sequence with the highest total read count as representative. Representative sequences are then written to a FASTA-file.

K-mer Counting K-mers are counted using a custom-made Python script. The script takes as input the k-mer size k, a SELEX round FASTA-file, and the FASTA-file containing all representative sequences. An array covering the whole k-mer space (4^k) is initialized with zeroes. Sequences of the SELEX round file are considered for k-mer counting, if they can be found in the representatives file. Sequences are scanned from left to right with a frame size of k to get all k-mers. For every k-mer the corresponding entry in the k-mer space array is incremented

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Scheme of the selex-assess nucleotide composition analysis step.

(b) Scheme of the selex-assess enrichment analysis step.

Fig. 8: Schemes of the actions performed in *selex-assess*. Shown actions extend the UML-diagram of *selex-assess* as seen in Figure 11 in the results section (4.1.2). Green rectangles represent sequence files (FASTA, FASTQ), pink rectangles represent plots and tabular files, and round edge rectangles represent analysis steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

by one. K-mers are only counted once per sequence. The k-mer count array is then written to a tabular csv-file.

Fisher Exact Testing Enrichment testing of k-mers is done using a custom-made Python script that performs Fisher Exact tests, using the *fisher_test* method as provided in the *scipy.stats* library[64]. The script takes the k-mer count files of two SELEX rounds and the length k. First, the k-mers are loaded into memory. Then the script runs through a loop for every k-mer. A contingency table is made for every k-mer which is used as input for the *fisher_test* function. The function has the alternative hypothesis set to 'greater'. The resulting p-value is z-transformed using the percent point function *scipy.stats.norm.ppf*. The transformed p-value (called the z-value) is then multiplied by -1. The z-values are limited to ± 20.
The output consists of the k-mer, the absolute counts of it, the p-value and the negated z-value.

Aptamer Scoring Aptamer scoring is done using a custom-made Python script. The script takes a FASTA-file containing all random regions and the output of the Fisher exact testing step.
The k-mer enrichment table is read and put into memory using a dictionary, as well as all distinct random region sequences. Then every sequence is scanned from left to right. All k-mers encountered in the sequence are put into a dictionary and the corresponding k-mer enrichment scores are assigned to them. The sequence scores are then calculated by averaging all k-mer scores in the dictionary.
Shifting scores are calculated for a short section of 5 k-mers. The shifting scores are initialized as: $s_{min} = +\infty$ and $s_{max} = -\infty$. Just as with the overall aptamer score, sequences are scanned from left to right. For every position i, 5 consecutive kmers are extracted and the score s_i is calculated. Then s_i is compared to s_{max}. If it is greater than the current value, s_{max} is assigned the value of s_i (and vice versa for s_{min}).
As output sequences, the compared SELEX round numbers and the associated scores are printed to the shell and written to a tabular csv-file.

3.5. Secondary Structure-Based Motif Detection: selex-blaster

The workflow *selex-blaster* was designed to cluster sequences based on their predicted secondary structure and find single-stranded motifs in these clusters.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The workflow starts with a dereplication step, as described in the *selex-assess* workflow. The representative sequence for every mutational family is determined, as described in the *selex-kmer* workflow. Sequences are folded and used to create a simlarity matrix using masked BLAST. Then the similarity matrix is clustered using MCL. The resulting clusters are scanned for motifs using MEME.

Secondary Structure-Based BLAST The secondary structure-based pairwise alignment (Figure 9) works as follows. Folding prediction is done using *RNAfold* 2.4.17 included in the toolkit ViennaRNA[3], using the packed ssDNA folding model by Mathews et al.,2004[77]. For this the forward and reverse primer are attached to the random regions using the POSIX tool 'sed'. Secondary MFE structures are used for masking, which are extracted using 'awk' and put into a tabular csv-file. Masking is performed using a custom Python script. The script reads the csv-table sequence-wise and checks for every nucleotide in the sequence whether it is bound in the secondary structure. Bound nucleotides are printed in lowercase and unbound nucleotides are printed in uppercase. Sequences are also hard masked using the character 'N' instead of lowercase.
BLAST 2.12.0[76] is used in 'blastn-short' mode with reward and penalty values '`-reward 1 -penalty -4 -gapopen 1 -gapextend 2`', as these seemed to work reasonably well. A masked BLAST database is created using the lowercase-masked FASTA-file. A cut-off of 20000 is used for the E-value to include lower-significance hits as well. Multiple single-core BLAST queries are executed in parallel using small chunks of the N-masked FASTA-file. The search results are then combined into one tabular csv-file using the POSIX tool 'cat'. Query sequence, hit sequence and alignment length are extracted using the POSIX tool 'cut'. Self-hits are excluded by scanning every hit and comparing query and hit sequence, using a custom Python script.

Graph Clustering The resulting similarity graph is clustered using *MCL* 14.137[40], short for Markov Clustering Algorithm. By default an inflation factor of 1.4 is used. A FASTA-file is created for every cluster using a custom Python script.

Motif Detection *MEME* 5.3.0[36, 28] is used to detect motifs in the cluster files. MEME is configured to find a maximum of 3 motifs with a minimal length of 4. Motif detection on the reverse-complement strand is disabled. Motif candidates are compared to a randomized model by MEME. The result consists of one motif file in HTML/MEME-format per cluster, which can be opened using a current web browser or used for further analysis using tools from the MEME-Suite.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

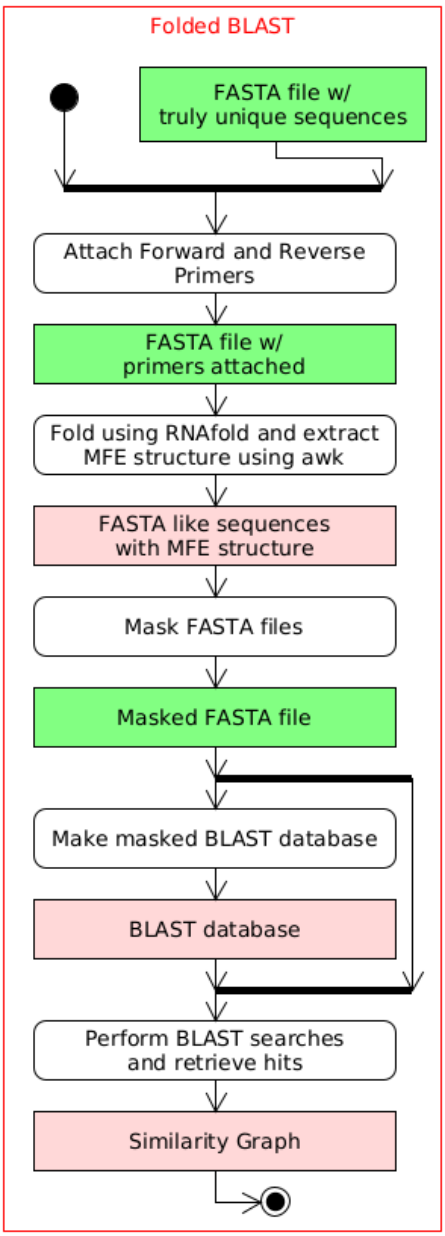


Fig. 9: Scheme of the secondary structure-based pairwise alignment procedure in the workflow *selex-blaster*. The shown action extends the UML-diagram of *selex-blaster* as seen in Figure 13 in the results section (4.1.4). Green rectangles represent sequence files (FASTA, FASTQ), pink rectangles represent plots and tabular files, and round edge rectangles represent analysis steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4. Results and Discussion

The results and discussions part of this thesis is divided into two sections. First, the bioinformatic analysis pipelines developed during this master project are presented and discussed in detail. In section two, the results from data analyses of three different SELEX experiments (EF01, EF05, and EF07) employing the developed pipelines are shown and discussed. Related wetlab work (SELEX experiments and NGS sequencing) was conducted by Claudia Kolm and Isabella Cervenka[14, 78].

4.1. Bioinformatic Analysis Pipelines

In this thesis project, four bioinformatic analysis pipelines were developed for the analysis of in-house performed bacterial whole-cell SELEX experiments, namely i) selex-ngs-prep, ii) selex-assess, iii) selex-kmer and iv) selex-blaster.
In the following subsections, the steps the workflows take during execution are outlined including thresholds and parameters which are expressed in *italics*.
With every pipeline a 'nextflow.config' file is provided, in which default values are stored, as well as a custom-made configuration wizard script, that can be used to generate experiment specific configuration files. Thresholds and other parameters described are default values and specific to the SELEX experiments conducted in 2019. Schemes of the workflows are presented as UML diagrams.

Data Availability Workflows are hosted on GitHub:

- *selex-ngs-prep*: `https://github.com/hovercat/selex-ngs-prep`
- *selex-assess*: `https://github.com/hovercat/selex-assess`
- *selex-kmer*: `https://github.com/hovercat/selex-kmer`
- *selex-blaster*: `https://github.com/hovercat/selex-blaster`

4.1.1. Preprocessing and Quality Assessment: *selex-ngs-prep*

The workflow *selex-ngs-prep* (Figure 10) handles the data preparation and quality assessment of next generation sequencing files from whole-cell bacterial

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

SELEX experiments, which were sequenced on an Illumina MiSeq platform. The pipeline works with demultiplexed paired-end files in FASTQ format.

Selex-ngs-prep expects sequences to consist of random regions with two fixed adapters (SELEX forward and reverse primers) attached, one on each end.

The resulting sequence files can be used as input for other tools or for the SELEX analysis pipelines presented in this thesis.

Trimming The primer regions are removed before progressing to further data analysis stages, as for most analyses the random region sequences were sufficient. Due to the heavy amplification over the course of the SELEX experiment, adapters and random regions can be subject to mutations, insertions or deletions, trimming the adapters rigidly was not an option. Therefore, an error-tolerant adapter scanning approach was used. All sequences starting with a forward primer, containing a variable length random region, and a reverse primer are considered valid and have their primers cut. Sequences not meeting these requirements are discarded.

Trimming is done using cutadapt, which is a versatile adapter trimmer offering great customizability. Cutadapt is configured to specifically look for the *forward* and *reverse primers* in the sequences, performing global alignment for the forward primer and semi-global alignment for the reverse primer. A default *error-threshold* of 20% was chosen for primer regions to be successfully recognized.

Quality Filtering Sequences stemming from SELEX can only be used if the full random region is of a high enough quality. Any sequence with an average quality score below a certain *minimumg average phred quality* should be discarded, which was set to 30 by default. The tool fastp[75] was used for this task.

Paired-End Read Merging Sequences from SELEX experiments are relatively short, e.g. 86 nucleotides (40nt random region and 23nt primer regions), so single-read sequencing is sufficient. Paired-end sequencing was performed as the technology was available. As for every strand two redundant sequences exist, paired-end read merging served as a quality increasing step. In the paired end-read merging step, every sequence pair is merged into a single sequence, as long as both strands are sufficiently similar. By default, the *maximum numbers of mismatches* is set to allow 1 mismatch. In hindsight, the maximum number of mismatches could have been set higher, as the used tool (fastp) corrects mismatches. Error correction can be disabled using the *base correction flag*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

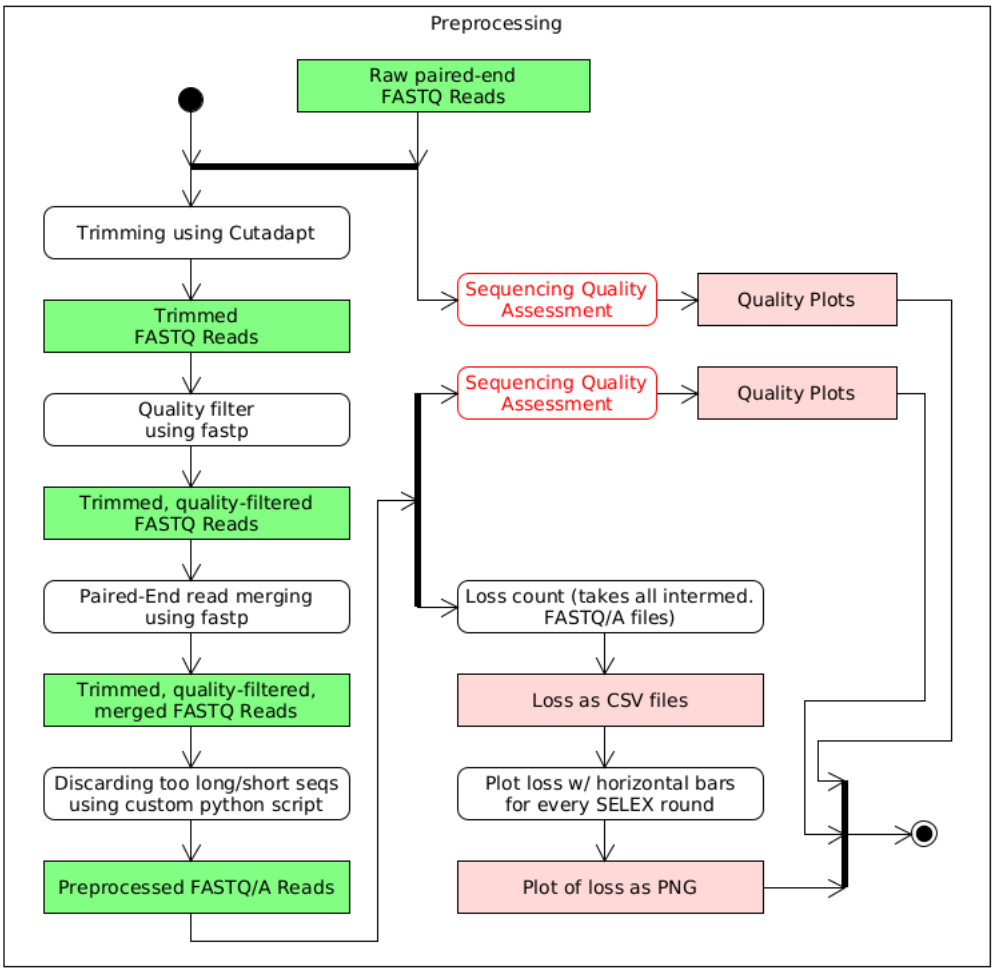


Fig. 10: Scheme of the *selex-ngs-prep* workflow as an UML diagram. Green rectangles represent sequence files (FASTA, FASTQ), pink rectangles represent plots and tabular files, and round edge rectangles represent analysis steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Random Region Length Restriction As sequences undergo insertions and deletions during SELEX they may differ significantly in length. Thus, only reads with a random region of length 40 ± 3nt were included by default. In the configuration file the expected *exact length*, the *minimum length* and the *maximum length* can be set. Accidentally litigated forward and reverse primers were encountered in SELEX experiment EF07 (4.2.3), which were discarded at this step.

Preprocessing Loss Assessment Over the course of the preparation procedure a number of reads are lost, which is visualized with horizontal stacked bar charts. One plot is used to show the exact share of loss at every preprocessing for every SELEX round. Another plot shows whether there are differences between SELEX round data set sizes, as it uses unscaled numbers. Optimally, all data sets should have similar size.

Sequencing Quality Assessment Sequencing quality profiles for every SELEX round are plotted using the function 'plotQualityProfiles' provided in the DADA2[47] toolkit. Sequencing quality is assessed for the unprocessed paired-end reads, as well as for the fully preprocessed reads, to allow for visual comparison. The plots are saved to PNG-files and embedded into an HTML-file using RMarkdown.
The workflow executes these steps in sequential order. In hindsight, a sequential approach may hinder detection of error sources as sequences may get discarded early on. For example, the adapter trimming step discards all sequences which have no primers attached. However, if the sequence is of too low quality, it is discarded as well. It would be better to do the steps individually and then overlap the resulting data sets.

4.1.2. SELEX Assessment: *selex-assess*

The workflow *selex-assess* (Figure 11) was developed to analyse sequencing data from in house performed bacterial whole-cell SELEX experiments by determining read counts (sequence frequency), normalizing them to the total number of reads in the population (reads per million), assessing nucleotide composition, ranking reads and determining singleton to duplicate ratios. Overall, it allows for a global assessment of sequence diversity, tracing of population dynamics and identification of putative aptamer candidates based on sequence abundances. Another important asset to check was that the starting ssDNA library is unbiased, as the whole experiment depends on it. The library should consist of singleton sequences only and be unbiased regarding nucleotide distribution.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

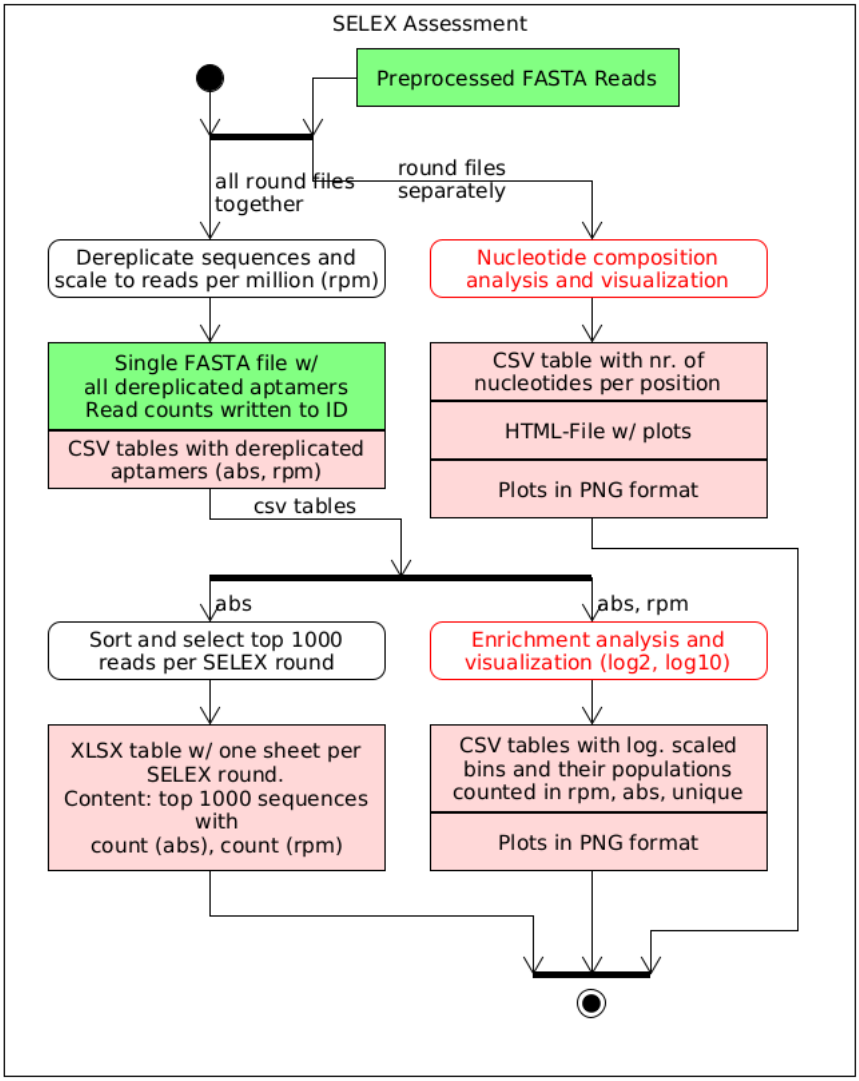


Fig. 11: Scheme of the *selex-assess* workflow as an UML diagram. Green rectangles represent sequence files (FASTA, FASTQ), pink rectangles represent plots and tabular files, and round edge rectangles represent analysis steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

***Top n* Aptamer List** At the start of HT-SELEX, aptamers mainly were chosen based on the highest read counts. The same procedure is used, when aptamer DNA pools are sequenced via Sanger sequencing. This method is still used today to some extent, though it is limited, as the most frequent aptamer sequences may not necessarily be the best performing binders[59].

