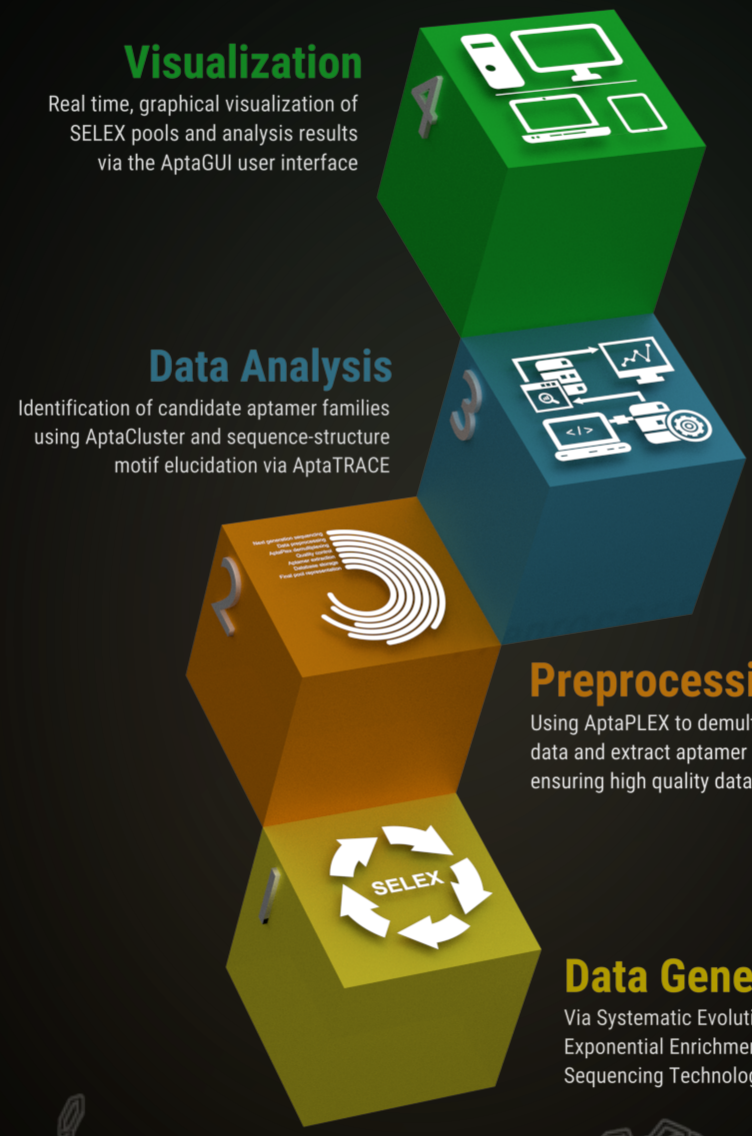
APTAMERS IN THE AGE OF BIG DATA

DEVELOPMENT AND APPLICATION OF ALGORITHMIC SOLUTIONS IN THE FIELD OF HIGH-THROUGHPUT SYSTEMATIC EVOLUTION OF LIGANDS BY EXPONENTIAL ENRICHMENT

PhD Thesis | Jan Hoinka



APTAMERS IN THE AGE OF BIG DATA

DEVELOPMENT AND APPLICATION OF ALGORITHMIC SOLUTIONS IN THE FIELD OF HIGH-THROUGHPUT SYSTEMATIC EVOLUTION OF LIGANDS BY EXPONENTIAL ENRICHMENT

Dissertation

zur Erlangung des Doktorgrades

der Technischen Fakultät

der Albert-Ludwigs-Universität Freiburg im Breisgau

03. August 2016

von

M.Sc. Bioinformatiker (Univ.)

Jan Hoinka

Dekan (Dean)
Prof. Dr. Georg Lausen
Databases and Information Systems
Department of Computer Science
University of Freiburg
Germany

Vorsitz (Chair)
Prof. Dr. Schindelhauer
Rechnernetze und Telematik
Department of Computer Science
University of Freiburg
Germany

Beisitz (Tenure)
Prof. Dr. Kuhn
Algorithms and Complexity
Department of Computer Science
University of Freiburg
Germany

Betreuer (Supervisor)
Prof. Dr. Rolf Backofen
Bioinformatics
Department of Computer Science
University of Freiburg
Germany

Gutachter (Reviewer)
Prof. Dr. Jan Baumbach
Bioinformatics
Dept. of Mathematics and Computer Science
University of Southern Denmark
Denmark

Mentor (Mentor)
Dr. Teresa M. Przytycka
Senior Investigator
Computational Biology Branch
National Center for Biotechnology Information
National Institutes of Health
United States of America

Datum der Promotion (Date of Promotion)
03. August 2016

Prologue

This dissertation would never have been possible without the support of the many wonderful people I have the honor of calling my family, friends, and colleges.

First and foremost, I am deeply beholden to my parents Hans Joachim Hoinka and Felicias Geyer who supported me in every possible aspect, have never once doubted me, and gave me the strength to continuously push myself further in my professional and personal development.

In the same way, I owe gratitude to my siblings Lea Hoinka and Nicolai Hoinka for their continuous encouragement, patience, and care throughout the years. My deep appreciation also goes to my uncle Eberhart Geyer and my aunt Kimberly Beers for their support and care. Thank you for always being there for me!

Expressing my affection, appreciation, and devotion for Natalia Acevedo Luna in words, will never do justice to how deep my feelings for her truly are. Her love, her unshattered believe in me, her honesty, her overall philosophy on life, and her persistent moral support and care have made me a better person and give me the strength to persist, both professionally and on a personal level, every single day. This thesis is dedicated to her.

I am truly indebted to my mentor Dr. Teresa Przytycka for her consistent support and guidance throughout my doctoral and pre-doctoral studies in her lab. Her outstanding advising skills, allowing me to pursue with passion the research topics I was most interest in while guiding me through the scientific discovery process and never once limiting my own modus operandi, genuinely set her apart as a senior investigator. Furthermore, I will always be appreciative of her non-trivial ability of defining a precise scientific goal while at the same time setting well defined research boundaries in which I could freely explore my own approaches and draw my own conclusion hence growing as a researcher and person. Teresa Przytycka is the type of mentor every doctoral fellow could only wish for having during his/her academic development.

I would like to express my sincere gratitude to Prof. Dr. Rolf Backofen for providing me the opportunity of performing my PhD studies under his supervision. His contribution to this thesis through countless and fruitful discussions not only substantially enriched my scientific expertise but also taught me the value of objective argumentation to improve my overall scientific thinking on a completely new level.

This dissertation would not have been possible without the countless collaborators I had the pleasure to work with over the past five years, many of which have since become much more than just colleges. These include, but are not limited to Eli Gilboa, Paloma Giangrande, Marit Nilsen-Hamilton, Vittorio DeFrancisis, Rebecca Whelan, John Burnett, and Zuben Sauna. My special thanks go to Alexey Berezhnoy and Olivier Martinez for their friendship and patience while providing me with the required skills to understanding the biological aspects of this

work. Without you, the quality and extent of this dissertation would not have been the same.

In addition, the support, professionalism and collaboration of my colleges and friends that are the group members of Dr. Przytyckas lab was, and remains, truly extraordinary in every aspect. Thank you Yoo-Ah Kim, Dong-Yeon Cho, Damian Wojtowicz, Yijie Wang, and especially The Phuong Dao for your amazing character, you encouragement, and teamwork.

Finally, I would like to thank the multitude of people I have gotten to know during my time at NIH and I am honored to call my friends. You have made this dissertation possible in ways that are impossible to convey in writing. Above all, my friendship with, and appreciation for, Victoria Porter genuinely deserve special mentioning. Meeting her has been a true privilege and joy.

Abstract

Aptamers are short, 15-150 nucleotide long RNA/DNA molecules capable of binding, with high affinity and specificity, a specific target molecule via sequence and structure features that are complementary to the biochemical characteristics of the target's surface. The spectrum of aptamer targets spans from small organic molecules, over transcription factors and other proteins or protein complexes, to the surfaces of viruses and entire cells. This broad range of targets makes aptamers suitable candidates for a variety of applications including molecular biosensors, drug delivery systems, and antibody replacement.

Aptamers are typically identified via the High-Throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX) protocol. HT-SELEX leverages the well established paradigm of *in vitro* selection by repetitively enriching a pool of initially random RNA/ssDNA sequences (species) with those that strongly bind a target of interest. Specifically, based on the assumption that a large enough initial pool of randomized (oligo)nucleotides contains some species with favorable sequence and structure allowing for binding to the target, these binders are then selected for through a series of selection cycles. Each such selection cycle involves (a) incubating the pool with the target, (b) partitioning target-bound species from non-binders and (c) removing the latter from the pool, followed by (d) elution of the bound fraction from the target, and (e) amplifying the remaining sequences, one portion of which forms the input for the subsequent round and the remainder is sequenced.

However, optimal utilization of the HT-SELEX process has lagged behind the wide range biomedical applications due to the lack of dedicated computational approaches. The key challenges in HT-SELEX data analysis include the identification of target-affine aptamers and aptamer families which are selected for, as well as the elucidation of common sequence-structure binding motifs. In addition, next-generation sequencing of the aptamer pools enables studies of important properties of the SELEX protocol itself, such as the mutational landscape of aptamer sequences due to error prone amplification (PCR) which also require to be informed by computational tools. Finally, user friendly, aptamer oriented software for demultiplexing and quality control of the raw sequencing data is also of great relevance.

To close this gap we, have developed several novel computational methods designed to tackle these challenges and to elucidate previously unappreciated properties of the SELEX protocol. For standardizing the preprocessing of raw sequencing data, we introduce AptaPLEX, a standalone and platform independent demultiplexer and quality control tool specifically designed for HT-SELEX data. Next, in order to study the effect of the selection pressures exerted during SELEX as well as to test our algorithms, we developed AptaSIM, aimed at realistically recreating the general purpose SELEX protocol *in silico*. AptaSIM simulates error-prone PCR and many aspects of *in vitro* experiments such as sampling effects, the presence or absence of pool contaminants, and aptamer affinity. To aid the identification of aptamer candidates, we introduce AptaCLUSTER, a novel technique that scales well with next generation sequencing data to efficiently cluster aptamer families in each selection round and to trace their behavior throughout consecutive cycles. Building on the results

of AptaCLUSTER we then provide the first in-depth analysis of the mutational landscape of HT-SELEX experiments and propose a theoretic model capable of discriminating favorable mutants from those which decrease the binding affinity to the target (AptaMUT). Finally, with our new AptaTRACE algorithm we tackle the challenging task of sequence-structure motif identification in HT-SELEX data. AptaTRACE is built on the idea of tracing the dynamics of the SELEX process itself to uncover motif-induced selection trends. It is robust enough to be applicable to a broad spectrum of RNA/ssDNA HT-SELEX experiments independent of the target's properties, and capable of elucidating an arbitrary number of binding sites along with their corresponding structural preferences.

The results of our approaches can be visualized via AptaGUI which provides, to the best of our knowledge, the first graphical, multi-user, and platform independent user interface for navigating high throughput sequencing data from HT-SELEX experiments.

Zusammenfassung

Aptamere sind kurze, aus 15-150 Nukleotiden bestehende RNS- bzw. DNS-Moleküle, die mit hoher Affinität und Selektivität an spezifische Zielmoleküle binden. Hierbei wird sich der Komplementarität der Sequenz- und Struktureigenschaften dieser Aptamere zu den biochemischen Oberflächeneigenschaften des Ziels bedient. Das Spektrum der Zielmoleküle reicht von kleinen organischen Molekülen über Transkriptionsfatoren, Proteinen und Proteinkomplexen bis hin zu Virusoberflächen und ganzen Zellen. Aufgrund dieser vielfältigen Einsatzmöglichkeit stellen Aptamere vielversprechende Kandidaten für eine Vielzahl von Anwendungen wie Biosensorik, Drug Delivery-Systeme sowie Antikörperersatz dar.

Zur Identifizierung von Aptameren wird üblicherweise auf das High-Throughput Systematic Evolution of Ligands by Exponential Enrichment Protokoll, kurz HT-SELEX Protokoll, zurückgegriffen. HT-SELEX nutzt das etablierte Paradigma der *in vitro* Selektion bei der sukzessiven Anreicherung eines zu anfangs aus zufälligen RNA/ssDNA-Sequenzen bestehenden Pools mit jenen Sequenzen, die ein gewünschtes Zielmolekül entsprechend stark binden. Vorausgesetzt, dass ein genügend großer Pool aus willkürlich gewählten (Oligo)Nukleotiden auch solche Spezies mit zur Bindung des Zielmoleküls vorteilhafter Sequenz und Struktur enthält, werden diese Infolge einer Serie von Selektionszyklen schließlich herausgefiltert. Jeder dieser Zyklen besteht aus (a) der Inkubation des Pools mit dem Zielmolekül, (b) der Klassifizierung in bindende bzw. nicht bindende Apatamere und (c) dem Verwerfen der letzteren, gefolgt von (d) dem Eluieren der bindenden Spezies vom Zielmolekül und (e) der Amplifizierung der übrig gebliebenen Sequenzen, welche den Pool für den nachfolgenden Durchlauf des Selektionszyklus darstellen.