The workflow returns the *top n* reads (by default 1000) for every SELEX round and puts them into an xlsx-table. Xlsx-format allows for multiple sheets in one file, so all SELEX round can be put into one file. For every SELEX round, a sheet is created containing the *top n* sequences, absolute counts and counts scaled to RPM (reads-per-million). RPM counts were useful to show the share every aptamer takes per SELEX round, and to calculate enrichment between two rounds. Every sheet in the xlsx-table is ordered in descending order.

Nucleotides Composition Assessment A scheme of the 'Nucleotide composition analysis and visualization' step is depicted in Figure 8a in 3.3.

The share every nucleotide takes along the random region per position can be used as an indicator to check for library randomness. Typically, the ssDNA library is expected to have an equal distribution of nucleotides at every position. The nucleotide distribution is visualized using stacked bar charts with one bar for every position in the random region. Additionally, nucleotide distribution is visualized for all SELEX rounds separately to show the changes per position.

Enrichment Analysis The enrichment analysis step (see Figure 8b in 3.3) serves as an analysis tool to determine whether sequences got enriched. Successful SELEX processes can be described by observing sequence enrichment over the SELEX process in combination with increased ssDNA pool binding affinity. Optimally, one could expect that with every SELEX round the heterogeneity of the data set decreases and the read counts of binding aptamers increases. Due to the possibility of PCR artifacts (inserts, deletes or mutations) it is possible that sequence variants may emerge, especially when distinct sequences are already dominating the pool population[79]. If SELEX pressure is properly applied, sequence variants will go 'extinct' and no new families except for spontaneous mutations will be added.

To show an increase in sequence duplication, logarithmic binning is applied, as during SELEX, sequence counts may be subject to exponential growth. Bins are spaced logarithmically (base 2 and 10) to make exponential growth perceivable. $\log_2$ was sufficient to show small increases, while $\log_{10}$ showed to be useful for larger steps. This step was also useful to check unbiasedness of the ssDNA library, as it optimally should contain singleton sequences only.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

SELEX rounds (and ssDNA library) are compared by plotting a stacked bar for every SELEX round on the x-axis.
It was observed that plots based on rpm-scaled counts were not useful, as comparatively smaller data sets got scaled much more than others, and therefore seemed to be more enriched. Subsampling should have been done before binning, as the input data sets were not always of equal size and therefore plots were slightly skewed.

4.1.3. K-mer-based Aptamer Scoring: selex-kmer

The workflow *selex-kmer* (Figure 12) adapts a k-mer-based scoring approach used in MPBind from Jiang et al., 2014[39]. A k-mer is a subsequence of length k. For instance, the sequence `GTTCAT` consists of the 4-mers `GTTC`, `TTCA` and `TCAT`. A length of 6 was chosen for k by default. The *selex-kmer* workflow produces scoring tables for every possible combination of SELEX rounds, consisting of an overall k-mer enrichment score, a maximum shifting score and a minimum shifting score for every sequence.
Jiang et al.[39], used k-mer enrichment scores to estimate binding potential of sequences, based on how strongly k-mers were enriched from round to round. The ground principle of k-mer enrichment based scoring lies in the assumption that only small sections of the aptamer nucleotide sequences are binding specifically to the target structures[19]. The aptamers are forming three-dimensional structures, partially folding to bind with themselves, minimizing free energy. Sections of an aptamer which are not bound to itself and therefore exposed to the outside, are commonly called looping regions. If a target is available for binding, the loop is expected to bind to the target[19]. Aptamers with looping regions which are sequence-wise similar should bind to the same target structure and therefore get replicated more often, even if otherwise not related. Thus, an enrichment of k-mers in the looping region seems plausible.
Unfortunately, MPBind v2.1 (written in Python 2.4.3) did not work as expected without rewriting parts of it. A short examination of the code showed that instead of importing custom scripts as libraries, MPBind executes these subsequent scripts by spawning new processes using the 'python' alias with the 'commands' library. As 'python' is the alias for the newest installed version of python by default, MPBind executed the scripts, which are written in Python 2, using Python 3, even if the program was started using an installation of Python 2. Besides that, even though the employed algorithm scales linearly, MPBind did not finish with the data sets of the SELEX experiments in a sufficient time. The reason for that may lie in excessive spawning of subprocesses. Therefore, parts of their method were adapted with performance in mind.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

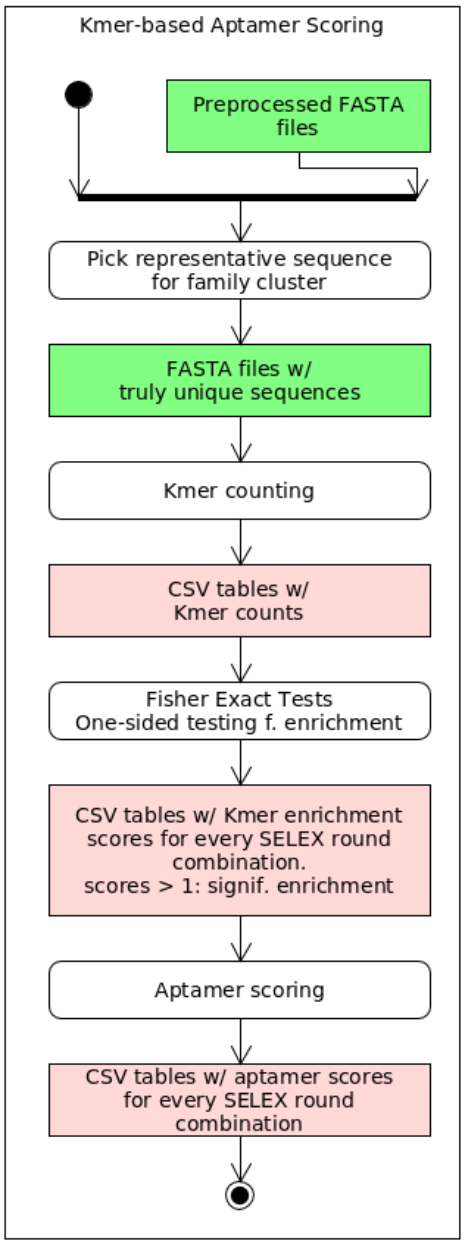


Fig. 12: Scheme of the *selex-kmer* workflow as an UML diagram. Green rectangles represent sequence files (FASTA, FASTQ), pink rectangles represent plots and tabular files, and round edge rectangles represent analysis steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

It was also hypothesized that scores could be used to estimate the binding affinity of sequences when used differentially i.e. by comparing scores for selection and counter-selection data sets.

Working Principle First, to avoid bias of overexpressed k-mers due to mutated sequence, a representative sequence for every mutational family is chosen. Then, k-mers are counted for every SELEX round. Every k-mer is only counted once per sequence to avoid bias due to long single nucleotide stretches, as seen in sequences such as `"GTTTCGGGGGGGGGGGGGAACACATTTGTGTAACAAACAGTC"` (found in *EF07*). A one-sided Fisher Exact test is performed for every k-mer for every combination of SELEX rounds to estimate enrichment significance. The resulting p-value is z-transformed and multiplied by -1, so enriched k-mers would have a positive score. To avoid overly strong influence of outlier k-mers, z-values were limited to ± 20.

K-mer based sequence scores are then calculated for every sequence by summing up the enrichment scores of every included unique k-mer and then taking the average.

Special interest was put into recognizing short regions of sequences which may show exceptionally high or low scores, by calculating scores for these regions. Theses scores are referred to as shifting scores and are calculated for regions consisting of 5 k-mers. The maximum and minimum shifting scores of every sequence are kept.

4.1.4. Secondary Structure-Based Motif Detection: selex-blaster

The workflow *selex-blaster* (Figure 13) was designed to cluster sequences based on predicted secondary structures and find single-stranded motifs in the discovered clusters.

A method developed by Song M. et al., 2019[41] was adapted for the workflow. They developed a tool which calculated a compound score for every sequence. The compound score combined k-mer-based scores, cluster sizes and structural stability of sequence-based clusters. Their approach for sequence-based clustering was adapted for this project.

During the SELEX experiments some sequences got enriched that interacted heavily with the forward and reverse primers. The random region of these sequences evolved to resemble a reverse complement of the flanking SELEX primers, and folding prediction showed that they folded into tightly bound structures. To reduce the influence of these sequences a secondary-structure

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

based clustering approach, based on unbound looping regions, was proposed and implemented.
The adapted method consists of four sequential steps, looping region-based masking of sequences, creation of a similarity graph using pairwise alignment, graph clustering and an EM-algorithm for motif detection.
The workflow starts with a dereplication step (see *selex-assess* workflow), and determines representative sequences for every mutational family(see *selex-kmer* workflow) to avoid retrieving very tight clusters.

Structure-based Sequence Masking

Working principle The *forward* and *reverse primer* have to be provided for folding prediction. The *temperature* of the folding environment was set to 21°C by default. The number of *CPUs* is set to 4 by default. The *energy model file path* for DNA folding has to be set manually in the config file. The energy model for DNA folding by Mathews et al., 2004[77], was used, which is packaged with ViennaRNA.
MFE (minimum free energy) secondary structures are predicted using RNAfold[3]. The MFE structures are used to mask sequence files, making bound nucleotides unavailable in the subsequent pairwise alignment step.
For pairwise alignment BLAST[76] is used, as it is an established tool for pairwise alignment, using heuristic search method with high performance. BLAST is used to run a search of the masked sequence data set on itself. The focus of the search lies on very short matches, as the masking steps only leaves unbound looping regions available for the search. As the matched regions are of very short length (4-12nt), the E-value was not useful, and instead the alignment length is used. The result of this step is a similarity graph, in which only aptamers with similar unbound looping regions are associated.
Clustering is then done using MCL[40] (Markov Clustering Algorithm), an unsupervised and highly performant clustering algorithm based on graph weight dependent flow. When testing the workflow, MCL was able to cluster graphs with over $5 * 10^6$ edges in decent time. MCL offers different parameters for clustering, which influence the overall clustering granularity, the most important parameter being the *inflation factor*, which is set to 1.4 by default. A FASTA-file is created for every cluster. For more detail on MCL see appendix at A.2.
For motif detection the tool MEME from the MEME-suite is used[36, 28]. The MEME-Suite offers a wide variety of motif detection and enrichment analysis tools originally based on the EM-algorithm. For more detail on MEME see appendix A.3. At first STREME[37] was used for motif detection as it makes use of a linearly scaling algorithm and is therefore better suited for next generation

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

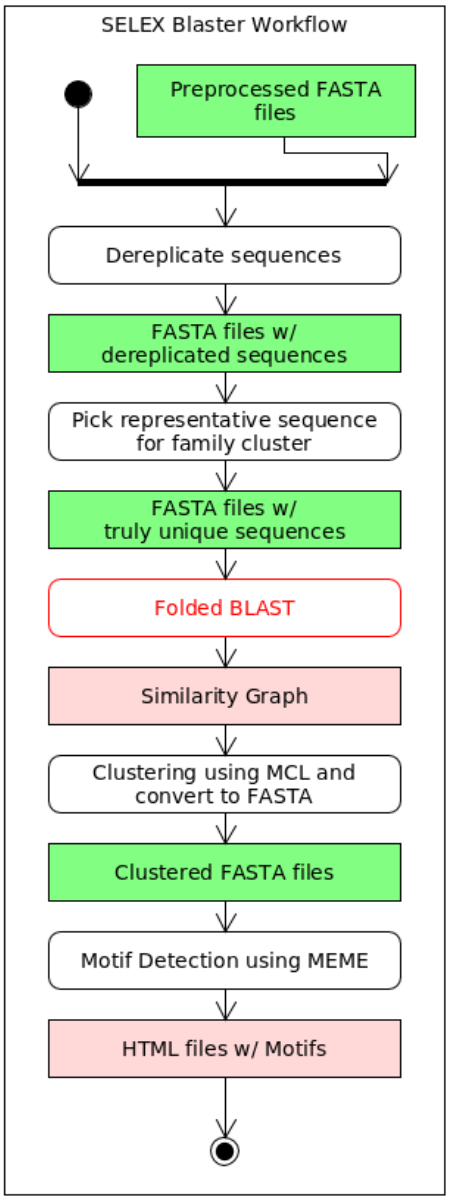


Fig. 13: Scheme of the *selex-blaster* workflow. Green rectangles represent sequence files (FASTA, FASTQ), pink rectangles represent plots and tabular files, and round edge rectangles represent analysis steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

sequencing data. When sequence masking was added at a late stage of development, cluster sizes got sufficiently small (10^0-10^3 sequences) to use MEME instead.
The result of this step is a motif file in HTML-format for every cluster, which can be opened in any conventional web browser.
In hindsight, using only one structure per sequence may limit the usefulness of this approach, as ssDNA does not have one fixed structure, and structure ensembles as seen in APTANI[38] or AptaMotif[21] could be used. As BLAST was used on masked sequences, matches were rather short, and E-values and bitscores were unreliable, so alignment length was used as clustering metric, which was not optimal. A way to overcome this could be to use a soft clustering based method using structure ensembles.

4.1.5. Discussion of Workflow Development

Workflow Management Workflow management tools have become a necessity in bioinformatics to allow for reproducible data analysis. Workflow managers allow for easier development, customization and maintenance, compared to complicated shell scripts or hard coded programs.
The presented pipelines were implemented using the workflow manager Nextflow. It is regarded as one of the top workflow managers in bioinformatics currently. Different workflow managers, namely Nextflow, Luigi and Snakemake, were initially tested. Nextflow has shown to be comfortable and intuitive to work with in multiple settings. Moreover, it can be easily extended using the package manager conda and works well using cluster computers.

Repository Hosting and Documentation All workflows developed in this thesis are managed using remote git repositories. Detailed info on execution, installation and configuration is provided in README pages included with every workflow. Using remotely hosted repositories offered many advantages compared to keeping all files on local data storage, such as: less risk of data loss, traceability of code changes, possibility to collaborate with others, having a central place to present the pipelines, and remote working from almost any machine.

Scripting Languages Development of own tools was preferred when the task at hand was easy-to-implement, such as k-mer counting or sequence dereplication, or if existing tools did not work as required, e.g. due to performance issues, as seen in MPBind[39], the form of results, reproducibility, customizability, or due to performing unnecessary analyses, as with smart-aptamer[41]. Existing tools were used for analysis or data preparation whenever feasible.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

By default, scripts in Nextflow process are executed as 'bash' and therefore tools included in POSIX environments are accessible. For easy-to-implement tasks, such as converting from FASTQ to FASTA or extracting data from tabular files, tools such as *awk* and *sed* were used.
Python 3 was used for computationally demanding tasks. One reasons to opt for Python was its broad support from the bioinformatics community, offering many useful libraries. Also, due to Python being a scripting language, compilation was not needed. Computationally demanding tasks tended to run much faster using Python, compared to the R scripting language.
The statistical scripting language R 3.6[68] was used for data wrangling and plotting tasks, as high-quality libraries for data wrangling and visualization are available.

4.2. Analysis of whole-cell HT-SELEX Experiments

All bacterial whole cell-SELEX experiments were performed with a chemically synthesized ssDNA library purchased from Integrated DNA Technologies (Coralville, USA). The ssDNA library was designed to have a 40-nucleotide random region (N40, equal distribution of A, T, G, and C), which is flanked with constant primer binding regions (forward primer: 5-TAG GGA AGA GAA GGA CAT ATG AT, reverse primer 5-TCA AGT GGT CAT GTA CTA GTC AA-3) at the 5' and 3' end. Sequencing of the SELEX pools (cell-bound ssDNA pools from different SELEX rounds) was done using an Illumina MiSeq with an expected output of roughly $2 * 10^6$ sequences (MiSeq Reagent Micro Kit v2, 300-cycles)[14]. Data analysis tasks were executed in a POSIX environment using Manjaro Linux 21.0.4.

Data Availability Raw FASTQ-files of SELEX EF05 can be downloaded from the Sequence Read Archive (SRA) at NCBI under the accession number PRJNA615076. For access to the raw FASTQ-files of SELEX EF01 and SELEX EF07 contact C. Kolm[14].

4.2.1. SELEX EF01

SELEX EF01 was performed to generate DNA aptamers against *Enterococcus faecalis*. After 9 consecutive SELEX rounds with increasing selection pressure[78], ssDNA pools of each round were subject to next generation sequencing to determine whether sequence enrichment took place. According to

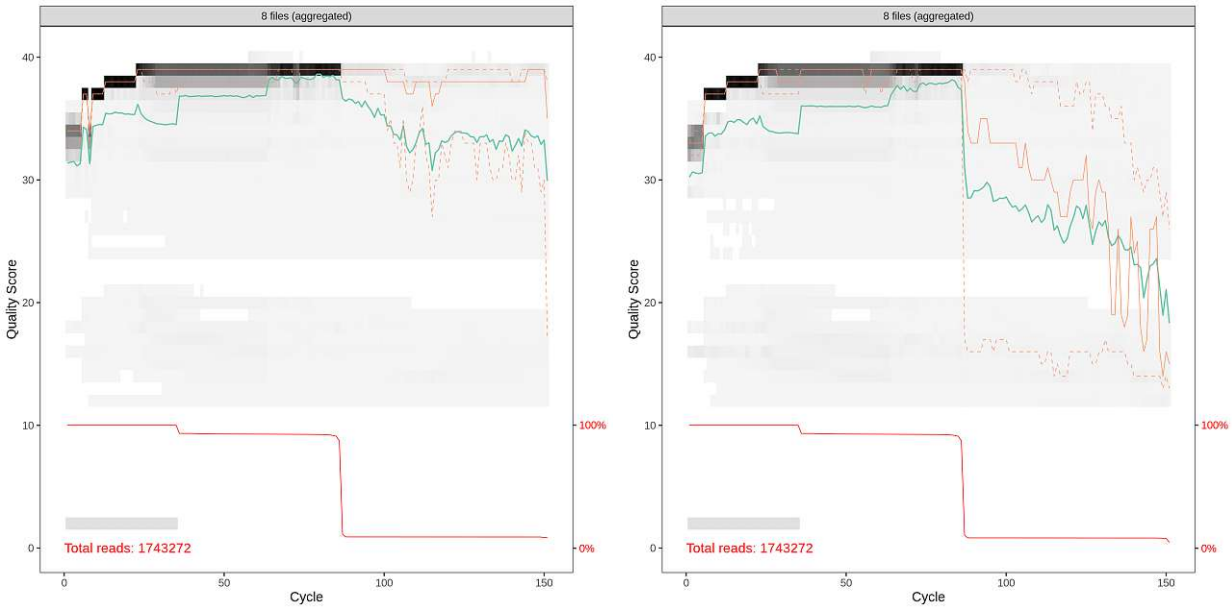
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

qPCR measured amounts of recovered ssDNA sequence and remelting curve analyses, no significant changes in ssDNA pool binding and heterogeneity were observed[78].

4.2.1.1. Results from workflow *selex-ngs-prep*

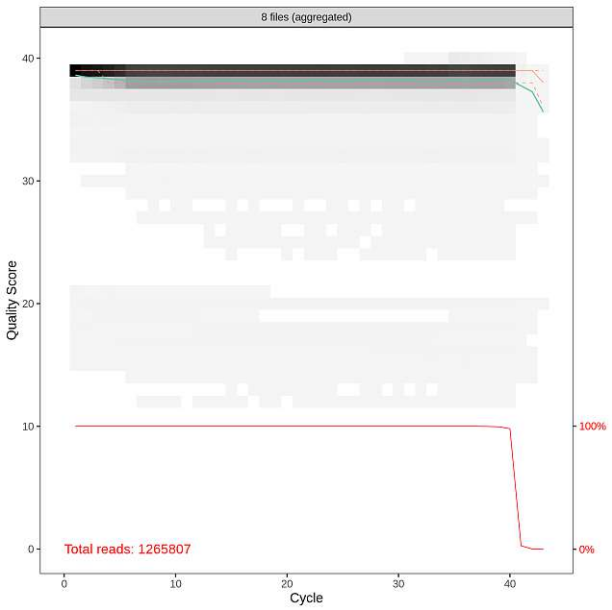
In total, 1,743,272 reads were generated and successfully demultiplexed by the Illumina MiSeq platform (R00, R02-R09).
Before preprocessing, the following parameters were set: i) a maximum error rate of 20% for successful adapter recognition, ii) every sequence had to show an average sequencing quality of at least 30 along the random region, iii) for paired-end read merging 1 mismatch was allowed and mismatch error-correction was enabled and iv) random regions were allowed to be 37 to 43 nucleotides long.
The number of reads per SELEX round was sufficiently well distributed, with R4 posing the greatest outlier from the mean with 4.04% deviation, as can be seen in Table 2.
Sequencing quality is plotted in Figures 14a and 14b. They show that almost all reads are around the designated length of 86 nucleotides, which can be seen at the drastic drop of the red line. The dotted orange lines show the 25th and 75th percentile and are both above 30 at all times in the region of interest. Overall, the quality values tended to be lower at the beginning of the reads and steadily increased. However, combining the mean quality value with the quality heatmap in the background shows that only very short sequences tended to be of low quality, which can be observed by an increase of mean quality at positions 40, when those reads are not considered for the plot anymore. Figure 14c shows the sequencing quality after preprocessing to show an average quality of around 38.
Table 3 and Tables 21, 23 and 22 in the appendix show the effect preprocessing had on the numbers of sequences. Figures 15a and 15b plot the preprocessing loss.
Overall, adapter trimming introduced an average loss of 10.97%. This step also discarded very short sequences and sequences of extremely low quality with no adapters detected. Quality filtering introduced an average loss of 17.23%, with a substantial amount of sequences removed (58.24%) in Round R8. At the quality trimming step on average 11.37% were lost (excluding R8). Paired-end merging and size limiting introduced very low numbers of loss (0.28%, resp. 0.20%).
The quality filter outlier of SELEX round R8 was especially severe. R8 had the lowest read count from the beginning, so either the sequencer was not able to properly demultiplex reads of R8 or there was another problem during library preparation, e.g. the barcoding step.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Forward Reads

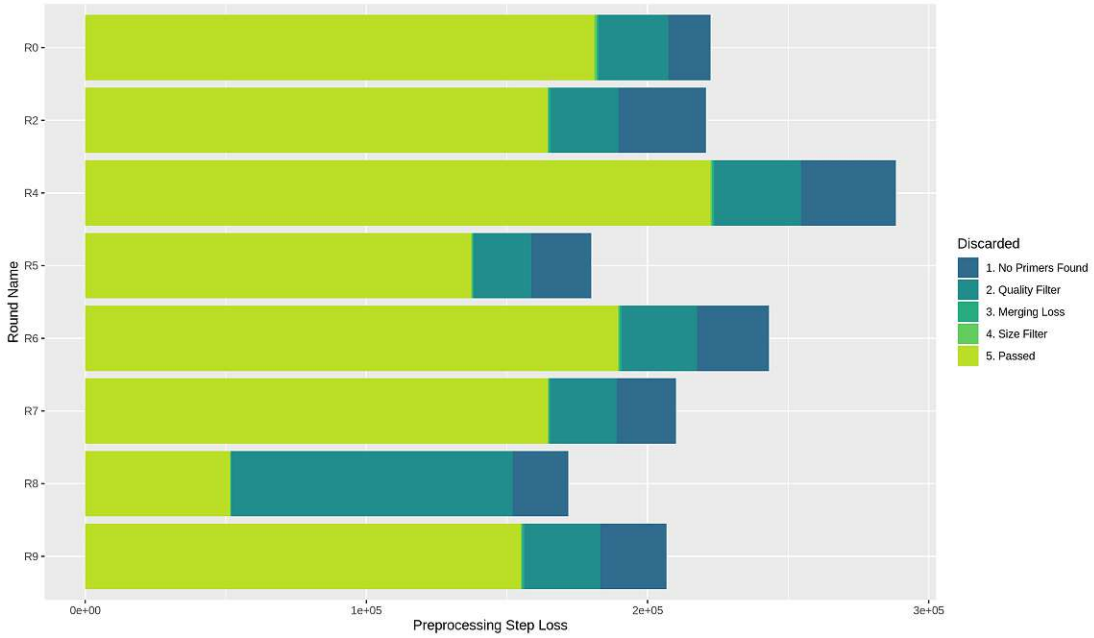
(b) Reverse Reads



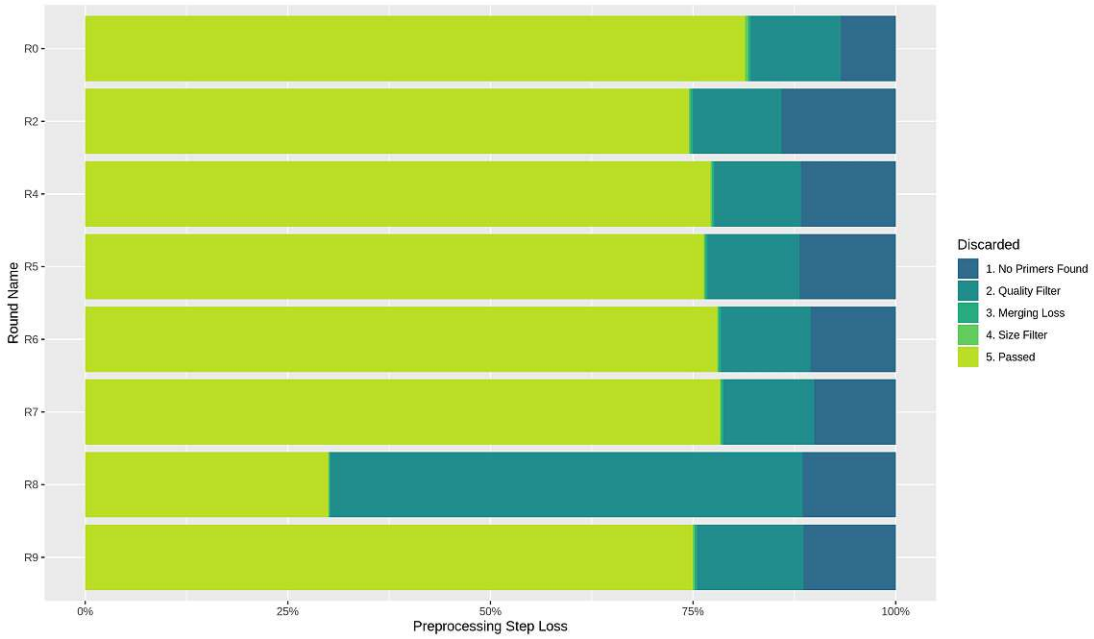
(c) Preprocessed Reads

Fig. 14: Sequencing quality profiles of EF01 raw and preprocessed reads. Median quality is plotted as a straight orange line while the mean quality is plotted as a straight green line. The background shows a heatmap for the overall quality distribution. Dotted orange lines show the 25th and 75th percentile of quality. The straight red line shows the share of considered reads.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Absolute numbers



(b) Scaled to 100%.

Fig. 15: Loss of preprocessing steps in EF01.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

	Total reads	Share in NGS run	Dev. from mean
R0	222414	12.76%	0.26 %
R2	220753	12.66%	0.16 %
R4	288262	16.54%	4.04 %
R5	180032	10.33%	2.17 %
R6	243162	13.95%	1.45 %
R7	210076	12.05%	0.45 %
R8	171799	9.85%	2.65 %
R9	206774	11.86%	0.64 %
	1743272	100.00%	

Tbl. 2.: Sequenced reads for SELEX EF01 (R0-R9).

Discarded	Trimming	Filter	PE-Merge	Length-limit
R0	6.77%	11.14%	0.28%	0.38%
R2	14.09%	10.92%	0.33%	0.19%
R4	11.66%	10.72%	0.29%	0.19%
R5	11.85%	11.38%	0.31%	0.16%
R6	10.50%	11.08%	0.28%	0.17%
R7	10.04%	11.23%	0.27%	0.15%
R8	11.51%	58.24%	0.15%	0.11%
R9	11.38%	13.12%	0.33%	0.23%
Avg	10.97 %	17.23 %	0.28 %	0.20 %

Tbl. 3.: Share of discarded reads (in %) during every preprocessing step for SELEX EF01. Shares are taken for the original data set size.

4.2.1.2. Results from workflow *selex-assess*

Nucleotide distribution and sequence enrichment in the cell-bound ssDNA pools of SELEX round R2-R9 were then determined using the workflow *selex-assess*. For visualization and simple interpretation, the workflow returns a series of tables and plots.

Figures 17a to 17c show the distribution of nucleotides at each position of the 40-nucleotide long random region in the ssDNA library and the ssDNA pools of SELEX rounds R4 and R9, while Figure 17d shows the overall changes in nucleotide composition over the SELEX rounds.

Results indicated that that the initial ssDNA library composition was slightly increased in adenine with 29.1% of all nucleotides. Guanine (24.0%) and thymine (24.8%) were close to 25% while cytosine was slightly decreased (22.2%). Over the course of SELEX EF01 cytosine rose to a share of 31.7%, while adenine and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

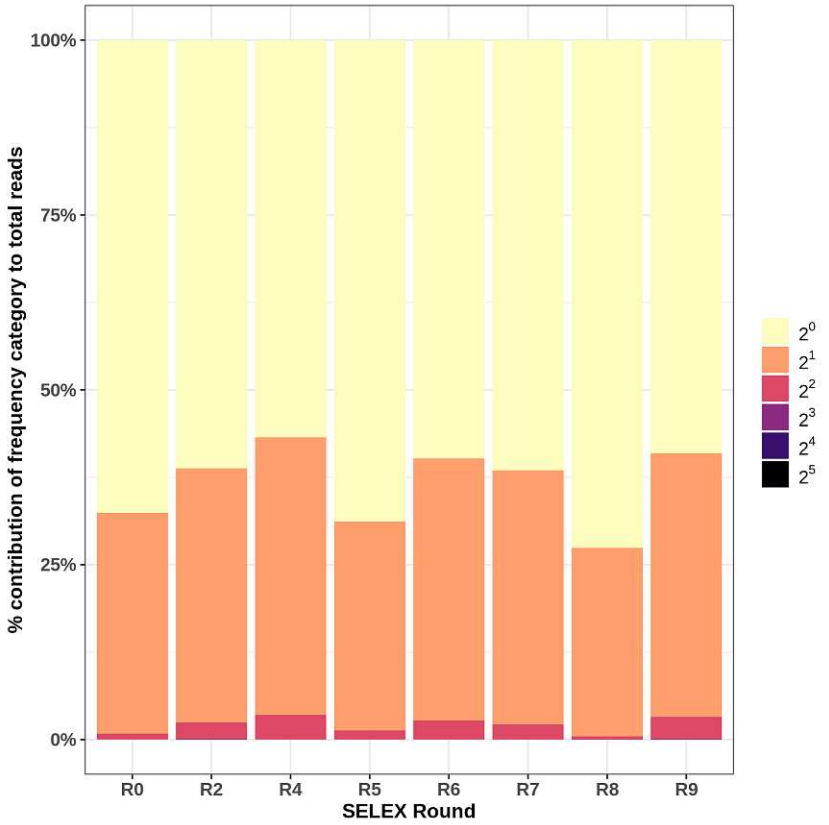
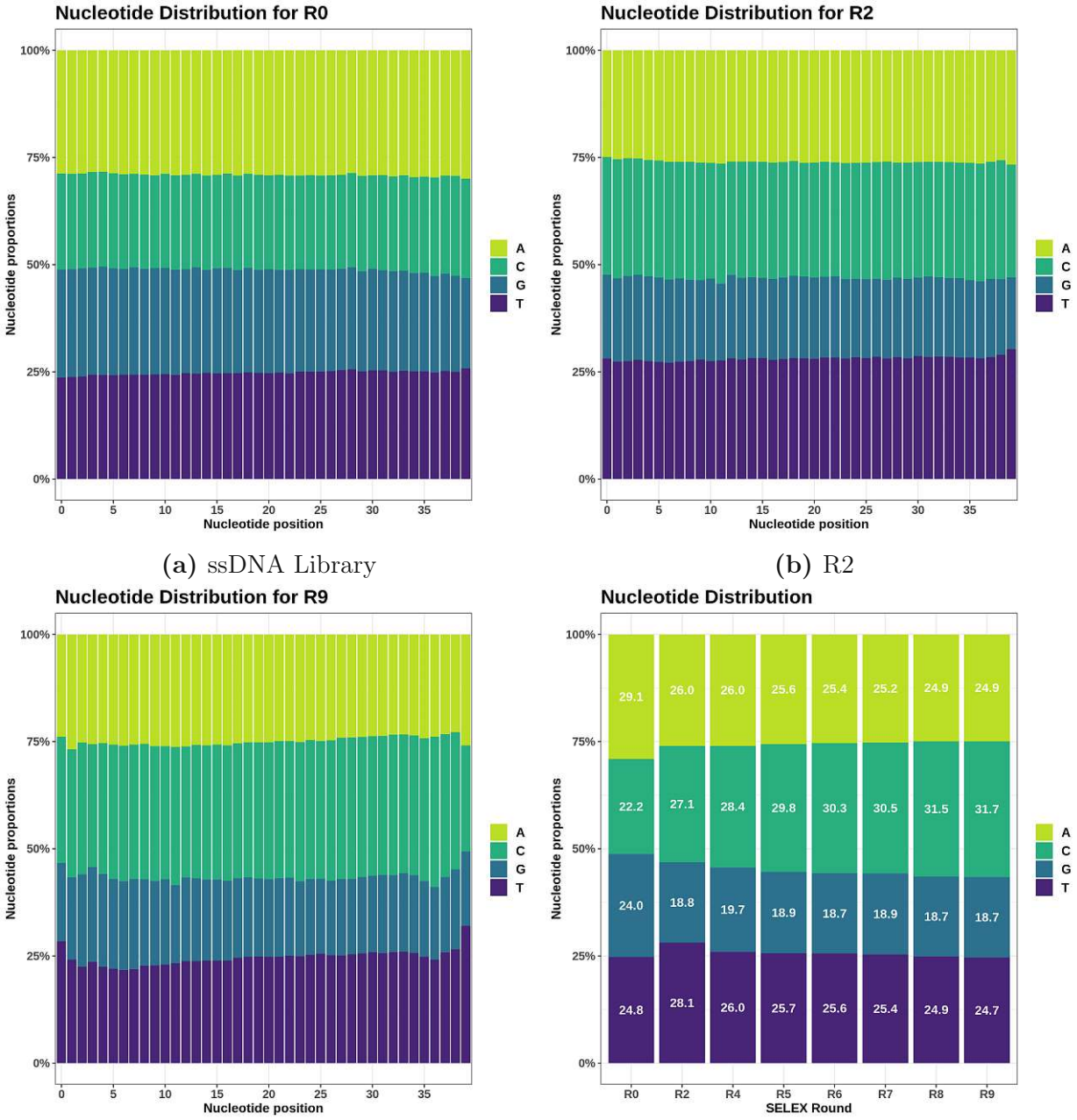


Fig. 16: Assessment of SELEX EF01 (R0-R9) in terms of sequence enrichment and frequency. Sequence enrichment can be observed. The darker a bar, the higher the replication number of the reads represented by them. Bars are as tall as the total read count of the sequences in them.

guanine declined to 24.9% and 18.7% respectively. Thymine increased to 28.1% in R2 and then slowly declined to a share of 24.7% in the last round.
The established workflow *selex-assess* revealed that no substantial enrichment of sequences was observable over the course of the bacterial whole-cell SELEX EF01, as seen in Figure 16. Thus, no further analyses were conducted and the SELEX experiment was aborted.
Appendix table 24 shows the top 25 reads for the last SELEX Round of EF01, which however were not chemically synthesized and tested in binding assays.
In conclusion, NGS data analysis confirmed the observations made by qPCR-based remelting curve analyses (no significant changes in ssDNA pool binding and heterogeneity) and thus verified the usability of the workflow for in-line monitoring of the SELEX process.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) ssDNA Library

(b) R2

(c) R9

(d) Nucleotide composition dynamics of EF01

Fig. 17: Nucleotide distributions of EF01 at each position of the 40-nucleotide long random region of the ssDNA library(a) and SELEX rounds R4(b) and R9(c), as well as changes in the nucleotide composition in the random regions over the selection rounds R0-R9(d).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

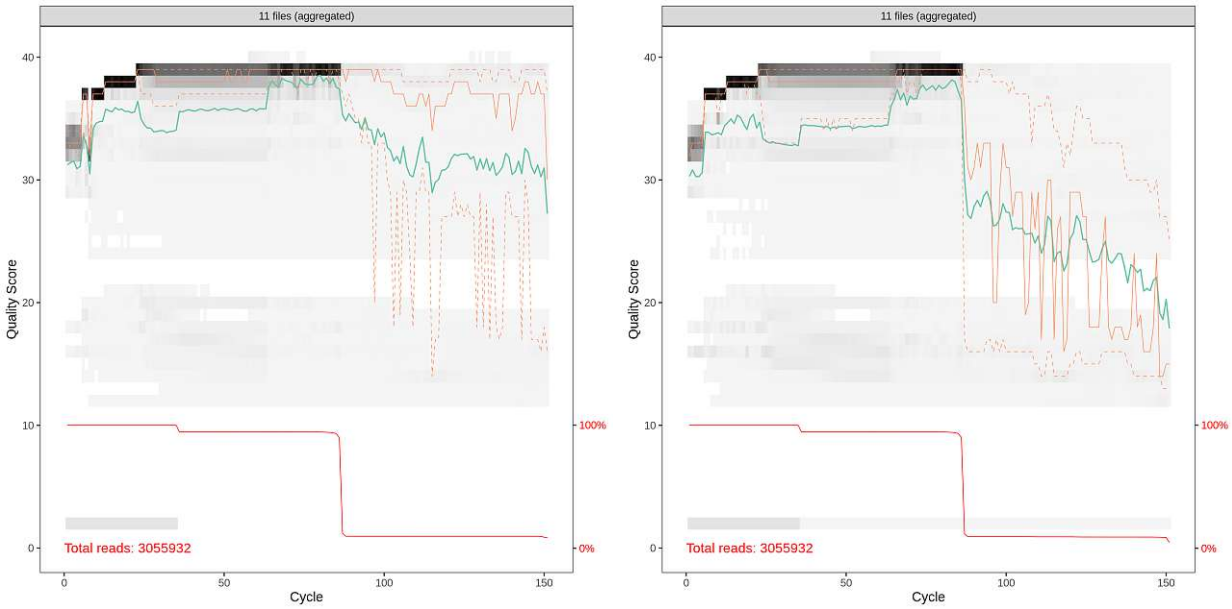
4.2.2. SELEX EF05

Like SELEX EF01, SELEX EF05 was performed to generate DNA aptamers against *E. faecalis*. In contrast to SELEX EF01 however, 11 consecutive SELEX rounds were performed and according to qPCR-based remelting curve analyses, changes in ssDNA pool populations were observed, indicating a potential enrichment of sequences[78]. To confirm these results and to identify potential aptamer candidates, ssDNA pools from SELEX rounds R02-R11 were subject to next generation sequencing and data analysis[78].