Allerdings hinkt die optimale Verwendung von HT-SELEX Prozessen der Vielfalt an biomedizinischen Anwendungen hinterher, was auf das Fehlen geeigneter computerbasierter Herangehensweisen zurückzuführen ist. Die Schlüsselherausforderungen der HT-SELEX Datenanalyse beinhalten die Identifizierung der zu selektierenden, Target-affinen Aptamere und Aptamerklassen sowie die Aufklärung der üblichen Sequenz/Struktur-Bindungsmotive. Weiterentwicklungen im Bereich der Sequenzierung des Aptamerpools liefern aber auch Informationen über das HT-SELEX selbst: So können Mutationslandschaften der Apatamersequenzen, welche aufgrund von fehleranfälligen Amplifizierungen (PCR) entstehen, adaptiv in den Prozess mit eingebunden werden. Schlussendlich ist eine benutzerfreundliche, aptamerorientierte Software für das Demultiplexing und das Qualitätsmanagement der rohen Sequenzierungsdaten von Relevanz.

Um diese Lücke zu schließen, wurden verschiedene computerbasierte Methoden entwickelt, die sich zum Einen der zuvor beschriebenen Herausforderungen annehmen und die zum Anderen weitere, bis dato unerkannte Möglichkeiten der SELEX Protokolle ermitteln sollen. Zur Standardisierung der Prozessierung von Sequenzierungsrohdaten wird die Software AptaPLEX eingeführt. Hierbei handelt es sich um einen autonomen und plattformunabhängigen Demultiplexer und Qualitätsmanager, der speziell für HT-SELEX Daten designet wurde. Als nächstes, um den Effekt des ausgeübten Selektionsdrucks während eines SELEX Prozesses

zu studieren sowie um die hier vorgestellten Algorithmen zu testen, wurde AptaSIM entwickelt, welches *in silico* auf die realistische Simulation des grundlegenden Prinzips des SELEX Protokolls abzielt. AptaSIM simuliert fehleranfällige PCR und viele weitere Aspekte der in vitro Experimente wie z.B. der Samplingeffekt, die An- bzw. Abwesenheit von Poolverunreinigungen sowie Aptameraffinitäten. Zur erleichterten Identifizierung von Aptamerkandiaten wird der Algorithmus AptaCLUSTER eingeführt. Diese neuartgie Technik skaliert gut mit Next Generation Sequenzierungsdaten zur effektiven Kategorisierung von Aptamerfamilien innerhalb der jeweiligen Selektionszyklen sowie zur Nachverfolgung ihres Verhaltens während der aufeinanderfolgenden Zyklen. Auf den Ergebnissen von AptaCLUSTER aufbauend, wird die erste detaillierte Analyse von Mutationslandschaften der HT-SELEX Experimente in form von AptaMUT bereitgestellt und ein theoretisches Modell vorgeschlagen, welches zwischen vorteilhaften Mutationen und solchen mit abschwächender Bindungsaffinität zum Zielmolekül zu unterscheiden vermag. Schließlich nehmen wird sich mit unserem neuen AptaTRACE-Algorithmus der herausfordernden Aufgabe der Sequenz/Struktur-Motividentifikation anhand von HT-SELEX Daten angenommen. AptaTRACEist auf die Idee hin entwickelt worden die Dynamik des SELEX Prozesses selbst nachzuvollziehen, um somit die strukturinduzierten Selektionstrends zu identifizieren. Der Algorithmus ist robust genug, um unabhängig von den Eigenschaften des Targets auf ein breites Spektrum von RNS/ssDNS HT-SELEX Experimenten angewendet zu werden. Desweiteren ist er in der Lage eine beliebige Anzahl an Bindungsstellen und deren zugehörige strukturelle Präferenzen zu erkennen und aufzuklären.

Die Resultate unserer Herangehensweise können via AptaMUT visualisiert werden, welches nach unserer Kenntnis die erste graphische, multi-user und plattformunabhängige Benutzeroberfläche zur Handhabung von High Throughput Sequenzierungsdaten aus HT-SELEX Experimenten ist.

List of Publications

This thesis is based on the following publications:

[1] **Jan Hoinka**, Phuong Dao, Yijie Wang, Mayumi Takahashi, Jiehua Zhou, Fabrizio Costa, John Rossi, John Burnett, Rolf Backofen, and Teresa M Przytycka. AptaTRACE Elucidates Aptamer Sequence-Structure Motifs in HT-SELEX Experiments. *Cell Systems*, (Manuscript Accepted for Publication), 2016

[2] **Jan Hoinka** and Teresa Przytycka. AptaPLEX – A dedicated, multithreaded demultiplexer for HT-SELEX data. *Methods*, 2016

[3] **Jan Hoinka**, Phuong Dao, and Teresa M Przytycka. AptaGUI—A Graphical User Interface for the Efficient Analysis of HT-SELEX Data. *Molecular Therapy—Nucleic Acids*, 4(10):e257, 2015

[4] **Jan Hoinka**, Alexey Berezhnoy, Phuong Dao, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Research*, 43(10):5699–5707, 2015

[5] **Jan Hoinka**, Phuong Dao, Yijie Wang, Mayumi Takahashi, Jiehua Zhou, Fabrizio Costa, John Rossi, John Burnett, Rolf Backofen, and Teresa M Przytycka. AptaTRACE: Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments. *Research in Computational Molecular Biology : Annual International Conference, RECOMB: Proceedings*, 2016

[6] **Jan Hoinka**, Phuong Dao, and Teresa M Przytycka. AptaGUI—A Graphical User Interface for the Efficient Analysis of HT-SELEX Data. *Molecular Therapy—Nucleic Acids*, 4(10):e257, 2015

[7] Agata Levay, Randall Brenneman, **Jan Hoinka**, David Sant, Marco Cardone, Giorgio Trinchieri, Teresa M Przytycka, and Alexey Berezhnoy. Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment. *Nucleic Acids Research*, 43(12), 2015

[8] **Jan Hoinka**, Alexey Berezhnoy, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. *Research in Computational Molecular Biology : Annual International Conference, RECOMB: Proceedings*, 8394:115–128, 2014

[9] **Jan Hoinka**, Elena Zotenko, Adam Friedman, Zuben E. Sauna, and Teresa M. Przytycka. Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, 28(12), 2012

List of Oral Presentations

Individual topics of this thesis have been presented at the following conferences:

April 2016	10th RNA Consortium, City of Hope, Duarte, California, USA Title: AptaTRACE - Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments
April 2016	20th Annual International Conference on Research in Computational Molecular Biology (RECOMB),Santa Monica, California, USA Title: AptaTRACE - Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments
June 2014	GTC Non-Coding RNAs and RNAi Reserach & Therapeutics Conference, San Diego, California, USA Title: AptaSuite - An In-Silico Lab to Analyze HT-SELEX Aptamer Pools and Lessons from its Application
April 2014	18th Annual International Conference on Research in Computational Molecular Biology (RECOMB),Pittsburgh, Pennsylvania, USA Title: AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application
February 2014	9th RNA Consortium, Santa Barbara Univerity, Stanta Barbara, California, USA Title: AptaGUI - Visualizing the Identication and Analysis of Aptamer Families using HT-SELEX Data
October 2013	Aptamers in Medicine and Perspectives, Naples, Italy Title: HTSAptamotif - Identification and Analysis of Aptamer Families using HT-SELEX Data.
January 2013	8th RNA Consortium, City of Hope, Duarte, California, USA Title: HTSAptaMotif - An in-silico Approach to Sequence-Structure Motif Identification for HTS SELEX Derived Aptamers.
June 2012	International Symposium of Molecular Biology (ISMB) 2012, Long Beach, California, USA Title: Identification of Sequence-Structure RNA Binding Motifs for SELEX-derived Aptamers.

Writing style

In a modern, interdisciplinary research environment, team work and collaborative projects form a fundamental basis for the advancement of scientific achievements. This is especially true for the field of computational biology and bioinformatics which requires a close cooperation with wet-lab experimentalists for biological validation and expert knowledge. In addition, the complexity of today's computational approaches often requires a solution best achieved through joint team effort. In recognition of these research aspects, I decided to write this dissertation in the plural form. Thus, "we" is used throughout this thesis even for those parts that were conducted solely by myself.

In view of this, I contributed fundamentally to all the publications forming the foundation of this thesis. My responsibilities included, but were not limited to, the development of the theoretical aspects regarding the here presented methods, their subsequent refinement and implementation, as well as generating the results and designing the experimental evaluation pipelines with our collaborators whenever possible. Furthermore, I played a pivotal role in writing the manuscripts which, in conjunction with the above contributions, is reflected in the first authorship of the publications this dissertation is built on.

The Graduate Partnership Program at the National Institutes of Health

This dissertation is the result of a successful partnership between the Albert-Ludwigs-Universität Freiburg, Germany, under the supervision of Prof. Dr. Rolf Backofen, and Dr. Teresa Przytycka from the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH), USA. The collaboration was formalized through an Individual Partnership of the NIH Graduate Partnerships Program (GPP). The NIH Office of Intramural Training & Education (OITE) hosts the GPP, which is designed to bring PhD graduate students to the NIH Intramural Research Program for dissertation research. Participants enjoy the academic environment of a university, the extensive research resources of the NIH, and the breadth and depth of the research programs of both the host university and the NIH Intramural Research Program (IRP). The goal is to create a different kind of graduate experience, one that focuses on training the next generation of scientific leaders by emphasizing communication and collaboration skills, integration of information, and interdisciplinary investigation.

As such, this work was supported by the NIH Intramural Research Program. Specifically, for the duration of my PhD studies, I was situated at the Computational Biology Branch (CBB) at NCBI/NIH in Bethesda, Maryland, USA and mentored on site by Dr. Przytycka, Senior Investigator and head of the Algorithmic Methods in Computational and Systems Biology research group.

Contents

List of Figures

List of Tables

1

Introduction

> *A totally blind process can by definition lead to anything; it can even lead to vision itself.*
>
> Jacques Monod, *Chance and Necessity*, 1972

1.1 Motivation

One of the fundamental principles enabling the existence of any organic entity known to science consists in the process of biomolecular recognition. This mechanism, by which biomolecules recognize and bind to their molecular targets with typically high affinity and specificity, drives a highly concerted interplay of various intermolecular interactions between proteins, DNA and RNA, and (un)organic ligands. In turn, a multitude of biotechnologies and pharmaceutical approaches relay on the development of ligands that influence these interactions [9, 10, 11, 12]. Successful design of such ligands therefore requires a in-depth understanding of the biomolecular recognition process as well as appropriate pipelines allowing for an efficient generation of target-specific molecules.

Even in an era of high-throughput screening technologies, designing a ligand with the desired binding properties to a specific molecular target is highly challenging, time consuming and cost intensive. Given a target of interest, such a process typically involves testing hundreds of thousands of compounds for binding, the isolation and biochemical characterization of a lead compound, and the subsequent refinement of its binding properties [13]. Besides computational optimization techniques [14], the latter is usually achieved through a process known as combinatorial chemistry in which millions of slight derivatives from the original ligand are synthesized at random and tested for improved binding [15]. Clearly, the main bottleneck in such a discovery process is the vast amount of affinity assays that need to be performed for each potential compound. Therefore, alternative solutions for the design of high-affinity binders to arbitrary biomolecules which overcome the above mentioned short-

coming have been envisioned. Such a solution requires a) an experimental procedure which enables rapid, direct and systematic selection of strong-binding ligands and b) a corresponding biomaterial which can undergo this selection, is flexible enough to span a large structural and biochemical space, and can be synthesized, handled, and stored in a time sensitive and inexpensive manner.

Interestingly, ribonucleic acids (RNAs) provide an ideal candidate for the latter requirement. Besides their traditional role as a passive member of the synthesis machinery converting information in form of DNA into proteins, a large number of novel, non-coding RNAs involved in a vast array of processes ranging from catalytic reactions (snoRNAs [16]), over gene and metabolic regulatory roles (snRNAs [17, 18], miRNAs and siRNAs [19, 20, 21, 22]), to biosensors in form of riboswitches [23] have been, and are still being discovered. While the above list of RNA types only represents a small subset of the entire known RNA universe, it does emphasizes the structural dynamics and range of functions ribonucleic acids play in biological systems. Consequently, their ability to target with specificity a vast array of molecules and to perform precise operations *in vitro* and *in vivo* has prompted researchers to utilize and repurpose RNAs for their own objectives.

RNAs binding a specific molecular target are typically identified through the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) protocol [24]. Although the specifics vary depending on the target, SELEX leverages the well established paradigm of *in vitro* selection by repetitively enriching a pool of initially random sequences (species) with those that strongly bind a target of interest. Specifically, based on the assumption that a large enough initial pool of randomized (oligo)nucleotides contains some species with favorable sequence and structure allowing for binding to the target, these binders are then selected for through a series of selection cycles. Each such cycle involves (a) incubating the pool with the target, (b) partitioning target-bound species from non-binders and (c) removing the latter from the pool, followed by (d) elution of the bound fraction from the target, and (e) amplifying the remaining sequences via polymerase chain reaction (PCR) to form the input for the subsequent round (see Figure 2.1). After a target-specific number of selection cycles, the final pool is then used to extract dominating, putatively high-affinity species, via traditional cloning experiments, computational analysis, and binding affinity assays. Notably, the iterative displacement of non-binders from the library in combination with simultaneous enrichment of target affine species in the pool allows for a drastic reduction in the number of required binding assays as compared to the conventional approach described above. Here, only a select number of ligands from the final selection round are characterized for their binding properties and are often further post-processed *in vitro* to meet additional requirements such as improved structural stability or reducing the size of the sequences to the relevant binding region. Those sequences conforming to the specifications of their intended application are called aptamers, derived from the Latin word *aptus* 'to fit' and the Greek word *merus* for 'piece'.