4.2.2.1. Results from workflow *selex-ngs-prep*

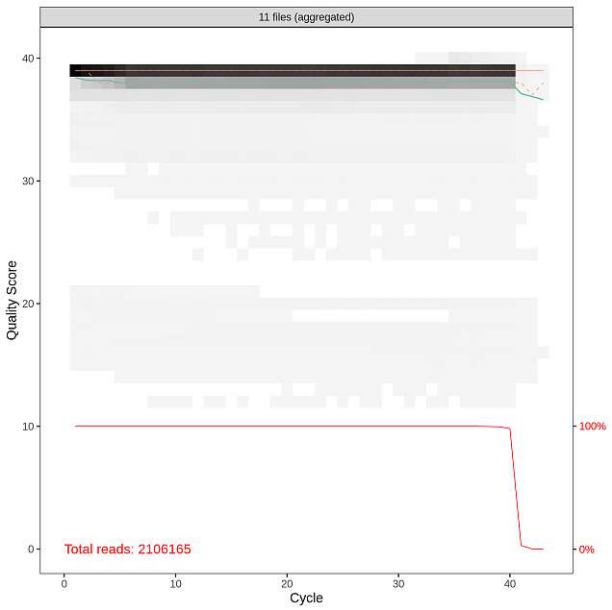
In total, 2,948,294 reads were produced and successfully demultiplexed by the Illumina MiSeq platform (R02-R11). The data set for the ssDNA library (referred to as R00) was taken from the EF01 sequencing run and increased the sequence count to 3,055,932 reads, which were put into the *selex-ngs-prep* pipeline. The number of reads per SELEX round was again sufficiently well distributed, indicating that library preparation and pooling of SELEX samples was performed properly. The plots in Figures 18a and 18b show the sequencing quality. Figure 18c shows the sequencing quality after preprocessing with an average quality of 38.
Table 5 and Tables 25, 26 and 27 in the appendix show the effect preprocessing had on the numbers of sequences. Figures 19a and 19b plot the preprocessing loss, including the ssDNA library R00. The preprocessing parameters used were identical to the ones of SELEX experiment EF01, given in 4.2.1.1.
Overall, adapter trimming introduced an average loss of 10.29%. This step also discarded very short sequences and sequences of extremely low quality with no adapters detected. Quality filtering introduced an average loss of 21.04%, which was on average higher than in EF01. In the quality filter step no outliers were observable. In contrast, paired-end merging and size limiting introduced very low numbers of loss (0.24% and 0.11% respectively).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Forward Reads

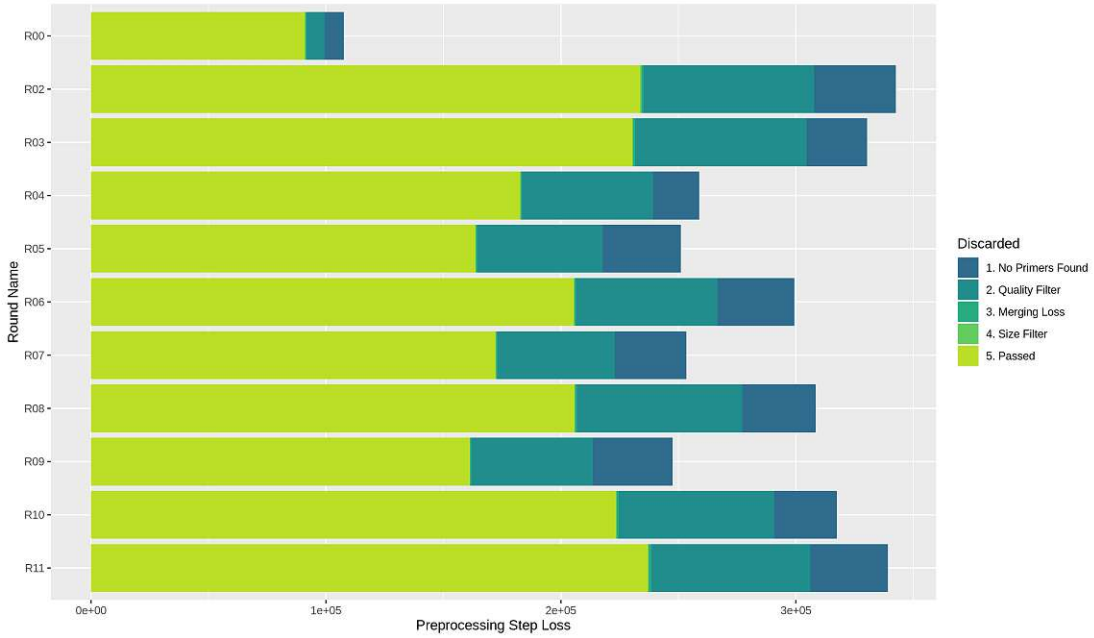
(b) Reverse Reads



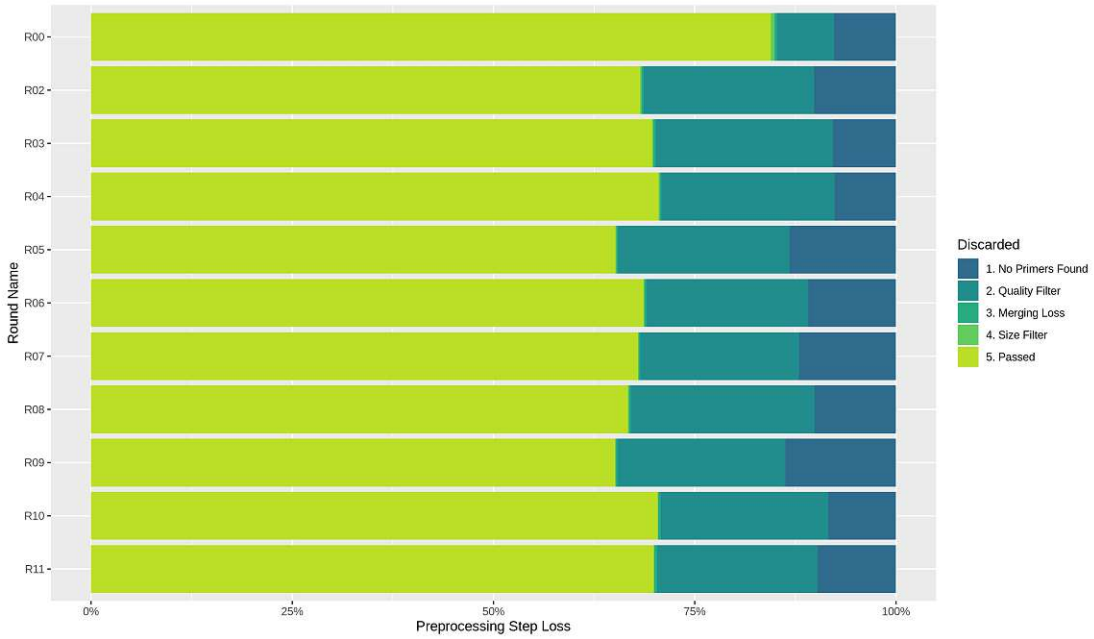
(c) Preprocessed Reads

Fig. 18: Sequencing quality profiles of EF05 raw and preprocessed reads. Median quality is plotted as a straight orange line, while the mean quality is plotted as a straight green line. The background shows a heatmap for the overall quality distribution. Dotted orange lines show the 25th and 75th percentile of quality. The straight red line shows the share of considered reads.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Absolute numbers



(b) Scaled to 100%.

Fig. 19: Loss of sequences after preprocessing steps in EF05.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	Total reads	Share in NGS run	Dev. from mean
R02	342616	11.62%	1.62 %
R03	330387	11.21%	1.21 %
R04	258891	8.78%	1.22 %
R05	251043	8.51%	1.49 %
R06	299303	10.15%	0.15 %
R07	253449	8.60%	1.40 %
R08	308397	10.46%	0.46 %
R09	247589	8.40%	1.60 %
R10	317454	10.77%	0.77%
R11	339165	11.50%	1.50%
	2948294	100.00%	

Tbl. 4.: Sequenced reads for SELEX EF05 (R02-R11).

Discarded	Trimming	Filter	PE-Merge	Length-limit
R02	10.14%	21.21%	0.25%	0.16%
R03	7.83%	22.04%	0.27%	0.14%
R04	7.62%	21.52%	0.20%	0.13%
R05	13.25%	21.29%	0.21%	0.11%
R06	10.90%	20.14%	0.21%	0.10%
R07	12.04%	19.73%	0.19%	0.09%
R08	10.11%	22.83%	0.24%	0.10%
R09	13.75%	20.86%	0.20%	0.09%
R10	8.41%	20.83%	0.29%	0.10%
R11	9.73%	19.95%	0.34%	0.09%
Avg	10.38%	21.04%	0.24%	0.11%

Tbl. 5.: Share of discarded reads (in %) during every preprocessing step for SELEX EF05. Shares are taken for the original data set size.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2.2.2. Results from workflow *selex-assess*

Nucleotide distribution and sequence enrichment in the cell-bound ssDNA pools of SELEX round R02-R11 were then determined using the workflow *selex-assess*. For visualization and simple interpretation, the workflow returns a series of tables and plots.

Figure 21a to 21c show the distribution of nucleotides at each position of the 40-nucleotide long random region in ssDNA pools of SELEX rounds R02, R06, and R11, while Figure 21d shows the overall changes in nucleotide composition over the SELEX rounds. As described in EF01, the initial ssDNA library was slightly biased towards sequences with elevated adenine(29.1%) and reduced cytosine nucleotide bases (22.1%). Over the SELEX process, the nucleotide distribution then changed by an increase of cytosine- and thymine-rich sequences of 8.0% and respectively 3.8% in SELEX round R11 (Figure 21c). Likewise, changes in nucleotide composition in the random regions were determined (Figure 21d).

The workflow *selex-assess* revealed that sequences were enriched over the course of the bacterial whole-cell SELEX EF05. From SELEX round R07 on, an increased number of sequences with > 10 reads were detected, while the proportion of unique sequences gradually decreased (Figure 20). Thus, next generation sequencing data confirmed the observations made by qPCR-based remelting curve analyses, which suggested a decrease in pool diversity and an increase of enriched sequences from R09 on.

In Table 6, the top 25 reads encountered in the last round R11 of SELEX EF05 from the workflow *selex-assess* are given, while Table 7 contains all potential aptamer candidates, which were then selected, chemically synthesized and experimentally screened for target binding. Amongst them, aptamer EF05-508 showed high affinity and specificity for *E. faecalis* target cells[14].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

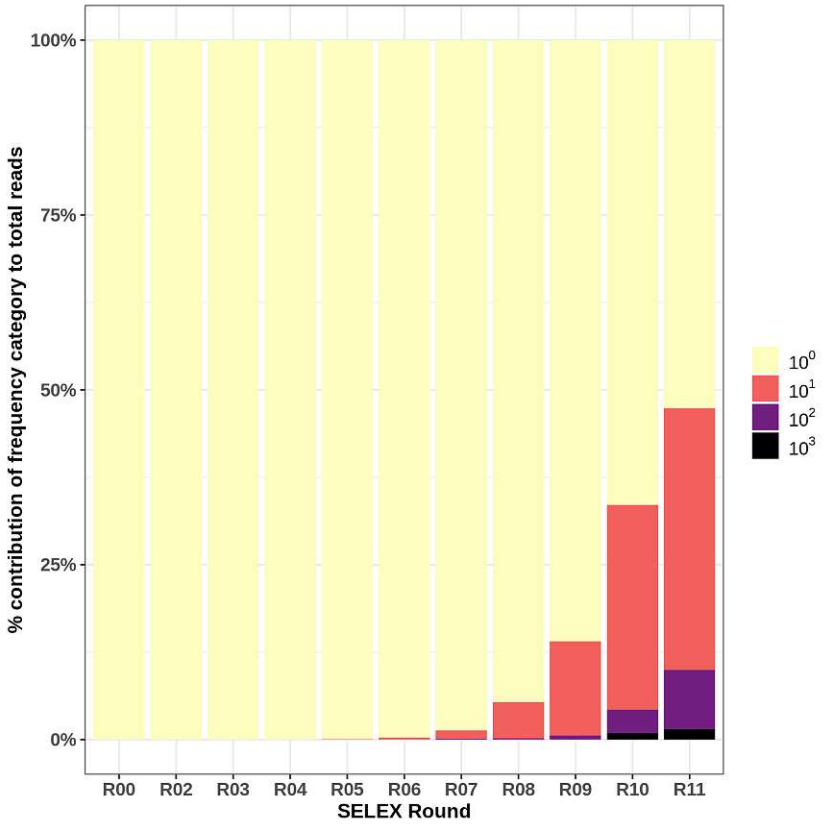
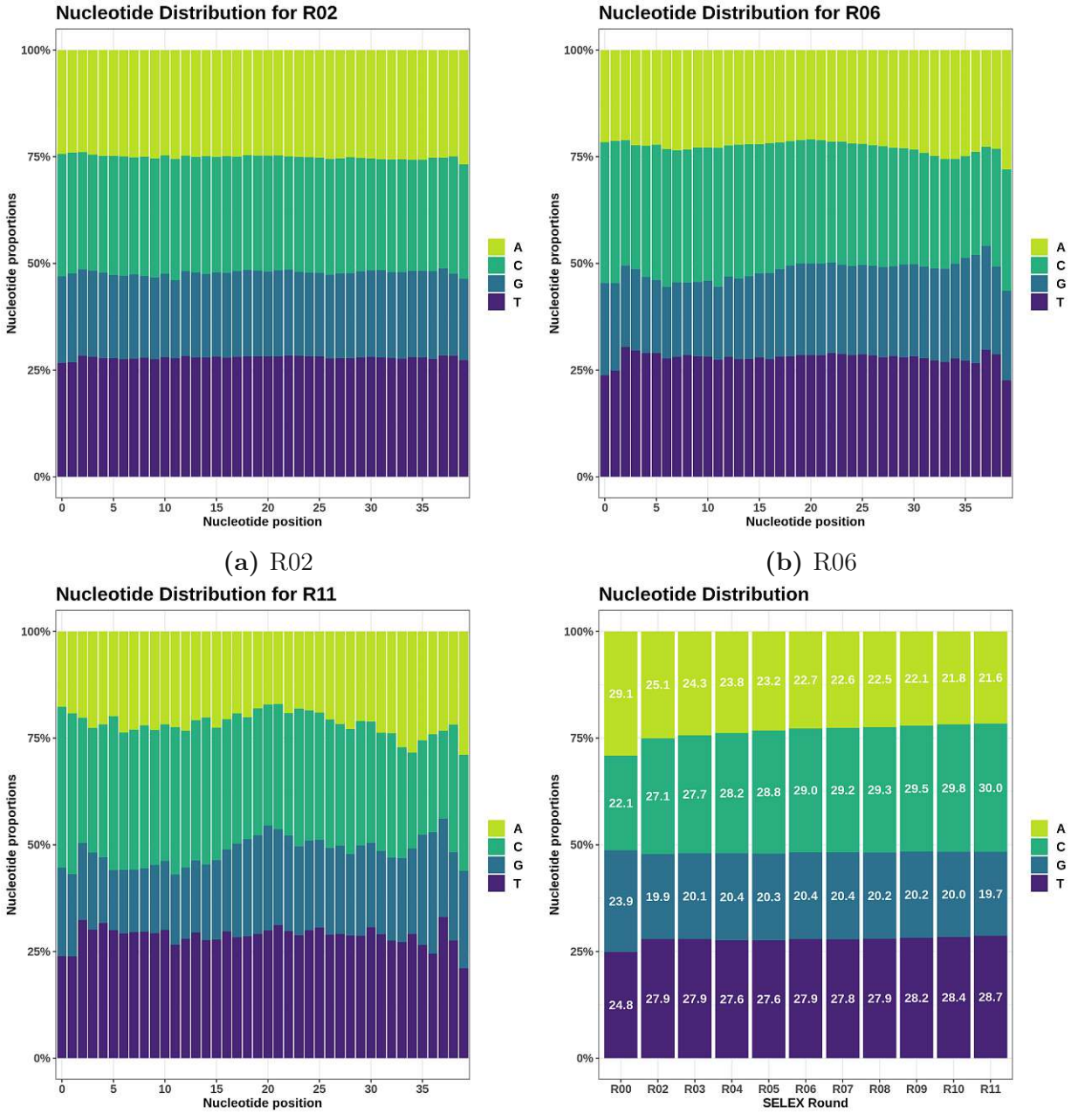


Fig. 20: Assessment of SELEX EF05 (R00-R11) in terms of sequence enrichment and frequency. Sequence enrichment can be observed. The darker a bar, the higher the replication number of the reads represented by them. Bars are as tall as the total read count of the sequences in them.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) R02 **(b)** R06

(c) R11 **(d)** Nucleotide composition dynamics of EF05

Fig. 21: Nucleotide distributions of EF05 at each position of the 40-nucleotide long random region of SELEX rounds R02(a), R06(b) and R11(c), as well as changes in the nucleotide composition in the random regions over the selection rounds R02-R11(d).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Rank	Count	Random Region	Tag
1	3538	TTTCTCAACGGGACCATCACTTACCTCAAGTACTTGGACG	EF05-501
2	557	GTCAACTCATTTATGGTGCTCCTCGTACCTCAGGTGGTTA	EF05-503
3	547	CCGGCTATCTCCCTACCGTGGCCGAGTACCTCAAACGTTT	EF05-502
4	392	CTGCCTGACTTCATAATGCTTCTTCTCCCTGTGGTACTTG	
5	390	GGCTCTCTGGTCTTCAAGGCCCATGATTACAGTCAGATCA	
6	360	CCTTAGGTCTTAACTATCAGGCGGTCTGTATCAATTCGAT	
7	330	GGCAGGGAGCGACCGGGTCATTGATTATCTGTCAAAGTGT	
8	322	GGCCATACCTCGTGCCTTCTGTGATCATCTCTATCAATTG	EF05-504
9	316	CTCAATCATCAAGGGTCTACTTCCCGCTTGTGGGCCATTC	
10	298	CCTCTCTCTTACTGCTACTGGGCAGGGTACTCAATTACGT	EF05-505
11	279	CGGTCCCGACTCAATATTGTTCCCTCCCCTTATCAGGCGG	EF05-506
12	270	TCCTCTAATCAACTCTATGCCTTATCCCCTTGGTCAGGAC	EF05-507
13	250	CAGGTCTCGTCCCTTGTGGAACAGGAATACTGGCATCACA	
14	249	CATGGCTCCCTCTTCAACTTCAAGTCAGTGATCTGTCAAA	
15	249	CCTCTGTCAGAGCCATCATAGAGACTATGATCCCTGGTCA	
16	241	CCTGGATCATCGATGGCAAAAGCGCATTCCCGGCATGTGGC	
17	234	ACGCACATCATGAATTGGCCACTCATCACTTTATCGTGGT	
18	225	ATGGCCTAGTTCTGCCCCCGGGGACATAGCTCAAACGCGA	
19	223	GCATTACATCAACACTATGACCATTCGTGACCGGAGTGGC	
20	220	TTCGAATTCATCTAGTGTCAATCATCATCCCTGGTCATTC	
21	218	ACTGGCCTTGACACCCTGTTGTGGCTTGATGACAATAACA	EF05-508
22	213	CCTGTCTCGTCCTAAGTAATGGTTTCATGTAACCTCAACT	
23	211	GGCCATCCCCCAATCGCGGTGGGCTATGCACCTCAACAAG	
24	205	CCTTCGCCTCTCTACAAGGGCGCAATGCTTCGCTCAATCGT	
25	204	TTCCTCCTCCTCTGACTGTTGTTGTCGGTAATATCAATCC	

Tbl. 6.: Top 25 Reads from last enrichment round of SELEX EF05.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Tag	Random Region
EF05-501	TTTCTCAACGGGACCATCACTTACCTCAAGTACTTGGACG
EF05-502	CCGGCTATCTCCCTACCGTGGCCGAGTACCTCAAACGTTT
EF05-503	GTCAACTCATTTATGGTGCTCCTCGTACCTCAGGTGGTTA
EF05-504	GGCCATACCTCGTGCCTTCTGTGATCATCTCTATCAATTG
EF05-505	CCTCTCTCTTACTGCTACTGGGCAGGGTACTCAATTACGT
EF05-506	CGGTCCCGACTCAATATTGTTCCCTCCCCTTATCAGGCGG
EF05-507	TCCTCTAATCAACTCTATGCCTTATCCCCTTGGTCAGGAC
EF05-508	ACTGGCCTTGACACCCTGTTGTGGCTTGATGACAATAACA
EF05-509	CCTCACTCTTGACCCAAAGTGCATGCTCTATTCATTCGGA
EF05-510	GCTTCTGTGCACATTAAGGCACTCGTCTTCACTGTGGTTC
EF05-511	CCTAACTCACTTACCAGCACGAGGTGCCTGTACCATCAAT
EF05-512	CTCTCATCACAGGAATTTGAATTTCCCTTGTGGACAGTAA
EF05-513	GATGTGAATTCCGTCCCTTGGTCAGACACTTCAACACCGG
EF05-514	TCTCGACGCTATGATCAAGACGCAGTATGATGGCACATCA
EF05-515	TTAACCCTCATTTAATGGCCGCGTCAATCCGCAAAGGGTC
EF05-516	TTCCTTCGCAGGACACCGATGGCCAGGCGCGAGTCAATAT
EF05-517	TCCTATGGCCGCATCCCTTCAAGGACAAGCTCACAAGAAT