Formally, aptamers are defined as short RNA/DNA molecules capable of binding, with high affinity and specificity, a distinct target molecule via sequence and structure features that are complementary to the biochemical characteristics of the target's surface. In particular, aptamers have recently regained substantial momentum in the biotechnology industry, medical sciences, and academia alike. While only 117 aptamer related publications were added

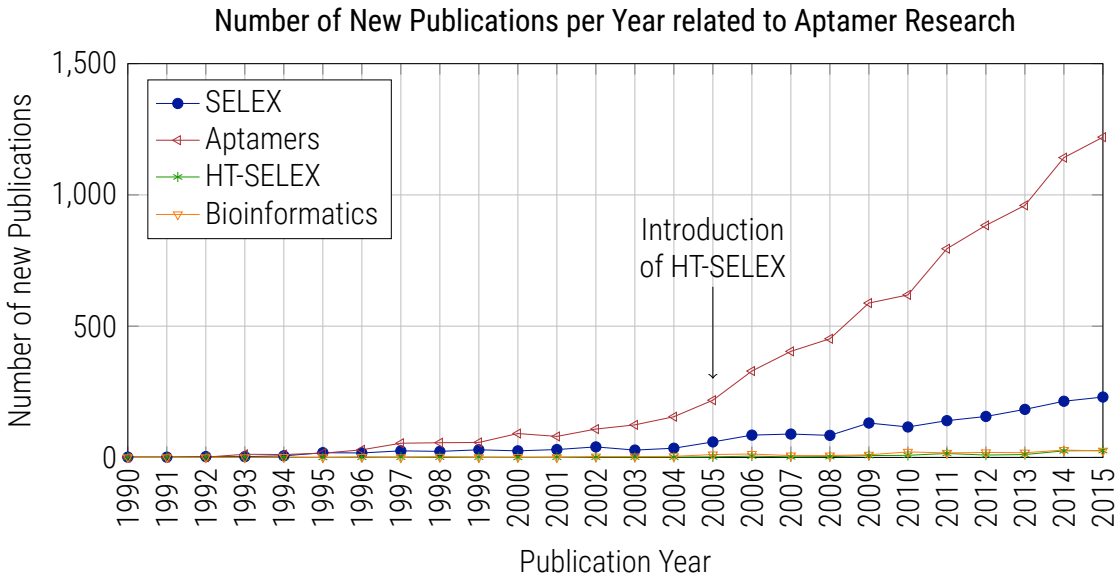


Figure 1.1: Number of aptamer related publications according to PubMed based on the occurrence of keyword terms (see legend) in the abstract of all publications in the database. Shown are the non-cumulative values of new publications per year between the introduction of the SELEX protocol in 1990 and 2015 as indexed by PubMed. The back vertical arrow indicates the introduction of the HT-SELEX protocol and the subsequent substantially larger publication rate in the field of aptamer research.

to PubMed in the year 2000, this number has since roughly doubled every 5 years, with 290 records added in 2005 alone, 764 additional inclusions in 2010, and as many as 1501 new manuscripts indexed in 2015 (see Figure 1.1). This astonishing trend is in part attributable to the considerable diversity of possible targets which span from small organic molecules [25], over transcription factors [26] and other proteins or protein complexes [27], to the surfaces of viruses [28] and entire cells [29, 30, 31]. The broad range of targets makes aptamers suitable candidates for a variety of applications ranging from molecular biosensors [32], to drug delivery systems [33], and antibody replacement [34] to just name a few.

Another reason for the resurgence of interest in aptamer research relates to the utilization of affordable next-generation sequencing technologies along with traditional SELEX. This novel protocol, called HT-SELEX, combines Systematic Evolution of Ligands by Exponential Enrichment with high-throughput sequencing. In HT-SELEX, after certain (or all) rounds of selection (including the initial library), aptamer pools are split into two or more samples. The first sample serves as the starting point for the next cycle whereas the latter are sequenced or stored for future reference. The resulting sequencing data, consisting of 2-50 million sequences per round, is then analyzed *in silico* in order to identify candidates that are selected for [4, 35].

The massive amount of sequencing data produced by HT-SELEX opens the opportunity for the study of many aspects of the protocol that were either not accessible in traditional SELEX or that could be realized more accurately given hundreds of millions of data points. Key challenges in HT-SELEX data analysis include the identification of target-affine aptamers and aptamer families (species related to each other by sequence) which exhibit exponential enrichment, as well as the elucidation of common sequence-structure binding motifs. In ad-

dition, next-generation sequencing of the aptamer pools enables studies of important properties of the SELEX protocol itself, such as the mutational landscape of aptamer sequences due to error prone amplification (PCR) which also require to be informed by computational tools. Finally, user friendly, aptamer-oriented software for demultiplexing and quality control of the raw sequencing data is also of great relevance.

However, development of universal methods for the analysis of HT-SELEX data is challenged by the vast diversity of selection conditions such as temperature, salt concentration and the number of targets in the solution to just name a few. Further, each of the stages (a-e) comprising one selection cycle can be accomplished by a variety of technologies. Even more importantly, the complexity of the target is also of great relevance. As an illustration, it has been shown that *in vitro* selection against transcription factors and other molecules requires only a small number of selection rounds in order to produce high quality aptamers [36, 26], likely due to their evolutionary optimization to efficiently recognize specific DNA/RNA targets. On the other side of the spectrum, in the case of CELL-SELEX, a variation of SELEX in which the pool is incubated with entire cells expressing a target of interest on their surface, the number of required selection cycles and the amount of non-specific binders that emerge during selection is significantly larger [29]. Indeed, such a target can in general accommodate a multitude of binding sites, each exposing different binding preferences and leading to a parallel selection towards unrelated binding motifs [30].

Taken together, the number of existing technologies and variations HT-SELEX can be performed with is reflected in the quality and properties of the resulting high-throughput sequencing data and represents a clear challenge for the development of robust and bias-resistant computational tools. Particularly the optimal utilization of the HT-SELEX process has lagged behind the wide range of its biomedical applications due to the lack of appropriate *in silico* methods (Figure 1.1 "Bioinformatics"). In order to close this gap, here, we propose a broad set of flexible computational approaches tailored towards the efficient analysis of aptamer selection trends while leveraging the entirety of data produced by modern SELEX experiments.

1.2 General Objectives

Given the potential of aptamers and their extensive application possibilities in the biomedical, pharmaceutical, and research community alike, we have developed a comprehensive *in silico* suite known as AptaTOOLS with the purpose to streamline the identification of aptamers with specific properties from HT-SELEX data. In addition, our research has revealed previously under-appreciated properties of the SELEX protocol itself which have the potential to significantly optimize the cost of selection procedure in terms of required labor and financial burden. Specifically, this thesis advances the field of aptamer research with the following computational contributions:

1. We tackle the challenging task of sequence-structure motif identification in HT-SELEX

data in form of our AptaTRACE algorithm. Our method is built on tracing the dynamics of the SELEX process itself to uncover motif-induced selection trends. AptaTRACE is robust enough to be applicable to a broad spectrum of RNA/ssDNA HT-SELEX experiments independent of the target's properties, and capable of elucidating an arbitrary number of binding sites along with their corresponding structural preferences. Furthermore, we show how our method can lead to time-and-cost optimized SELEX protocols.

2. We introduce AptaCLUSTER, a novel technique to efficiently determine aptamer families in each selection round and to trace their behavior throughout consecutive selection cycles. In contrast to traditional clustering techniques, AptaCLUSTER scales well with next generation sequencing data. In addition, we show that AptaCLUSTER can reveal properties of the selection process which have previously not been appreciated and which can be utilized for and optimized aptamer discovery process.

3. We provide the first in-depth computational analysis of the mutational landscape of HT-SELEX experiments and propose a mathematical model, AptaMUT, capable of discriminating favorable mutants from those which have a detrimental effect on the binding affinity to the target.

4. In order to study more general properties of the SELEX protocol, we developed AptaSIM. AptaSIM is aimed at realistically recreating selections *in silico* and can simulate many aspects of *in vitro* experiments such as error-prone PCR, sampling effects, the presence or absence of pool contaminants, and aptamer affinity. Using AptaSIM not only allows to quantify several of those parameters in real data, but also serves as a benchmarking and validation framework for the remaining approaches presented in this thesis.

5. We present our graphical user interface AptaGUI, which provides, to the best of our knowledge, the first graphical, multi-user, and platform independent navigation tool designed for high throughput sequencing data from HT-SELEX experiments. AptaGUI includes, among others, support for AptaCLUSTER, AptaMUT, and AptaSIM and allows for easy integration of additional features due to its modular design.

6. For standardizing the preprocessing of the raw sequencing data, we introduce AptaPLEX, a standalone and platform independent demultiplexer and quality control algorithm tailored towards HT-SELEX data.

Despite the computational nature of this thesis, our research also contains a strong biological focus. The general objective was to develop and apply *in silico* approaches to support and complement biological research on aptamer development, design, and refinement. This is reflected in the tight network of collaborations with wet-lab experimental groups which performed the SELEX experiments and provided the data forming the base of this work. Indeed, out of the nine publications which form the foundation of this dissertation, seven were produced in close collaboration with multiple experimental research groups.

1.3 Thesis Outline

The content of this thesis covers a variety of computational, as well as biological questions involving aptamer research and the SELEX protocol. As such, we have put special emphasis in structuring this dissertation into self-contained chapters which can be read and evaluated individually without requiring the knowledge from the remaining parts. In order to obtain a general overview, it is therefore sufficient to read the introductions and conclusions of each section. The general outline can be summarized as follows:

- Chapter 1 introduces the general topic of the dissertation , provides a brief biological background of aptamers and their selection process, and defines the scope of this work with respect to the computational achievements described therein.

- Chapter 2 provides the reader with a more detailed description of the experimental methodologies regarding the SELEX protocol and each of its selection steps. Special attention is paid to elucidating how the most commonly used experimental methods to achieve each of these steps affect the landscape of the resulting sequencing data. Furthermore, we motivate why taking these effects into account is crucial for the development of accurate and robust computational tools.

- Chapter 3 showcases our approach to standardize the processing of raw sequencing data produced by modern HT-SELEX experiments for downstream analysis. Specifically, we present our demultiplexing algorithm, AptaPLEX which is capable of efficiently extracting aptamers from raw sequencing data and assigning these to their corresponding selection cycles while adhering to strict quality control routines.

- Chapter 4 details our computational approaches for data-driven aptamer analysis, divided into four sections. Each section is concerned with a specific aspect and/or biological question and our corresponding computational solution. These correspond to AptaSIM, AptaCLUSTER, AptaMUT, and AptaTRACE as described in the General Objectives 2-5, outlined above in Section 1.2.

- Chapter 5 is concerned with AptaGUI, our approach on visualizing the results of our approaches in an interactive, modular, and extensible manner. We motivate the need for such a system and highlight how it can be implemented into modern, security-focused IT infrastructures in a platform independent way.

- Chapter 6 concludes the entire thesis, explains general limitations of our approaches and offers suggestions on possible improvements and future work.

For readers with a special interest in the biological aspect of the work, additional details regarding the selection protocol are provided in the Appendix A. Similarly, Appendix B contains complementary materials and methods concerning our algorithmic approaches from Chapter 4.

2

Aptamers and the Systematic Evolution of Ligands by Exponential Enrichment

> *Things exist either because they have recently come into existence or because they have qualities that made them unlikely to be destroyed in the past.*

Richard Dawkins, *The Blind Watchmaker*, 1986

In this chapter, the general Systematic Evolution by Exponential Enrichment protocol and its individual steps are formally introduced to the reader. We begin with a general description of the SELEX procedure in sufficient detail as to appreciate the remaining chapters of this work. Furthermore, special emphasis is put on the selection conditions of the protocol and how these affect the landscape of the resulting pools from a computational perspective. In a similar fashion, an overview of the different technologies that can be utilized to accomplish each of the selection steps (a-e) and their impact on the sequencing data is provided. Finally, we introduce an excerpt of the most relevant SELEX variations that have been developed to date in order to accommodate selections against the different types of targets in an effort to elucidate the complexity of the data resulting from high-throughput sequencing these pools.