Tbl. 7.: Characterized aptamer sequences from EF05.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

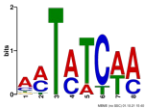
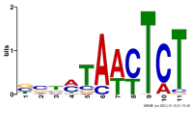
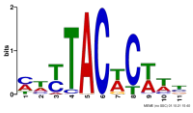
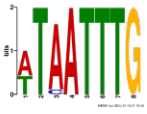
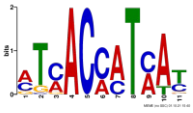
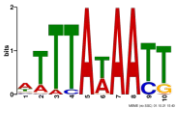
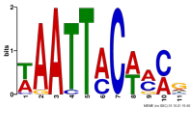
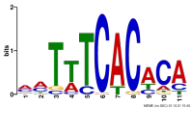
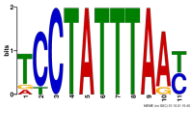
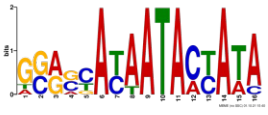
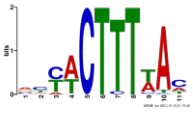
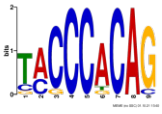
4.2.2.3. Results from workflow *selex-blaster*

Motif detection was performed on a data set that combined all round files of SELEX EF05 (R00-R11) by using the workflow *selex-blaster*, to find the looping regions responsible for binding.
The workflow reported to have found 14.247 clusters. Tables 8 and 9 show the top three motif logos for the first 20 clusters. Apparently, some motifs repeated to some degree in the found clusters. For example the sequence TAATA is contained in motifs c000007/1, c000008/2, c000011/1, c000012/1, c000014/1, c000015/2. Bitscore of discovered motifs was lower than expected.
The combined SELEX rounds dereplicated to 913.582 unique sequences, of which representative sequences were chosen for every mutational family cluster, resulting in a data set containing 873.737 sequences. The data set was masked after folding prediction and used to create a similarity graph using BLAST. The similarity graph created contained 45.169 sequences connected by 69.083 edges. Only around 5% of the input sequences were retained, probably due to the strict restrictions imposed by secondary-structure based masking. Clustering was done using an inflation factor of 1.2 for MCL. MEME (MEME-Suite 5.3.0) was used on all clusters that had at least 20 sequences (122 clusters).

4.2.2.4. Results from workflow *selex-kmer*

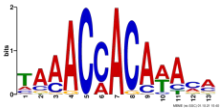
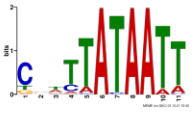
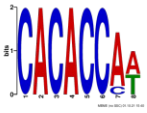
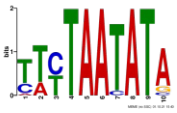
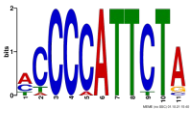
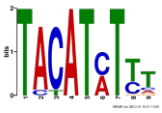
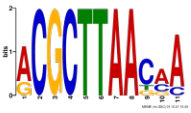
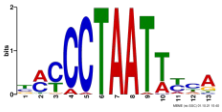
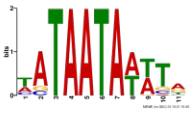
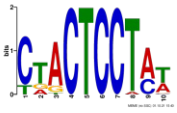
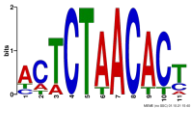
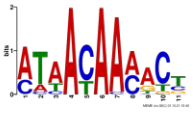
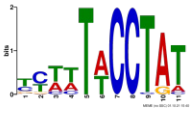
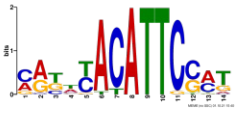
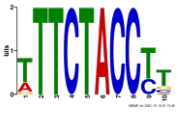
K-mer-based scoring was performed to estimate binding affinity of sequences, by using the workflow *selex-kmer* for 6-mers.
Scores were calculated for every combination of datasets including rounds R02-R11 and library R00. To avoid overrepresentation of k-mers, the data sets were dereplicated and filtered to only include representative sequences of mutational family clusters, similarly as described in the results of *selex-blaster*.
The *selex-kmer* workflow returned k-mer based sequence scores for every round combination. Scores consisted of the overall aptamer score, the lowest and the highest shifting score. The shifting score was based on 5 consecutive k-mers of length 6, so a stretch of 10 characters.
Table 10 shows the top 15 sequences chosen using the highest shifting scores of round R11, and contains sequences with their highest shifting scores for rounds R2, R5, R8 and R11. Table 11 shows motifs of the top 1000 sequences of R11, chosen by highest k-mer shifting scores. Motifs 1 and 3 have the consensus sequences CTCTCT, which is included in aptamer candidate EF05-505. Parts of motif 3 (consensus sequence AATCATCATG) can be found in EF05-504 (ATCAT) and EF05-512 (TCATCA).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Cluster	Seqs	Motif 1	#	Motif 2	#	Motif 3	#
c000000	135		113				
c000001	121		95				
c000002	106		76				
c000003	86		53		13		
c000004	74		28		20		11
c000005	73		40		15		
c000006	71		50				
c000007	70		42		8		
c000008	67		40		7		
c000009	64		43		12		

Tbl. 8.: Motifs detected using MEME in clustered data sets of EF05, continued in Table 9. Motifs of E-value ≤ 0.05 are shown.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Cluster	Seqs	Motif 1	#	Motif 2	#	Motif 3	#
c000010	64		39				
c000011	62		35		15		
c000012	61		15		13		10
c000013	60		16		24		7
c000014	60		33		11		
c000015	59		37		4		
c000016	56		24		16		
c000017	54		29		14		
c000018	52		37				
c000019	52		21		15		

Tbl. 9.: Continuation of Table 8.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Sequence	R2	R5	R8	R11
GCTGAAAGAGTTACAATGAAAGAGTCCGTCAACATCATCT	14.63	3.47	1.70	3.68
CCTATCCGCGATAACCTCCCGTATCGTGTCTCTCTCTCTC	18.04	4.77	2.23	3.43
ATCCATGAAAATGGTACTGCCATCGGCCCTCTCTCTCTCC	18.70	4.79	2.59	3.43
CTCCATCTACTGTTACGCCGTCTACTGTTATTGCATCATG	16.15	3.59	2.17	3.25
CCCGGCTGAATTTACATCTGGCATGGATTTACATCATCGA	13.85	3.84	2.24	3.19
TGACGTGGCGTATCCCCGCGTAACTATCTCTCTCTCTCGC	16.58	4.47	1.65	3.18
TTCCTCTCTCTTCTATAGTGGAATCAAACAGTAATTAAGA	18.23	4.32	2.44	3.06
CCCGCGTTTCTTCCTCTCTCTTCGCGCACCGTACACGTCT	18.26	4.32	2.44	3.06
GGTTTGCCGCGTTCGGGACCTCTCTCTTCCTGGCACCACG	18.11	4.32	2.44	3.06
TCGTAATCACACGATGCCTCTCTCTTCAGACAACCGGTGT	18.11	4.32	2.44	3.06
CACGTATCAAGGTTTGCCTCTCTCTTCTTAGGCTTGAAAC	18.11	4.32	2.44	3.06
GGAGACTCTTTAATACCTCTCTCTTCGGTTCTGCATGAAT	18.11	4.32	2.44	3.06
ACTAACCACATGTCCTTCCGCATCCTCTCTCTTCAAGGTT	18.11	4.32	2.46	3.06
CCTCTCTCTTCCATCCCCCCACTATTTTAAGACAGGTTTC	18.11	4.32	2.44	3.06
GTACATCACATGGTAGATCCTCTCTCTTCCACGCAAAACA	18.11	4.32	2.44	3.06

Tbl. 10.: Sequences maximizing the k-mer shifting score in the last round R11 of SELEX EF05.

Motif	**E-value**	**Sites**	**Width**
	3.1e-1462	777	15
	4.9e-295	108	10
	1.1e-061	102	9

Tbl. 11.: Motifs found in the top 1000 sequences maximizing the k-mer shifting score in round R11 for SELEX EF05. MEME 5.3.3[28] was used in anr-mode to find motifs.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2.3. SELEX EF07

In contrast to SELEX EF01 and SELEX EF05, SELEX experiment EF07 was performed to *in vitro* select species cross-reactive DNA aptamers that bind to intestinal enterococci. To that end, a toggle-SELEX experiment approach was used in which four enterococcal species, namely *Enterococcus faecalis*, *E. faecium*, *E. durans* and *E. hirae*, served as target cells in alternating rounds of SELEX, see Table 12. It must be highlighted that after nine rounds of SELEX it was no longer possible to produce enough ds/ssDNA for further rounds due to severe concatemer formation during the PCR amplification step (unpublished data). As a result, ssDNA library R00 and ssDNA pools from round R02-R09 were subject to next-generation sequencing in order to assess the SELEX pool populations and to identify potential aptamer candidates.

SELEX Round	**Target Species**
R01	*E. faecalis*
R02	*E. faecalis*
R03	*E. faecium*
R04	*E. durans*
R05	*E. hirae*
R06	*E. faecalis*
R07	*E. faecium*
R08	*E. durans*
R09	*E. hirae*

Tbl. 12.: Targets used in the Toggle-SELEX approach of EF07

4.2.3.1. Results from workflow *selex-ngs-prep*

In total, 783,823 sequences were produced and successfully demultiplexed by the Illumina MiSeq platform (R00, R02-R09). The number of reads per SELEX round was sufficiently well distributed, indicating that library preparation and pooling of SELEX samples was performed properly, however, a higher number of reads was expected. The raw data showed that the Illumina MiSeq was not able to demultiplex 484,794 sequences. The plots in Figures 22a and 22b show the sequencing quality. Figure 22c shows the overall quality after preprocessing, to show an average quality of around 38. Sequences tended to be of higher quality than in EF01 and EF05. Just as in EF01 and EF05 a small amount of short sequences was in the data set. Also, compared to them, EF07 appeared to contain a lot more long length reads.

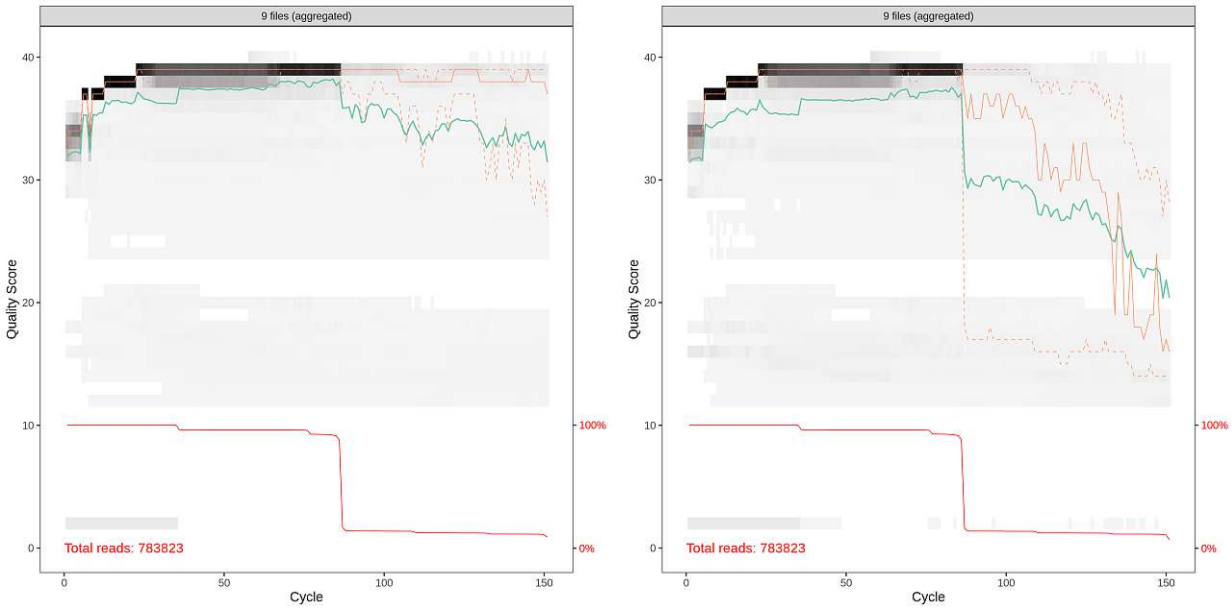
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	Total reads	Share in NGS run	Dev. from mean
R0	107638	13.73%	2,62 %
R2	89759	11.45%	0,34 %
R3	89697	11.44%	0,33 %
R4	88914	11.34%	0,23 %
R5	85909	10.96%	0,15 %
R6	96500	12.31%	1,20 %
R7	83246	10.62%	0,49 %
R8	57997	7.40%	3,71 %
R9	84163	10.74%	0.37%
	783823	100.00%	

Tbl. 13.: Sequenced reads for SELEX EF07 (R0-R9)

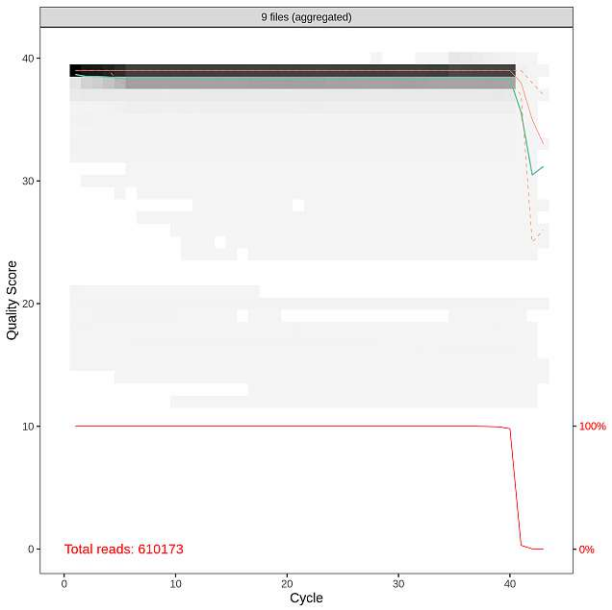
Table 14 and Tables 28, 30 and 29 in the appendix show the effect preprocessing had on the numbers of sequences. Figures 23a and 23b plot the preprocessing loss. The preprocessing parameters used were identical to the ones of SELEX experiment EF01, given in 4.2.1.1.
Overall, adapter trimming introduced an average loss of 10.38%, with R09 being an outlier with 21.14% loss. Quality filtering introduced an average loss of 7.72%, which was the lowest, compared to SELEX EF01 and SELEX EF05. Average paired-end read merging loss was low for rounds R00-R05 with 0.33% and started to increase with R6, rising to 16% in round R9. Therefore, the overall average loss for paired-end read merging was 4.38%. The loss stemming from random region length limitations was increasing as well starting with R7 to 3.25% for R9. The average loss was therefore 0.69%. The average length-limitation loss, when R7 to R9 were excluded, was 0.30%. The propagation of the error implies that the problem may stem from the SELEX experiment itself and not from sequencing.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) Forward Reads

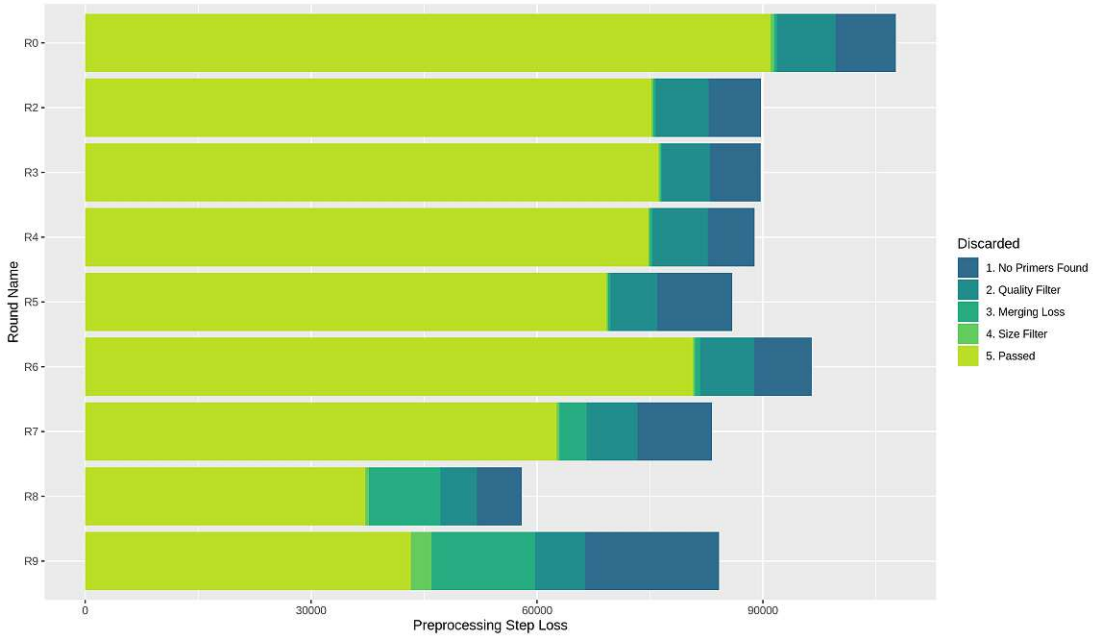
(b) Reverse Reads



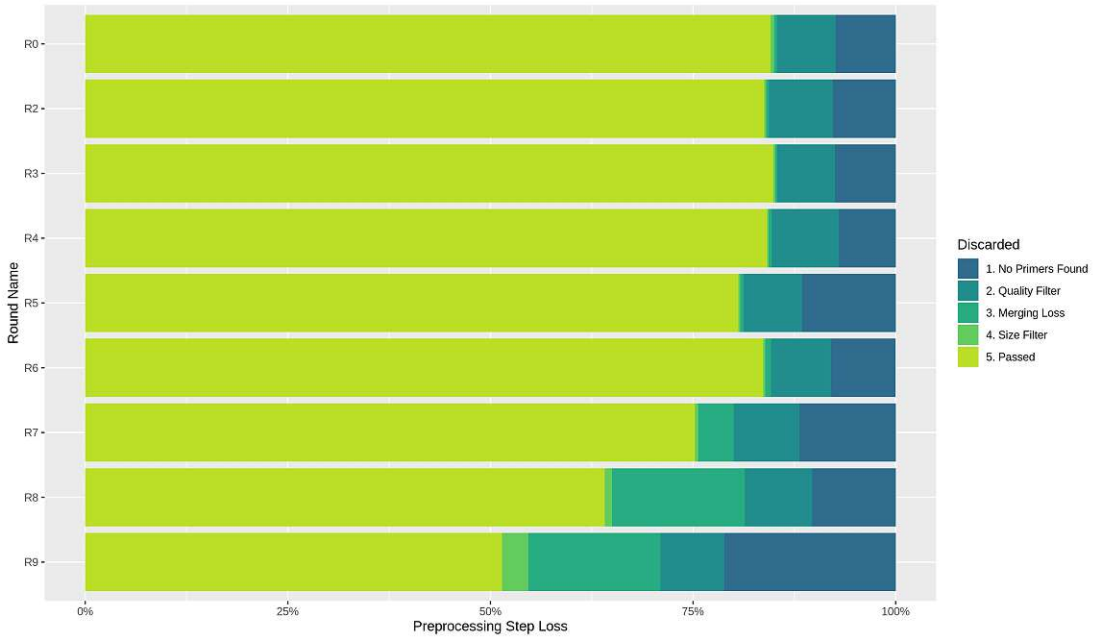
(c) Preprocessed Reads

Fig. 22: Sequencing quality profiles of EF07 raw and preprocessed reads. Median quality is plotted as a straight orange line, while the mean quality is plotted as a straight green line. The background shows a heatmap for the overall quality distribution. Dotted orange lines show the 25th and 75th percentile of quality. The straight red line shows the share of considered reads.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



(a) Absolute numbers



(b) Scaled to 100%.