2.1 General Description of the SELEX Procedure

Systematic Evolution of Ligands by EXponential Enrichment (SELEX) is an experimental technique allowing for the identification of aptamers - short, synthetic, single-stranded (ribo)-nucleic molecules selected to bind with high specificity almost any molecular target of interest [24, 37]. The binding targets aimed at with SELEX vary from small organic molecules [38, 39], through transcription factors [40, 41, 42] and other proteins and protein complexes [27], to viruses [28, 43] and cells [31, 44]. In addition, aptamers are chemically synthesized, can be well characterized by analytical methods, have limited toxicity, and are expected to

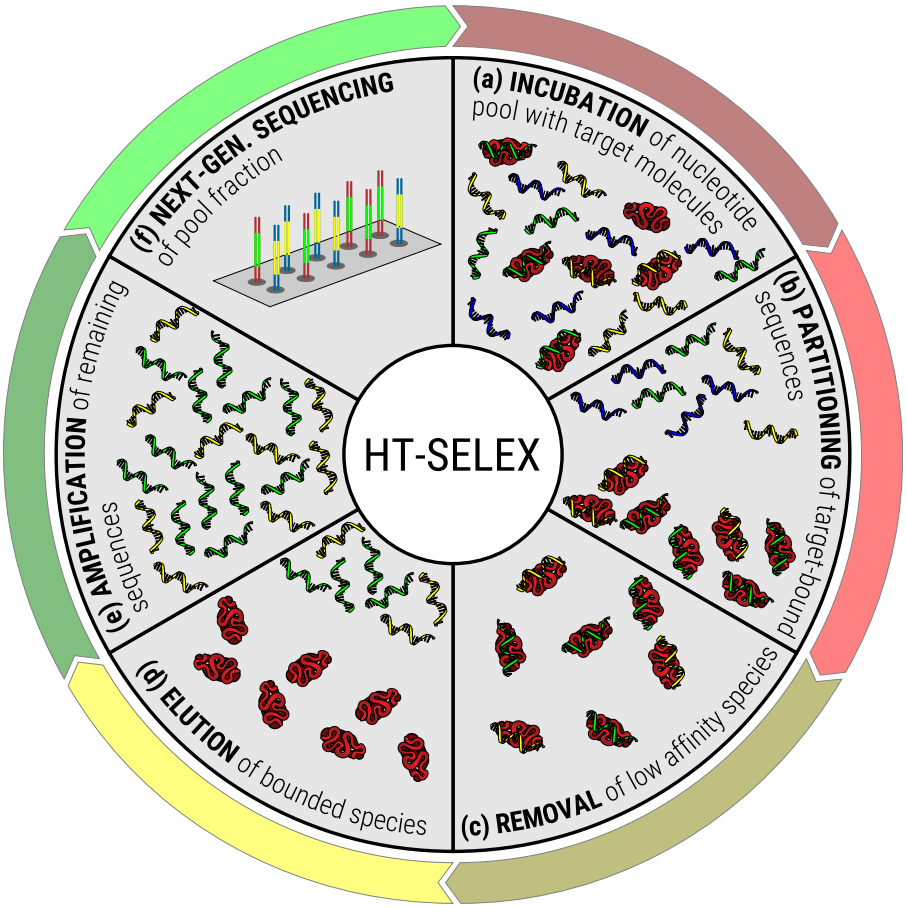


Figure 2.1: Overview of the principles of HT-SELEX. Schematic of the steps defining one selection cycle (clockwise): incubation of a sequence pool with the target, binding of target affine species, partitioning of target-bound and low-affinity species, target-bound elution, and amplification followed by high throughput sequencing.

be less or non-immunogenic as compared to their biologics counterparts such as antibodies. Until recently, the generation of aptamers took a black box approach where a traditional SELEX procedure iterates over four basic steps that together define one selection cycle: (a) incubation and binding, (b) partitioning and washing, (c) target-bound elution, and (e) amplification (Figure 2.1). The process starts with a sequence library of $10^7 - 10^{15}$ random molecules of fixed length flanked by constant primer sites to aid amplification. At the beginning of each cycle, such an RNA/ssDNA pool is incubated with a target of interest. At the end of each cycle, lower affinity binders are removed from the solution whereas bound aptamer molecules are eluted and amplified, forming the input for the consecutive round. The aptamer molecules that persist until the final cycle are then evaluated experimentally for binding affinity and optimized for specific properties, such as size or stability, depending on the intended application.

Massively parallel sequencing technologies have the potential to revolutionize the SELEX protocol by allowing sequencing of entire aptamer pools [36], leading to a novel protocol referred to as High-Throughput SELEX (HT-SELEX). In an HT-SELEX procedure some (or all) selection rounds are sequenced and computationally analyzed for potential binders. More

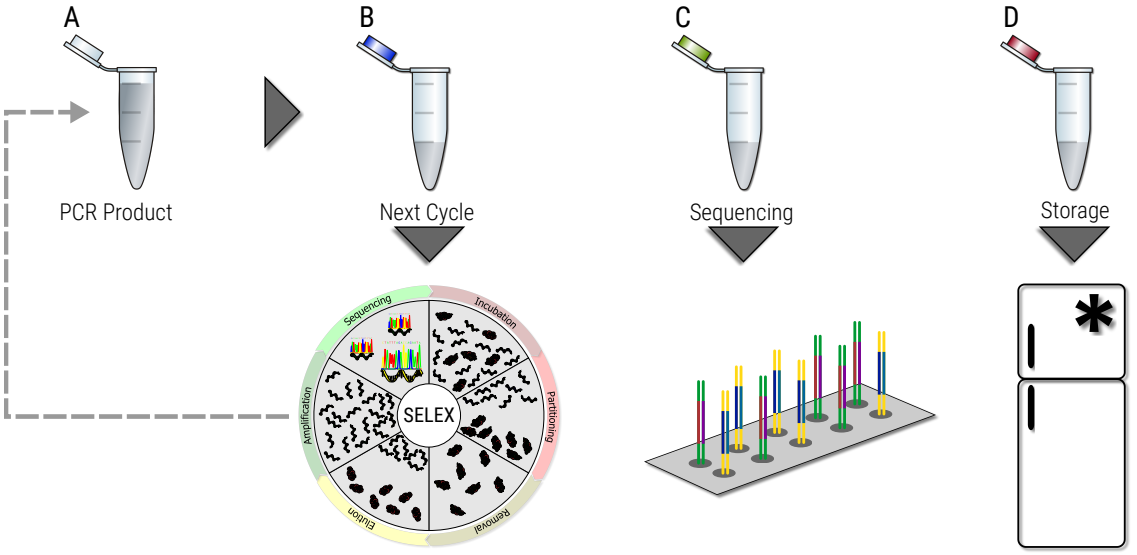


Figure 2.2: Schematic of the aliquotation process during HT-SELEX. **(A)** After each amplification step in a selection cycle, the PCR product arising from this procedure is partitioned into three distinct samples (aliquots) at ratios which are dependent on the experiment and the researchers requirements. **(B)** The first aliquot serves as input for the next selection round, whereas the remaining samples are either **(C)** sequenced for the purpose of computational analysis or **(C)** stocked in ultra-low freezers ($\leq$ 70 Degree Celsius) for reference purposes or further experiments.

precisely, after each amplification step, a sample of the aptamer pool is sequenced while another fraction serves as the input for the next cycle in a process known as aliquotation (see Figure 2.2 for details). Thus, unlike the traditional SELEX approaches in which only a handful of aptamers are sequenced and analyzed after the last cycle, HT-SELEX provides data for a global analysis of the selection properties and for the simultaneous discovery of a large number of candidates. This large amount of information has utility only in conjunction with suitable computational methods to analyze the data, sort the potential aptamers and identify candidate aptamers with properties consistent with the intended application.