Fig. 23: Loss of sequences after preprocessing steps in EF07.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Loss at	Trimming	Filter	PE-Merge	Length-limit
R0	7.42%	7.23%	0.34%	0.46%
R2	7.77%	7.87%	0.32%	0.25%
R3	7.49%	7.15%	0.22%	0.25%
R4	7.02%	8.27%	0.37%	0.21%
R5	11.56%	7.21%	0.40%	0.25%
R6	7.98%	7.39%	0.71%	0.28%
R7	11.88%	8.12%	4.36%	0.41%
R8	10.32%	8.30%	16.39%	0.88%
R9	21.14%	7.91%	16.31%	3.25%
Avg	10.29%	7.72%	4.38%	0.69%

Tbl. 14.: Share of discarded reads (in %) during every preprocessing step for SELEX EF07. Shares are taken for the original data set size.

Examination of Length/PE-merging Loss To assess the cause of the increased loss in the preprocessing the length-restriction script was modified to publish FASTA-files containing too long and too short sequences. The tool MEME 5.3.3[28] was used on the data set containing 429 too long sequences of round R9. MEME found a motif that strongly resembles the used reverse primer `TTGACTAGTACATGACCTCTTGA`, which is usually attached only once at the 3′-end of the sequences. Therefore, MEME was run with a maximum motif length of 23, and found that in 367 out of 429 sequences the motif was found (87.6%). As the reverse primer was already cut once in the trimming step, sequences containing the motif had the reverse primer attached at least twice and were therefore concatemers. The found motif logo can be seen in Figure 24.

To find the point in time at which the problem emerged, FIMO[80], included in MEME-Suite 5.3.3, was used to search for the motif in the other SELEX rounds data sets containing too long sequences. Table 15 shows that starting with SELEX round R2, one sequence was found to have two additional reverse primers attached. In R3 three sequences were found to have one additional primer attached, which were 12% of the discarded sequences in this step. R5 and R6 had an increase to 18 resp. 23 sequences containing primers with 24.7% resp.

Fig. 24: Motif logo of the enriched concatemer of SELEX EF07 round R9.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	Seqs	Seqs >43nt	%	Motifs found	Seqs w/ Motif	%	Motifs per Seq
R0	107638	11	0.01%	0	0	00.0%	
R2	89759	5	0.01%	2	1	20.0%	2.00
R3	89697	25	0.03%	3	3	12.0%	1.00
R4	88914	38	0.04%	0	0	00.0%	
R5	85909	73	0.08%	26	18	24.7%	1.44
R6	96500	100	0.10%	33	23	23.0%	1.43
R7	83246	176	0.21%	145	104	59.1%	1.39
R8	57997	177	0.31%	211	130	73.5%	1.62
R9	84163	429	0.51%	565	369	86.0%	1.53

Tbl. 15.: Increase of concatemers containing the motif logo representing the reverse primer. (Figure 24)

23.0% of all length-wise discarded sequences being concatemers. In round R7 an increase to 104 sequences (59%) were concatemers, which rose to 130 (73.45%) in R8, and to 369 (86.01%) in R9.

4.2.3.2. Results from workflow *selex-assess*

Nucleotide distribution and sequence enrichment in the cell-bound ssDNA pools of SELEX round R02-R11 were then determined using the workflow *selex-assess*. For visualization and simple interpretation, the workflow returns a series of tables and plots.
Figures 26a to 26c show the distribution of nucleotides at each position of the 40-nucleotide long random region in ssDNA pools of SELEX rounds R2, R5 and R9. Figure 26d shows the overall changes in nucleotide composition over the SELEX rounds.
The initial ssDNA library was resequenced for this run, and as described in SELEX EF01 and SELEX EF05, slightly biased towards sequences with elevated adenine(29.1%) and reduced cytosine nucleotide bases (22.1%).
Over the SELEX process, the nucleotide distribution changed by an increase of cytosine- and thymine-rich sequences of 7.6% and respectively 2.7% in SELEX round R6, to then decrease again (Figure 26d).
In Table 16, the top 25 reads encountered of the last round R09 of SELEX EF07 from the workflow *selex-assess* are given. The workflow *selex-assess* revealed that sequences were enriched over the course of the bacterial whole-cell SELEX EF07. From SELEX round R05 on, an increased number of sequences with > 10 reads were detected, while the proportion of unique sequences gradually decreased (Figure 25). Thus, NGS data confirmed the observations made by qPCR-based

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

remelting curve analyses, which suggested a decrease in pool diversity and an increase of enriched sequences.

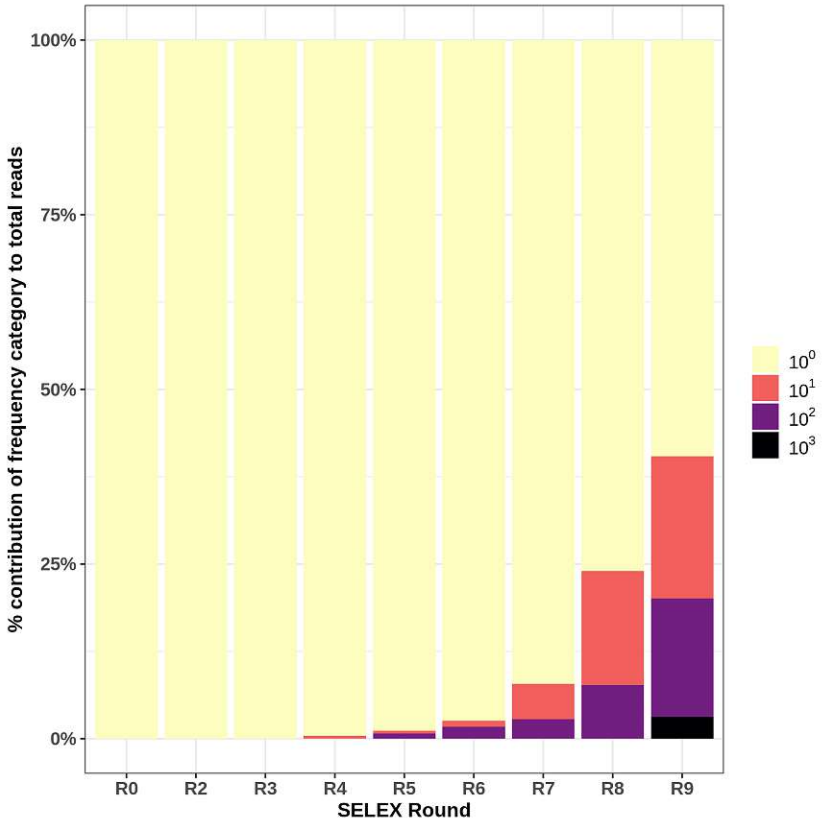
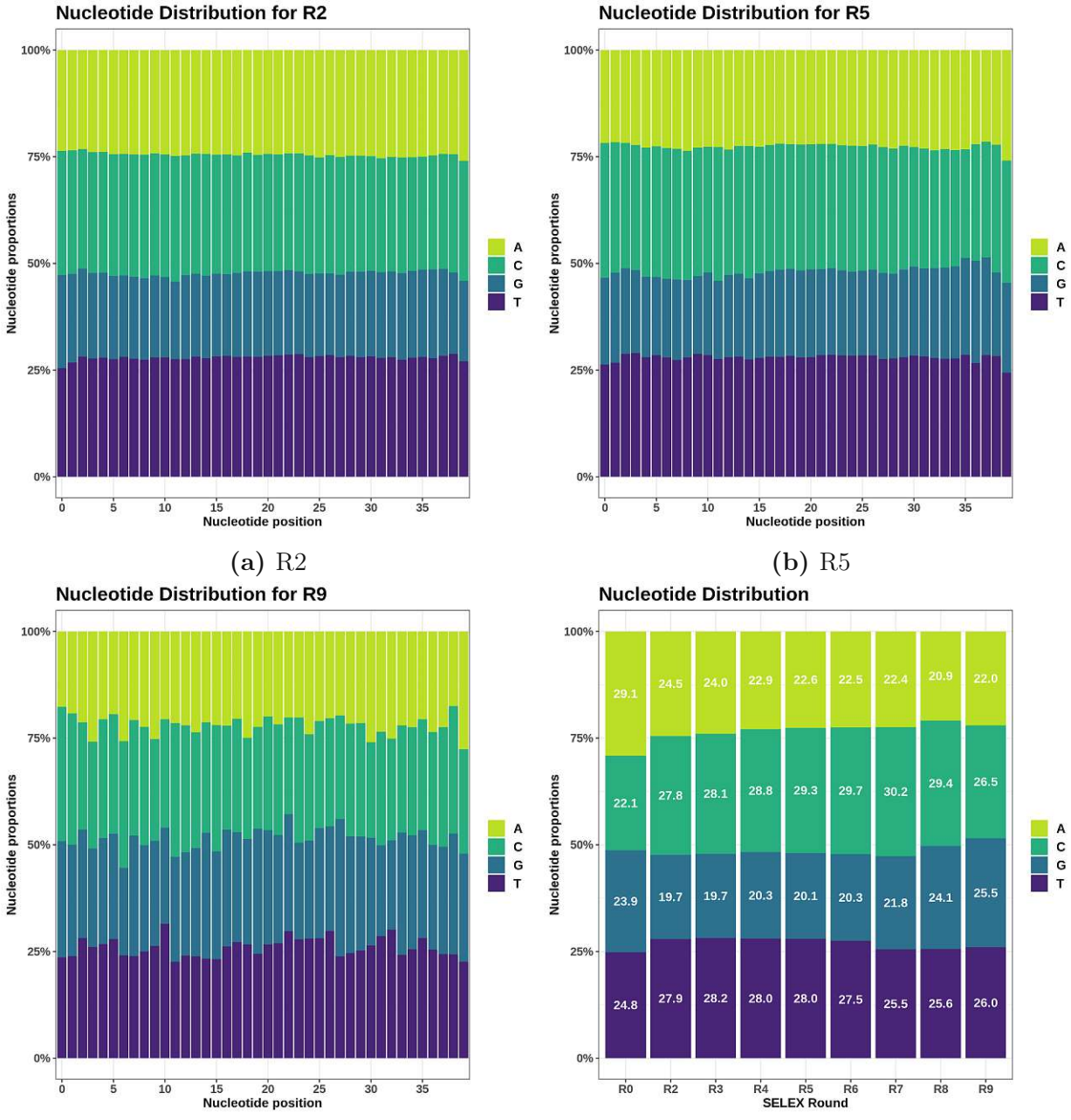


Fig. 25: Assessment of SELEX EF07 (R00-R09) in terms of sequence enrichment and frequency. Sequence enrichment can be observed. The darker a bar, the higher the replication number of the reads represented by them. Bars are as tall as the total read count of the sequences in them.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) R2 **(b)** R5

(c) R9 **(d)** Nucleotide composition dynamics of EF07

Fig. 26: Nucleotide distributions of EF07 at each position of the 40-nucleotide long random region of SELEX rounds R02(a), R07(b) and R09(c), as well as changes in the nucleotide composition in the random regions over the selection rounds(d).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Rank	Count	Random Region
1	1368	GCTATTCGCATCCGGCTGGGTGGCAGGGGGATTGGTAAGG
2	821	GTTTCGGGGGGGGGGGGAACACATTTGTGTAACAAACAGTC
3	626	GGGTGGTGGGTGGGGGGCAAGCTACTTGCCTATTTTCGTA
4	531	CTGAGACGGGTTCAGGGTGTGGTTGGAGGGAATGGGGCTG
5	464	GCGGGGAGAGGCGAAAGAAGCTGGGATGGAAGGGCGTAGG
6	460	TGACACACTTTCCTCAATTGGTACAATCACGTTCATCTCT
7	382	GTTTCGGGGGGGGGGGAACACATTTGTGTAACAAACAGTC
8	363	TTAATCATAATGACAGGGAGGGGGGTGGCGGTCGGCGGGT
9	351	CGAACCATAACCACGCCTAACATCTTCTTATACTCAGCGA
10	333	CGGGGGGTGGCAGCGCGGATACGTTTCCCTGCTTTGTGAC
11	317	AGCTCTTAGTTCGTTCGTACTCGCGGTTGGGGGGGGGCCG
12	280	CCGGCGTTTTTGAGGGGGTGGTGCTTGGGGGAACGGGCCA
13	274	AGCTCTTAGTTCGTTCGTACTCGCGGTTGGGGGGGGGCCG
14	219	GTTTCGGGGGGGGGGGGAACACATTTGTGTAACAAACAGTC
15	212	CGATTCATTTGTCAGCCTAATATTCCGAGGGACTAATGCT
16	201	CAAAGTCAAAAGACATACAGGCCGTCTGTACGTCCCTTCC
17	177	TGTATGAGCGGAAACTTCCTAGACCTGCATACAATCGCCT
18	167	TTCGGGGGGGTGGGGAATTCCATTGCTATGGTAGACTAAT
19	163	TACGCTTCGTTCCATTACTGCATGCAATGTACATACCTCA
20	145	TTCCGCAATTAGTTCCCCTAGTTACGCACGTCTAAATGTC
21	142	TACAGGCAGTGACACGACAAAGTATCCTACTTCTCGAGACG
22	141	CTTCAGACCGAAAATATGTGCATGGGGGACTTAATTTTGA
23	130	CCACATACCTGATTTGGTCGAAGTACTGCATGATTCTCCC
24	107	AGGCCTTGTTAAGATATCCTACATGTTTGTGTACTTAGGA
25	101	TGGCGTAACGGGTTGGGGGGGTGAGATATTCTAAGCATCT

Tbl. 16.: Top 25 Reads from last enrichment round of SELEX EF07.

4.2.3.3. Results from workflow *selex-blaster*

Motif detection was performed on a data set that combined all round files of SELEX EF07 (R00-R09) by using the workflow *selex-blaster*, to find the looping regions responsible for binding.
The workflow reported to have found 7.966 clusters. Tables 17 and 18 show the top three motif logos for the first 20 clusters. The discovered motifs had a higher bitscore compared to the ones in SELEX EF05.
The combined SELEX rounds dereplicated to 367.420 unique sequences, of which representative sequences were chosen for every mutational family cluster, resulting in a data set containing 357.254 sequences. The data set was masked after folding prediction and used to create a similarity graph using BLAST. The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

similarity graph created contained 52.999 sequences connected by 125.438 edges. Compared to SELEX EF05, the share of retained sequences was higher at around 15%. Clustering was done using an inflation factor of 1.2 for MCL. MEME (MEME-Suite 5.3.0) was used on all clusters that had at least 20 sequences (499 clusters).
graphicx

4.2.3.4. Results from workflow *selex-kmer*

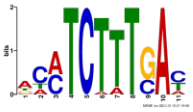
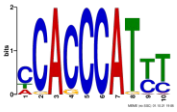
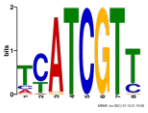
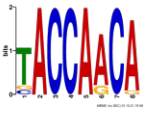
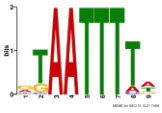
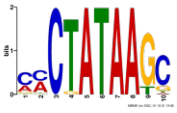
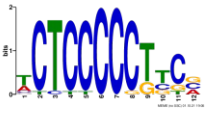
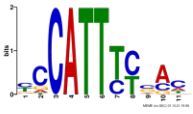
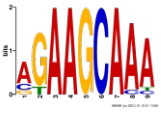
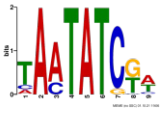
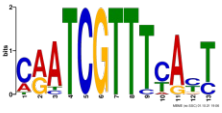
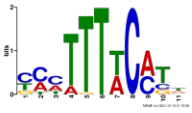
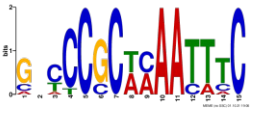
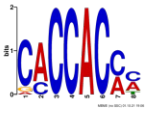
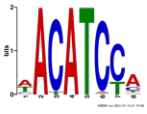
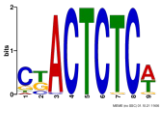
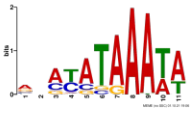
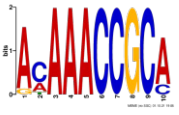
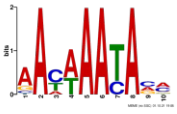
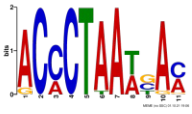
K-mer-based scoring was performed to estimate binding affinity of sequences, by using the workflow *selex-kmer* for 6-mers.
Scores were calculated for every combination of datasets including rounds R02-R09 and library R00. To avoid overrepresentation of k-mers, the data sets were dereplicated and filtered to only include representative sequences of mutational family clusters, similarly as described in the results of *selex-blaster*. The *selex-kmer* workflow returned k-mer based sequence scores for every round combination. Scores consisted of the overall aptamer score, the lowest and the highest shifting score. The shifting score was based on 5 consecutive k-mers of length 6, so a stretch of 10 characters.

Target-specific Shifting Scores An attempt was made to find sequences which may be specific to *E. faecalis* by maximizing k-mer shifting scores for that target and minimizing shifting scores for non-targets (see Table 12). Only sequences with a score difference below 5 for all targets were considered to increase confidence in predictions. Target specific scores were calculated for the remaining 8.870 sequences:

$$s_{E.faecalis} = \frac{s_{R2} + s_{R6}}{2} - \frac{s_{R3} + s_{R7}}{2} - \frac{s_{R4} + s_{R8}}{2} - \frac{s_{R5} + s_{R9}}{2} - |s_{R2} - s_{R6}|$$

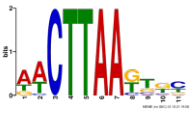
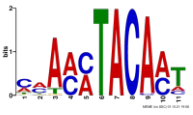
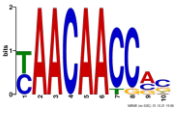
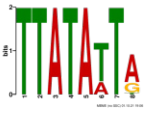
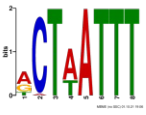
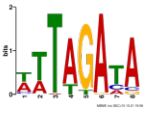
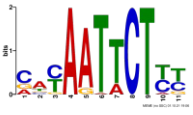
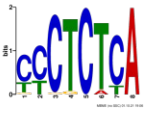
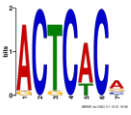
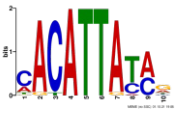
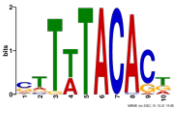
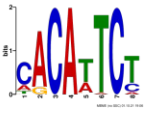
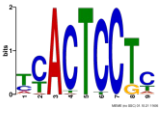
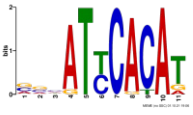
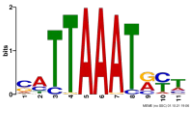
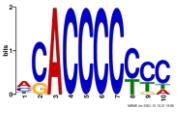
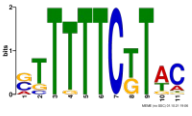
The difference between the target is subtracted to discourage too high differences. Table 19 shows the top 15 sequences ordered by the *E. faecalis* specific score. Table 20 shows motifs of the top 1000 sequences of *E. faecalis* targeted rounds, chosen using the target specific scores.
Similar analyses have been done for the other targets and are in the appendix. *E. faecium*: Tables 31 and 34. *E. durans*: Tables 32 and 35. *E. hirae*: Table 33 and 36.
The target specific shifting scores were calculated manually using an R-script. Target specific score calculation should be done automatically by the pipeline, if the pipeline is used for Toggle-SELEX, and should therefore be included in future updates. A summary of the top sequences for every target (in Toggle-SELEX) should be created.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Cluster	Seqs	Motif 1	#	Motif 2	#	Motif 3	#
c000000	294		26		26		33
c000001	261		19		29		20
c000002	248		49		22		20
c000003	233		47		26		15
c000004	223		59		33		13
c000005	221		46		69		10
c000006	217		97		15		10
c000007	214		70		35		25
c000008	214		39		58		11
c000009	207		75		18		22

Tbl. 17.: Motifs detected using MEME in clustered data sets of EF07, continued in Table 18. Motifs of E-value ≤ 0.05 are shown.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Cluster	Seqs	Motif 1	#	Motif 2	#	Motif 3	#
c000010	207		35		37		20
c000011	204		42		21		20
c000012	195		36		29		29
c000013	193		100		15		7
c000014	183		24		40		45
c000015	182		72		22		
c000016	176		46		41		12
c000017	165		48		26		18
c000018	165		43		39		
c000019	161		58		9		17

Tbl. 18.: Continuation of Table 17.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Sequence	*E. faecalis*	*E. faecium*	*E. durans*	*E. hirae*
CGAAATGAATTTTATAAAAGTCAATGTTAATTGTTAGAA	2.12	-3.61	-2.44	-3.54
GACAAATGTAGAAATGGCATAGGATTCTGTGAAATCGGAC	1.10	-5.16	-4.21	-3.75
GAACGATTGTAAGGAAATCATGGCAAGCATTTAGATTGAC	0.89	-5.43	-6.10	-7.33
CTGGATGTTACTCGAAAAACAGTGGCCATAATCGTAATCA	0.85	-4.57	-5.15	-5.30
ATTTATTTAGAATATGGAGTGGCCATAAACAAACATGGAC	0.79	-4.69	-4.75	-6.42
CTATGAAAATGTCAAACATGGCAATTACAAACAAATTCGC	0.76	-3.80	-4.20	-5.19
AGCATGTTACAAAGAAATTCAACAGTCGTTTTTATCGTTT	0.75	-3.84	-3.91	-3.31
TATGGGAAAAATCAGTTTTGTGACATCAAATATCAGCACT	0.72	-4.07	-4.88	-6.16
TTAAACAAAGACAAAAGAAAACCCAAAGTGGCCATGGATA	0.71	-4.26	-5.74	-6.37
TAGATCGAATAACATCGTGATAGATTCAACAAGGCATTAT	0.69	-4.50	-2.91	-4.13
CGGACATGGCTTAATATTCAAAACGAATGATTTGGTCTGA	0.69	-4.42	-4.02	-4.41
CCAAAATTACAGTCATGATAATTTAAAAACGGTATTGCTC	0.67	-6.76	-4.99	-6.96
TAGTTGAAATAGGTACAATGTTACCAAATAGTGGCCTAAC	0.67	-3.90	-3.67	-4.02
GATGATACCCAAGACGAAGTCAAACTCGGGAACTCTGAGG	0.65	-3.42	-3.47	-5.04
TCGGTACTATGACATAAATTTAGAAATATTTGTCGAGAGT	0.65	-5.17	-3.17	-4.65

Tbl. 19.: Sequences maximizing the k-mer shifting score for target *E. faecalis*.

Motif	E-value	Sites	Width
	2.2e-150	277	11
	2.8e-045	72	8
	3.0e-043	130	9
	2.4e-039	70	8
	6.2e-043	160	8
	3.5e-002	31	9

Tbl. 20.: Motifs found in the top 1000 sequences maximizing the k-mer shifting score for target *E. faecalis*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5. Conclusion and Outlook

Despite big strides in the last decade, the field of SELEX bioinformatics is still in an early stage of development. There is no 'one-fits-all' solution as the tools that can be used strongly depend on SELEX structure and the sought-after results. In this thesis project, automatic bioinformatic pipelines were developed to analyse in-house performed bacterial whole-cell SELEX experiments and to identify potential aptamer candidates. A major advantage of the developed pipelines over other current tools is the customizability as they can be easily extended with additional features by adding new processes in the Nextflow pipeline. The presented pipelines need no user-interaction and make data analysis results reproducible by using experiment specific configurations. Moreover, analysis results have proven useful for finding error sources and to support wetlab work (NGS library prep and SELEX protocols). They have shown that qPCR-based remelting curve analyses, which is done to check for enrichment, were also observable in the NGS data. The cause for the concatemer formation in SELEX EF07 was investigated, and it was shown that concatemers consisted of SELEX reverse primers.

The established pipelines *selex-ngs-prep* and *selex-assess* enable to preprocess NGS raw reads and assess ssDNA pool populations from various rounds of a SELEX experiment with regard to nucleotide composition, sequence enrichment and frequency-based ranking. Furthermore, first steps were made to establish structure-based clustering (*selex-blaster*) and k-mer-based aptamer scoring (*selex-kmer*) for Toggle-SELEX approaches. The pipeline *selex-blaster*, which was adapted from Song M. et al., 2019[41] was developed to perform clusterings. Compared to the original method, clustering is not done on the full length random regions, but instead based on looping regions, which were predicted using RNAfold[3]. Momentarily, *selex-blaster* could still could use some optimization due to ambiguity of the results by providing an overview of results for every cluster. To overcome the current limitations of the single structure-based clustering, adaptions using soft clustering on structure ensembles, similar to APTANI[31] and AptaCluster[27], with attention on looping regions could be implemented to make *selex-blaster* a feasible option for future SELEX experiments. The pipeline *selex-kmer* has shown potential for analyzing Toggle-SELEX experiments and calculates scores for short sequence stretches to highlight sequences containing well-enriched nucleotide stretches only

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

in a portion of their sequence. *Selex-kmer* has shown to be faster and more robust than the tool it was adapted from (MPBind[39]) and it would be interesting to add further k-mer-specific enrichment calculations. Both pipelines *selex-kmer* and *selex-blaster* should be benchmarked with other tools using HT-SELEX data sets including characterized aptamer sequences in the future. An interesting *in silico* approach, potentially enabling highly specific aptamer development, could be to combine single-step SELEX (Hoon et al.,2011[26]) with differential cell-SELEX (Meyer S. et al.,2013[58], Pleiko et al.,2019[59]), using multiple targets and target combinations of related bacterial species. Enrichment would be much smaller, compared to conventional SELEX procedures, thus biological and technical replicates would be required to make target-specific enrichment observable on the k-mer level. This could significantly reduce time investment needed for SELEX experiments and allow for the parallel development of aptamers specific to a variety of targets.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Bibliography

[1] Theodore R Allnutt et al. “Shortlisting aptamer candidates from HT-SELEX data”. In: *Aptamers* 2 (2018), pp. 36–44.

[2] Agata Levay et al. “Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment”. In: *Nucleic acids research* 43.12 (2015), e82–e82.

[3] Ronny Lorenz et al. “ViennaRNA Package 2.0”. In: *Algorithms for molecular biology* 6.1 (2011), pp. 1–14.

[4] Maciej Antczak et al. “New functionality of RNAComposer: application to shape the axis of miR160 precursor structure”. In: *Acta Biochimica Polonica* 63.4 (2016), pp. 737–744.

[5] Mariusz Popenda et al. “Automated 3D structure composition for large RNAs”. In: *Nucleic acids research* 40.14 (2012), e112–e112.

[6] Andrew B Kinghorn et al. “Aptamer bioinformatics”. In: *International journal of molecular sciences* 18.12 (2017), p. 2516.

[7] Craig Tuerk and Larry Gold. “Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase”. In: *science* 249.4968 (1990), pp. 505–510.

[8] Jinwei Zhang, Matthew W Lau, and Adrian R Ferré-D’Amaré. “Ribozymes and riboswitches: modulation of RNA function by small molecules”. In: *Biochemistry* 49.43 (2010), pp. 9123–9131.

[9] Matthew R Dunn, Randi M Jimenez, and John C Chaput. “Analysis of aptamer discovery and technology”. In: *Nature Reviews Chemistry* 1.10 (2017), pp. 1–16.

[10] Esther Frohnmeyer et al. “Highly affine and selective aptamers against cholera toxin as capture elements in magnetic bead-based sandwich ELAA”. In: *Journal of biotechnology* 269 (2018), pp. 35–42.

[11] David HJ Bunka et al. “Production and characterization of RNA aptamers specific for amyloid fibril epitopes”. In: *Journal of Biological Chemistry* 282.47 (2007), pp. 34500–34509.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[12] M Svobodova et al. “Selection of 2 F-modified RNA aptamers against prostate-specific antigen and their evaluation for diagnostic and therapeutic applications”. In: *Analytical and bioanalytical chemistry* 405.28 (2013), pp. 9149–9157.

[13] Yi Xi Wu and Young Jik Kwon. “Aptamers: The “evolution” of SELEX”. In: *Methods* 106 (2016), pp. 21–28.

[14] Claudia Kolm et al. “DNA aptamers against bacterial cells can be efficiently selected by a SELEX process using state-of-the art qPCR and ultra-deep sequencing”. In: *Scientific reports* 10.1 (2020), pp. 1–16.

[15] Tatjana Schütze et al. “Probing the SELEX process with next-generation sequencing”. In: *PloS one* 6.12 (2011), e29604.

[16] Andrew D Ellington and Jack W Szostak. “In vitro selection of RNA molecules that bind specific ligands”. In: *nature* 346.6287 (1990), pp. 818–822.

[17] Fabian Spill et al. “Controlling uncertainty in aptamer selection”. In: *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12076–12081.

[18] Min Young Song et al. “Broadly reactive aptamers targeting bacteria belonging to different genera using a sequential toggle cell-SELEX”. In: *Scientific reports* 7.1 (2017), pp. 1–10.

[19] Christian Schudoma et al. “Sequence–structure relationships in RNA loops: establishing the basis for loop homology modeling”. In: *Nucleic acids research* 38.3 (2010), pp. 970–980.

[20] Phuong Dao et al. “AptaTRACE elucidates RNA sequence-structure motifs from selection trends in HT-SELEX experiments”. In: *Cell systems* 3.1 (2016), pp. 62–70.

[21] Jan Hoinka et al. “Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers”. In: *Bioinformatics* 28.12 (2012), pp. i215–i223.

[22] Laia Civit et al. “Systematic evaluation of cell-SELEX enriched aptamers binding to breast cancer cells”. In: *Biochimie* 145 (2018), pp. 53–62.

[23] Michael Kohlberger and Gabriele Gadermaier. “SELEX: Critical factors and optimization strategies for successful aptamer selection”. In: *Biotechnology and Applied Biochemistry* (2021).

[24] Michael Blind and Michael Blank. “Aptamer selection technology and recent advances”. In: *Molecular Therapy-Nucleic Acids* 4 (2015), e223.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[25] Nam Nguyen Quang et al. “Time-lapse imaging of molecular evolution by high-throughput sequencing”. In: *Nucleic acids research* 46.15 (2018), pp. 7480–7494.

[26] Shawn Hoon et al. “Aptamer selection by high-throughput sequencing and informatic analysis”. In: *Biotechniques* 51.6 (2011), pp. 413–416.

[27] Jan Hoinka et al. “Aptacluster–a method to cluster ht-selex aptamer pools and lessons from its application”. In: *International Conference on Research in Computational Molecular Biology*. Springer. 2014, pp. 115–128.

[28] Timothy L Bailey et al. “The MEME suite”. In: *Nucleic acids research* 43.W1 (2015), W39–W49.

[29] Khalid K Alam, Jonathan L Chang, and Donald H Burke. “FASTAptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections”. In: *Molecular Therapy-Nucleic Acids* 4 (2015), e230.

[30] Michael Hiller et al. “Using RNA secondary structures to guide sequence motif finding towards single-stranded regions”. In: *Nucleic acids research* 34.17 (2006), e117–e117.

[31] Jimmy Caroli, Mattia Forcato, and Silvio Bicciato. “APTANI2: update of aptamer selection through sequence-structure analysis”. In: *Bioinformatics* 36.7 (2020), pp. 2266–2268.

[32] Shermin Pei, Betty L Slinger, and Michelle M Meyer. “Recognizing RNA structural motifs in HT-SELEX data for ribosomal protein S15”. In: *BMC bioinformatics* 18.1 (2017), pp. 1–14.

[33] Natalia Komarova, Daria Barkova, and Alexander Kuznetsov. “Implementation of High-Throughput Sequencing (HTS) in Aptamer Selection Technology”. In: *International Journal of Molecular Sciences* 21.22 (2020), p. 8774.

[34] Jan Hoinka, Rolf Backofen, and Teresa M Przytycka. “AptaSUITE: a full-featured bioinformatics framework for the comprehensive analysis of aptamers from HT-SELEX experiments”. In: *Molecular Therapy-Nucleic Acids* 11 (2018), pp. 515–517.

[35] Kevin R Shieh et al. “AptCompare: optimized de novo motif discovery of RNA aptamers via HTS-SELEX”. In: *Bioinformatics* 36.9 (2020), pp. 2905–2906.

[36] Timothy L Bailey et al. “MEME SUITE: tools for motif discovery and searching”. In: *Nucleic acids research* 37.suppl_2 (2009), W202–W208.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[37] Timothy L Bailey. “STREME: Accurate and versatile sequence motif discovery”. In: *Biorxiv* (2020).

[38] Jimmy Caroli et al. “APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data”. In: *Bioinformatics* 32.2 (2016), pp. 161–164.

[39] Peng Jiang et al. “MPBind: a Meta-motif-based statistical framework and pipeline to Predict Binding potential of SELEX-derived aptamers”. In: *Bioinformatics* 30.18 (2014), pp. 2665–2667.

[40] Stijn Marinus Van Dongen. “Graph clustering by flow simulation”. PhD thesis. 2000.

[41] Jia Song et al. “A sequential multidimensional analysis algorithm for aptamer identification based on structure analysis and machine learning”. In: *Analytical chemistry* 92.4 (2019), pp. 3307–3314.

[42] Robert C Edgar. “UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing”. In: *BioRxiv* (2016), p. 081257.

[43] Jacob T Nearing et al. “Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches”. In: *PeerJ* 6 (2018), e5364.

[44] Robert C Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19 (2010), pp. 2460–2461.

[45] Mark A Ditzler et al. “High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase”. In: *Nucleic acids research* 41.3 (2013), pp. 1873–1884.

[46] Ryoga Ishida et al. “RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information”. In: *Nucleic acids research* 48.14 (2020), e82–e82.

[47] Benjamin J Callahan et al. “DADA2: high-resolution sample inference from Illumina amplicon data”. In: *Nature methods* 13.7 (2016), pp. 581–583.

[48] Qing Yu et al. “Selection and characterization of ssDNA aptamers specifically recognizing pathogenic Vibrio alginolyticus”. In: *Journal of fish diseases* 42.6 (2019), pp. 851–858.

[49] Shixi Song et al. “Selection of highly specific aptamers to Vibrio parahaemolyticus using cell-SELEX powered by functionalized graphene oxide and rolling circle amplification”. In: *Analytica chimica acta* 1052 (2019), pp. 153–162.

[50] Jennifer Soundy and Darren Day. “Selection of DNA aptamers specific for live Pseudomonas aeruginosa”. In: *PLoS One* 12.9 (2017), e0185385.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[51] Soledad Marton et al. “Isolation of an aptamer that binds specifically to E. coli”. In: *PloS one* 11.4 (2016), e0153637.

[52] Padma Sudha Rani Lavu et al. “Selection and characterization of aptamers using a modified whole cell bacterium SELEX for the detection of Salmonella enterica serovar typhimurium”. In: *ACS combinatorial science* 18.6 (2016), pp. 292–301.

[53] Jihea Moon et al. “Comparison of whole-cell SELEX methods for the identification of Staphylococcus aureus-specific DNA aptamers”. In: *Sensors* 15.4 (2015), pp. 8884–8897.

[54] Nasa Savory et al. “Selection of DNA aptamers against uropathogenic Escherichia coli NSM59 by quantitative PCR controlled Cell-SELEX”. In: *Journal of microbiological methods* 104 (2014), pp. 94–100.

[55] Abdullah Tahir Bayraç and Sultan Ilayda Donmez. “Selection of DNA aptamers to Streptococcus pneumonia and fabrication of graphene oxide based fluorescent assay”. In: *Analytical biochemistry* 556 (2018), pp. 91–98.

[56] Zhaofeng Luo et al. “Developing a combined strategy for monitoring the progress of aptamer selection”. In: *Analyst* 142.17 (2017), pp. 3136–3139.

[57] AS Davydova et al. “In vitro selection of cell-internalizing 2-modified RNA aptamers against Pseudomonas aeruginosa”. In: *Russian Journal of Bioorganic Chemistry* 43.1 (2017), pp. 58–63.

[58] Susanne Meyer et al. “Development of an efficient targeted cell-SELEX procedure for DNA aptamer reagents”. In: *PLoS One* 8.8 (2013), e71798.

[59] Karlis Pleiko et al. “Differential binding cell-SELEX method to identify cell-specific aptamers using high-throughput sequencing”. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[60] Mayumi Takahashi et al. “High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency”. In: *Scientific reports* 6.1 (2016), pp. 1–14.

[61] William H Thiel et al. “Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection”. In: (2012).

[62] Masaki Takahashi. “Aptamers targeting cell surface proteins”. In: *Biochimie* 145 (2018), pp. 63–72.

[63] Paolo Di Tommaso et al. “Nextflow enables reproducible computational workflows”. In: *Nature biotechnology* 35.4 (2017), pp. 316–319.

[64] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* 17.3 (2020), pp. 261–272.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[65] Peter JA Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[66] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[67] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[68] R Core Team et al. “R: A language and environment for statistical computing”. In: (2013).

[69] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7. 2021. URL: https://CRAN.R-project.org/package=dplyr.

[70] Hadley Wickham. *tidyr: Tidy Messy Data*. R package version 1.1.3. 2021. URL: https://CRAN.R-project.org/package=tidyr.

[71] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[72] Adrian A Dragulescu, Maintainer Adrian A Dragulescu, and R Provide. “Package ‘xlsx’”. In: *Cell* 9.1 (2020).

[73] Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, 2017.

[74] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1 (2011), pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

[75] Shifu Chen et al. “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17 (Sept. 2018), pp. i884–i890. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty560. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/17/i884/25702346/bty560.pdf. URL: https://doi.org/10.1093/bioinformatics/bty560.

[76] Stephen F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[77] David H Mathews et al. “Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure”. In: *Proceedings of the National Academy of Sciences* 101.19 (2004), pp. 7287–7292.

[78] Isabella Cervenka. *Selection of DNA-aptamers for Enterococcus faecalis*. eng. Wien, 2020.

[79] Jan Hoinka et al. “Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery”. In: *Nucleic acids research* 43.12 (2015), pp. 5699–5707.

[80] Charles E Grant, Timothy L Bailey, and William Stafford Noble. “FIMO: scanning for occurrences of a given motif”. In: *Bioinformatics* 27.7 (2011), pp. 1017–1018.

[81] Timothy L Bailey, Charles Elkan, et al. “Fitting a mixture model by expectation maximization to discover motifs in bipolymers”. In: (1994).

[82] Timothy L Bailey and Charles Elkan. “Unsupervised learning of multiple motifs in biopolymers using expectation maximization”. In: *Machine learning* 21.1 (1995), pp. 51–80.

A. Appendix

A.1. Nextflow

Nextflow offers a domain specific language to allow for dynamic pipeline building. Its syntax is based on Groovy, a Java-derived scripting language. Nextflow can be used to easily combine data driven processes for preparation, analysis or visualization.