The development of robust computational tools however is challenged by the large number of different technologies and experimental conditions each of the selection steps can be performed with, which can potentially have a substantial effect on the composition of the resulting sequencing data. These include, but are not limited to, nucleotide biases in the initial library, the pH value and temperature of the incubation buffer affecting the folding properties of the species (and therefore aptamer-target binding), and the time span selected for incubation and washing. Next, the ratio of aptamer sequences to target molecules is also known to impact the selection process by increasing competition between target-affine binders. Furthermore, the PCR technology utilized during the amplification step, as well as the type of target itself, all contribute towards highly diverse data sets.

It is therefore not surprising that these experiment-specific details must be acknowledged when analyzing HT-SELEX data *in silico*. More importantly, for the development of universally applicable algorithms in the field of aptamer research, an in-depth understanding of these biases, their inner mechanics, and their extent, is paramount.

2.2 The Initial Sequence Library

The experimental design of SELEX is based on the assumption that a large enough pool of initially random candidate sequences is likely to contain nucleotide strands which fold into structures complementary in shape and biochemical properties to a targets binding site (apatope). In theory, it would be desirable that these sequences span the entire sequence space for a given nucleotide length n, i.e. that at least one instance of each of the 4^n possible sequences be present in the initial pool. In practice however, these pools typically consist of $10^{14} - 10^{16}$ individual molecules and as a consequence only a small fraction of all possible sequences occur in the initial library for larger n.

Each sequence constituting the pool and consequently present throughout the selection process consists of a randomized region of fixed nucleotide size flanked by constant primer regions required for the polymerase chain reaction during the amplification stages of the protocol. The choice of the randomized region size is dependent on multiple factors such as, but not limited to, the complexity of the target, the intended purpose of the aptamer, and ultimately the researchers experience. As a point in case, SELEX is routinely leveraged for the discovery of transcription factor binding motifs [45, 36]. These motifs, ranging between 4-10 nucleotides (nt) in size, tend to be rather small allowing for successful selections with randomized region sizes around 10-20 nt. On the other side of the spectrum, libraries of up to 60nt in size have been employed against more complex targets such as human alpha-Thrombin [46], a powerful player in the coagulation process of blood. Notably, choosing a particular randomized region size has two major consequences for the subsequent selection. First, it determines the density at which the sequence space can potentially be sampled. For a pool size of 10^{15} molecules for instance, a randomized region of $n = 25$ is still sufficiently small to cover the entire sequence space, however, choosing a larger value such as $n = 30$ results in sampling only 9.3×10^{-10} percent of the 4^{30} possible combinations of nucleotides. Second, the randomized region size provides an upper bound on the conformational space any sequence is able to attain via secondary and tertiary folding and can therefore be seen as a limiting factor for the structural diversity of the molecules.

The libraries utilized in SELEX experiments are typically synthesized chemically in an iterative procedure. Starting with a support platform (solid support/phase) onto which single but well-defined bases are attached (these will later on form the beginning of the primer region), the individual sequences are then grown by incubating the solid support with a solution containing a reactive mixture of specially modified adenine, cytosine, guanine, and thymine. Consequently, the nucleotides in the solution randomly couple with one of the already present sequences on the solid phase whereas the modification prevents them from reacting with additional components once attached to a particular strand. Next, the remaining bases in the solution are washed off, and the blocking modification is removed, leaving each sequence with one additional base. Finally, by repetitive incubation and washing, these sequences are grown to the desired length.

Notably, the synthesis of oligonucleotides at massive scales is a well developed and fully automated technology. This automation however can also introduce potential biases into

the library which must be taken into account when analyzing SELEX experiments *in silico*. One such bias relates to the composition of the nucleotide mixture which typically consists of an equal amount of each base type [47]. These nucleotides however are known to differ slightly in their reactivity to bind other bases and especially thymine is recognized to preferentially bind to itself. This can lead to sequences containing large stretches of thymine and as a consequence, to a non-uniform nucleotide distribution of the initial pool which tends to propagate through the subsequent selection process (see Figure 2.3 A-D). From the perspective of sampling the sequence space, this bias translates into an uneven representation of the possible aptamers in the pool in which areas corresponding to thymine-rich species are represented more densely. This potentially reduces the chances of synthesizing a target-affine species in the initial library. In order to mitigate this issue, next-generation synthesizers allow for manual adjustment of the base composition during synthesis, e.g. by reducing the amount of thymine according to its increased level of reactivity. While these hand-mixed approaches do better approximate a uniform distribution of the nucleotide composition in the resulting library, its success remains highly dependent on the scientists expertise of operating this technology and traces of this bias tend to remain in the initial library (see Figure 2.3 E).

2.3 Incubation of the Aptamer Library with the Target

Provided with either the initial library when starting a new selection, or any intermediate pool from a particular selection cycle, SELEX aims at reducing the species in the sequence pool to those which exhibit strong affinity to the target of interest. This is typically achieved by first creating a mixture of the denaturated aptamer sequences with Phosphate-buffered saline (PBS), a water-based salt solution containing sodium hydrogen phosphate, sodium chloride and, in some formulations, potassium chloride and potassium dihydrogen phosphate. This results in the renaturation of the aptamer sequences, i.e. their folding into secondary and tertiary structures depending on the nucleotide sequence and the intended pH value of the solution. Next, the aptamers suspended in solution are incubated with the target under the assumption that the diversity of aptamer sequences in the initial pool is large enough to produce at least one species that is complementary in structure and biochemical properties to the surface of the target. Similarly, pools of subsequent selection cycles are assumed to retain these target-affine binders which at this point either interact with distinct apatopes or compete for the same binding moiety on the target surface.

Five main factors of the incubation process are known to have a measurable effect on the overall selection and therefore the resulting sequencing data, the first two of which affect the structural properties of the aptamers whereas the latter regulate the selection pressure exerted on the system. The folding properties of aptamers depend, among multiple other factors, on temperature and the salt (NaCl) concentration of the solution [48]. A change in either of these parameters could therefore alter the structural landscape of a given pool and as consequence shift the affinities of aptamer families in a favorable or detrimental