A.1.1. Nextflow Processes

In Nextflow, multiple analysis tasks can be linked using processes with their input and output channels. The core part of a process is the embedded script. By default, the script is executed as Bash, however, it is also possible to execute scripts written in other languages, making it possible to directly embed R or python in a Nextflow process.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Input and output channels can be used to direct data to processes in a pipeline-like way. These channels can contain either files, values, or tuples of various combinations.
Nextflow processes can be decorated with directives to change their execution behaviour. Process directives can be used for a variety of process adjustments. For instance, they can be used to change the number of CPUs to use, the size of available working memory, time constraints, whether to retry failed processes, where to publish finished analysis files, or even whether the process should be executed on a remote cluster computer. The example process A.1 uses a process directive to create or check for a conda environment containing the bioinformatic analysis tool 'fastp' and also allows only for one CPU per process fork.
A process in Nextflow always consists of a unique name, an input channel and a script part.

```
/**
 * Example for printing out all text files in the current
 * directory
 */

notes = Channel.fromPath( '*.txt' ) // put txt files in channel '
    notes'

/**
 * process cat will call the UNIX command 'cat' to print the
 * content of every file in the channel 'notes'
 */
process cat {
  echo true // print process output to console
  cpus 1    // use one cpu only
  maxForks 15 // allow 4 concurrent processes

  input:
    file note_file from notes
  script:
  """
    cat ${note_file}
  """
}

```

Listing A.1. Rudimentary Nextflow example

A.1.2. Conda Integration

Nextflow workflows offer the possibility of importing required software and libraries utilizing the package manager conda. Tools and software are installed to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

temporary conda environments by using the *conda* directive as seen in Listing A.2. Conda was useful to decrease configuration and installation overhead.

```
/**
 * example process using fastp to filter files based on
 * quality
 */
process fastp {
  conda 'bioconda::fastp'
  input:
    file f_raw from fasta_files
  output:
    file f_filtered into fasta_files_filtered
  script:
  """
    fastp -i ${f_raw} --average_qual 30 -o ${f_filtered}
  """
}

```

Listing A.2. Conda integration in Nextflow

A.1.3. Configuration and Execution

Nextflow uses a clear text file, called 'nextflow.config' by default, to save configuration parameters. Parameters in the config files can be accessed globally across the pipeline. Using different configurations allows for easy modifications of the pipeline execution behavior. Configuration files can also be used to define different execution profiles. Profiles can be used to give different instructions, depending on whether the pipeline is run on a remote cluster, locally or for instance with a different number of CPUs. Whole workflows can be run on cluster computers, as tried on the Vienna Scientific Cluster. Nextflow also offers the possibility to only run subprocesses in cluster environments, making it feasible to outsource demanding tasks to more powerful machines.
Before workflow execution can be started, a valid configuration file must exit as well as all required input files. Workflows can be executed from the command line using the commands shown in Listing A.3. Nextflow stores temporary intermediate files in the directory 'work', which can be deleted after execution.

```
# run pipeline for the first time for SELEX experiment
# slx1.
nextflow run selex-blaster.nf -config slx1.config

# continue pipeline from where we left off (e.g. if changes
# were made to the script, to not start over).
```


Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

```
nextflow run pipeline_name.nf -config slx1.config -resume

```

Listing A.3. Execution of Nextflow Scripts

A.2. Clustering using MCL

The original idea behind the MCL algorithm[40] lies in random walks on weighted undirected graphs. Nodes represent aptamer sequences, while edges represent similarity. A walker starts at a random node in the graph, then takes an edge to walk to another node based on edge weight. The walker walks a number of steps until it stops and the process is repeated. Clusters are discovered, as the edges of a family will be traversed more often due to higher similarity. The MCL algorithm adapts this principle in a flow-like approach, alternating expansion and inflation steps.

A.3. Motif Detection using MEME

The tool MEME is based on an adaption of the EM-algorithm (Expectation-Maximization), which is a common method for data clustering. In the EM-algorithm many similarities to the k-means algorithm are apparent. However, every cluster's center point is calculated based on the probability of a data point to belong to the cluster. Therefore, the EM-algorithm algorithm performs soft clustering, in which data points are not exclusively assigned to a cluster.
Lawrence and Reilly[81] have adapted the EM-algorithm to be used for motif detection on sequences. They assumed every sequence to contain the sought motif once. As the motif is unknown, a guess is made on the position where it could be in every sequence. For every found subsequence a probability is calculated, whether it could be a motif. Based on these subsequences and their probabilities the motif model is updated. Probability calculations and choosing of new starting points can be repeated multiple times until the changes in the motif model are insignificant. In the tool MEME, Bailey and Elkan[82] extended their algorithm to also work on sequences containing no motifs or multiple motifs. MEME runs the EM-algorithm for only step, using different motif models as starting points. Then of these runs, an initial model, maximizing a consensus score, is chosen. This model is then used to run the EM-algorithm to convergence. The goal is to move away from local maxima to finding global maxima.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

A.4. EF01

Remaining	Total reads	Trimmer	Filter	PE-merging	Length-limit
R0	222414	207355	182569	181940	181085
R2	220753	189641	165524	164799	164375
R4	288262	254650	223738	222910	222376
R5	180032	158700	138215	137660	137374
R6	243162	217638	190697	190007	189599
R7	210076	188977	165381	164818	164497
R8	171799	152033	51980	51714	51532
R9	206774	183250	156120	155443	154969
	1743272	1552244	1274224	1269291	1265807

Tbl. 21.: Number of reads of SELEX EF01 round R0-R9 after the various preprocessing steps.

Discarded	Trimming	Filter	PE-Merge	Length-limit
R0	15059	24786	629	855
R2	31112	24117	725	424
R4	33612	30912	828	534
R5	21332	20485	555	286
R6	25524	26941	690	408
R7	21099	23596	563	321
R8	19766	100053	266	182
R9	23524	27130	677	474
	191028	278020	4933	3484

Tbl. 22.: Number of discarded reads of every preprocessing step for EF01.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Remaining	Total reads	Trimmer	Filter	PE-merging	Length-limit
R0	100.00%	93.23%	82.09%	81.80%	81.42%
R2	100.00%	85.91%	74.98%	74.65%	74.46%
R4	100.00%	88.34%	77.62%	77.33%	77.14%
R5	100.00%	88.15%	76.77%	76.46%	76.31%
R6	100.00%	89.50%	78.42%	78.14%	77.97%
R7	100.00%	89.96%	78.72%	78.46%	78.30%
R8	100.00%	88.49%	30.26%	30.10%	30.00%
R9	100.00%	88.62%	75.50%	75.18%	74.95%

Tbl. 23.: Read share remaining in EF01 after every preprocessing step.

Rank	Count	Random Region
1	33	TGACTAGTACATGACCATAGGGAAGAGAAGGACATATGA
2	26	GTGGCAGGTTACCCGAGAACCGAACCATACTCTCTCCCG
3	19	GTACAACACCTCAATAAGTCCGCGATAACGCGCAACAGTA
4	11	TTAGCCCCCCGAACCTCACTCACACATCTGCATACACTTT
5	10	CCGAGCTCCTGATGTGACGTCGGACTTCTTGACCACCG
6	9	CGGCCGCCCCAAGTCTCGTATATAGTCCCAACGCCTACAA
7	7	CATCTCCGGCTTGGCACCATCACCCAGACACACCACTAAT
8	7	TAGTTCCAAGGCAGCCCACCCCTACCCTCTCTCGACTCTA
9	7	ACGACACACACCACTCCATCTCCGCCGTCTCCTGCCAGCC
10	7	CACTCTCTCAGAAGCCAGCATCCCGCTCCACCTTTCGCCC
11	7	GCATTTGCGCTTACATCCAACGACTGTATACCTCGGACAC
12	7	CGTGAGCTGCAGTTATCGTTAAACTGGTACCAATTCGTTT
13	7	TGACTAGTACATGACCACTAGGGAAGAGAAGGACATATGA
14	7	TTACCAAATTCTGCACATCACCTCCACACCCGGCCGGCTG
15	7	GGATGAATCAGCGGGGCCGGGTAAGCAGATATGAGGACTC
16	7	CCCCTGTAGTTAGCCCCACACTATCTTGCTCTTTCTCACT
17	7	CATAACCACCGATAGTTTCTTAAGTACCCATCGATTCTTT
18	7	TCGCCATCGAGAACCTGATCATTCGAATTAGCTAAGGAGT
19	7	ACCCCCTGTACCCCCCTCCCAGCCGTAGCACGCCCCATGA
20	7	ACCTTGCATCAATTTACACGGCGACACAATCCCCACCCAA
21	6	GCCCGAGCTCCTGATGTGACGTCGGACTTCTTGACCACCG
22	6	CACCCTCCACGCTAGTACCACCCCACCTCCAGGCTATCCC
23	6	ACCGCCACATACTCACACTATGCCCACGAAACCAACCCTT
24	6	TATCGAAACAACACCCATACACCCAGCCTGATACCAACTC
25	6	TCGTCGTTAACTGAAATTTCCAGCGTTTTGGCAGATGTTG

Tbl. 24.: Top 25 Reads from last enrichment round of SELEX EF01.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

A.5. EF05

Remaining	Total reads	Trimmer	Filter	PE-merging	Length-limit
R02	342616	307863	235186	234321	233769
R03	330387	304505	231689	230803	230346
R04	258891	239157	183432	182908	182567
R05	251043	217770	164329	163804	163522
R06	299303	266687	206409	205787	205483
R07	253449	222924	172912	172435	172210
R08	308397	277222	206811	206056	205754
R09	247589	213555	161909	161412	161187
R10	317454	290747	224628	223704	223392
R11	339165	306175	238504	237347	237027
	2291675	2049683	1562677	1557526	1554838

Tbl. 25.: Number of reads of SELEX EF05 round R02-R11 after the various preprocessing steps.

Discarded	Trimming	Filter	PE-Merge	Length-limit
R02	34753	72677	865	552
R03	25882	72816	886	457
R04	19734	55725	524	341
R05	33273	53441	525	282
R06	32616	60278	622	304
R07	30525	50012	477	225
R08	31175	70411	755	302
R09	34034	51646	497	225
R10	26707	66119	924	312
R11	32990	67671	1157	320
	191028	278020	4933	3484

Tbl. 26.: Number of discarded reads of every preprocessing step for EF05.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Remaining	Total reads	Trimmer	Filter	PE-merging	Length-limit
R02	100.00%	89.86%	68.64%	68.39%	68.23%
R03	100.00%	92.17%	70.13%	69.86%	69.72%
R04	100.00%	92.38%	70.85%	70.65%	70.52%
R05	100.00%	86.75%	65.46%	65.25%	65.14%
R06	100.00%	89.10%	68.96%	68.76%	68.65%
R07	100.00%	87.96%	68.22%	68.04%	67.95%
R08	100.00%	89.89%	67.06%	66.82%	66.72%
R09	100.00%	86.25%	65.39%	65.19%	65.10%
R10	100.00%	91.59%	70.76%	70.47%	70.37%
R11	100.00%	90.27%	70.32%	69.98%	69.89%

Tbl. 27.: Read share remaining in EF05 after every preprocessing step.

A.6. EF07

Remaining	Total reads	Trimmer	Filter	PE-merging	Length-limit
R0	107638	99656	91876	91513	91017
R2	89759	82782	75718	75432	75206
R3	89697	82979	76565	76369	76147
R4	88914	82674	75318	74990	74802
R5	85909	75976	69783	69440	69225
R6	96500	88799	81663	80981	80713
R7	83246	73357	66599	62966	62621
R8	57997	52014	47199	37691	37182
R9	84163	66373	59719	45996	43260
	699660	638237	584721	569382	566913

Tbl. 28.: Number of reads of SELEX EF07 round R0-R9 after the various preprocessing steps.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Discarded	Trimming	Filter	PE-Merge	Length-limit
R0	7982	7780	363	496
R2	6977	7064	286	226
R3	6718	6414	196	222
R4	6240	7356	328	188
R5	9933	6193	343	215
R6	7701	7136	682	268
R7	9889	6758	3633	345
R8	5983	4815	9508	509
R9	17790	6654	13723	2736
	61423	53516	15339	2469

Tbl. 29.: Number of discarded reads of every preprocessing step for EF07.

Remaining	Total reads	Trimmer	Filter	PE-merging	Length-limit
R0	100.00%	92.58%	85.36%	85.02%	84.56%
R2	100.00%	92.23%	84.36%	84.04%	83.79%
R3	100.00%	92.51%	85.36%	85.14%	84.89%
R4	100.00%	92.98%	84.71%	84.34%	84.13%
R5	100.00%	88.44%	81.23%	80.83%	80.58%
R6	100.00%	92.02%	84.62%	83.92%	83.64%
R7	100.00%	88.12%	80.00%	75.64%	75.22%
R8	100.00%	89.68%	81.38%	64.99%	64.11%
R9	100.00%	78.86%	70.96%	54.65%	51.40%

Tbl. 30.: Read share remaining in EF07 after every preprocessing step.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Sequence	*E. faecalis*	*E. faecium*	*E. durans*	*E. hirae*
TGGTTTATGATACGCGGGAAGTAGGCGTGAAGGATAAAAG	-8.98	2.24	-1.48	-0.60
AGACAGAAACGCGAACTAGAGGATATATGGTTATTTGCGA	-5.96	2.14	-2.29	-2.13
ATGAAATTAATAACTAGGTAACAAATCGATAGTCGAAGAC	-3.75	1.93	-1.66	0.72
TGTGAGAGATGAGTATCGAGAAGGATAAAGTTTTAAGTAA	-7.29	1.92	-1.71	-0.92
TCGCAGATATGAAACAATAATGATAAACAATAGCGTGGGG	-8.66	1.86	-0.43	0.48
TAGAATTTATAGATTTGAGAACAAGTTGAAAAACGGGAAC	-8.84	1.78	0.09	1.14
TTGATATGGAGTAAACCAATAGACTAGTAATAAACCAGC	-3.90	1.77	-1.91	-1.22
TGCAAGCACTGAGAGGGACGACATATGGGCATAGTTGT	-5.36	1.68	-4.45	-4.67
AAAATGAGATTGGGTAGAAGGTATAACAATAGCAAGTATC	-9.68	1.66	-1.14	-1.06
ATAGAACACGCGAGCAGTTTAGATATAAATTGGACATTTC	-9.04	1.58	-3.85	-4.78
AAAAACTGCAAACGCGGGTATCGAAAGAGAATAACACAAC	-5.94	1.56	-3.27	-1.38
GAGTAGAATGATTAAGAACTGAACGGAATACAATAGTGGC	-7.65	1.55	-0.19	0.05
TGGTTAACGCGGCTAGGTACAGGAGTTATAAACAGAGATA	-8.46	1.52	-1.14	-0.34
ACTGAAACAACGAGCATATGCAAACAGTTATTTGGAGAAA	-5.34	1.51	-3.55	-2.32
GTAAACGTAGGCAAGTAGACTAGAGATGATACAGTAAATA	-7.93	1.50	-1.00	1.11

Tbl. 31.: Sequences maximizing the k-mer shifting score for target *E. faecium*.

Sequence	*E. faecalis*	*E. faecium*	*E. durans*	*E. hirae*
GATGGGTGGTGTACGGAAAAATAGTTGAAGATGCAAGAGC	-10.97	-4.95	2.73	-4.08
CTACATGGATAGGAAAGTTGGTGGTGCAAAGGGATTTACA	-10.30	-4.03	2.52	-4.58
ATGCGATATAAGGCCGGAGGTAAAGGCTGGAAGACATAGG	-11.07	-3.30	2.19	-2.98
GTGATGGTGGTCGAATAAGAAGTTATTATTAGTTGACTAC	-8.97	-3.43	2.16	-3.45
TTCGTTAGGATGATTATGTGGTGGTAAAGGGTAAGAAATT	-7.44	-4.32	1.72	-4.89
GATAGCGTATTGCAGGTGAATGTTGGTGGTGGATGAGGTC	-7.66	-4.93	1.48	-5.57
GTCAAAGTTTGATTAGGTATGTGGTGGTTTGTGCGAATCT	-5.79	-4.67	1.39	-5.03
GCATACGAATAGGAAAGCGATGTGAAGCGAGATAATGTAG	-8.67	-0.12	1.36	-0.18
ATAGGTATATGTGAGGTGTGTGGAATTTAAGAATGTTACC	-8.82	-2.86	1.33	-3.53
TTTGCGTGAATAAATATTGAAGTGAAGAATGAGTGTAAAG	-9.35	0.69	1.29	0.29
GCAAAAGTAAGTGTTAGGAATAAATATAATAAGCTACACA	-6.89	0.18	1.29	1.08
ATACGGGTGGTTTGTAAACGATGTAAGCAATGGAACACTC	-6.98	-4.02	1.28	-5.19
ATATTTTAAAACAAATTAGTTGGAAACGAGAACCGGGGCA	-8.30	-1.61	1.25	-2.56
GGAACTAAGATACCGGAAATATTATAACTAGGGAACCCCA	-8.68	-0.64	1.21	-1.53
CGACAATAAAATGGAAGCAATGCAACGGGGTCGGTTGAGA	-8.49	-1.60	1.15	-1.20

Tbl. 32.: Sequences maximizing the k-mer shifting score for target *E. durans*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sequence	*E. faecalis*	*E. faecium*	*E. durans*	*E. hirae*
GCTAGGATAGAAATTGAATTATTTAATGTACGAGTACAAA	-2.67	-0.24	-1.34	2.29
TCGAAGGAATACAAGCGTAAAACGATTAAAAAAAGACGTA	-6.15	1.48	0.65	2.07
CATAAAGGTCGAAAGTAGAAGTATATTAAAACAAATTAGC	-7.27	0.48	-0.57	1.51
CTAATGAAACGCAATAAGCGAATGTTTAAGATCACGT	-4.31	-0.17	-0.77	1.40
GAGAGATATAGGCAAAAGCGGTAAAGAATGTATAGATTTC	-8.40	0.59	0.30	1.38
GAAAATAAATGCAAACTGAGTAAAGCAAAGAAGTAAA	-6.18	-0.20	-0.77	1.37
GGTAAGAATTTAAGTATATTAGAAGAGTACAACGAAGATC	-7.85	0.48	0.14	1.35
CTGCGAGGTATTGAATAGAAGCAAAACAGCTAAAGAACAG	-9.49	-0.16	-1.19	1.33
CCAAAACAAGAGTAATGTGAAAGATACGTAGAGGTCCATG	-7.35	0.54	0.70	1.15
TAGAATTTATAGATTTGAGAACAAGTTGAAAAACGGGAAC	-8.84	1.78	0.09	1.14
GTAAACGTAGGCAAGTAGACTAGAGATGATACAGTAAATA	-7.93	1.50	-1.00	1.11
AGTGTAGGGATAGCAAAGGGTATTAGTAGGCAAATTTGGC	-8.08	1.31	-1.35	1.10
CGTGCGAAACAAAATTAGATGAAAATGAAGGTCGAACAGG	-5.97	-0.50	-1.17	1.09
CGCAGATTTAGTAGGGCCAGAAACAATATAAGGATGTAGG	-8.82	-0.41	-0.73	1.09
GCAAAAGTAAGTGTTAGGAATAAATATAATAAGCTACACA	-6.89	0.18	1.29	1.08

Tbl. 33.: Sequences maximizing the k-mer shifting score for target *E. hirae*.

Motif	E-value	Sites	Width
	7.6e-018	93	8

Tbl. 34.: Motifs found in the top 1000 sequences maximizing the k-mer shifting score for target *E. faecium*.

Motif	E-value	Sites	Width
	2.1e-031	135	8
	5.8e-012	85	8

Tbl. 35.: Motifs found in the top 1000 sequences maximizing the k-mer shifting score for target *E. durans*.

Motif	E-value	Sites	Width

Tbl. 36.: No motifs were found in the top 1000 sequences maximizing the k-mer shifting score for target *E. hirae*.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN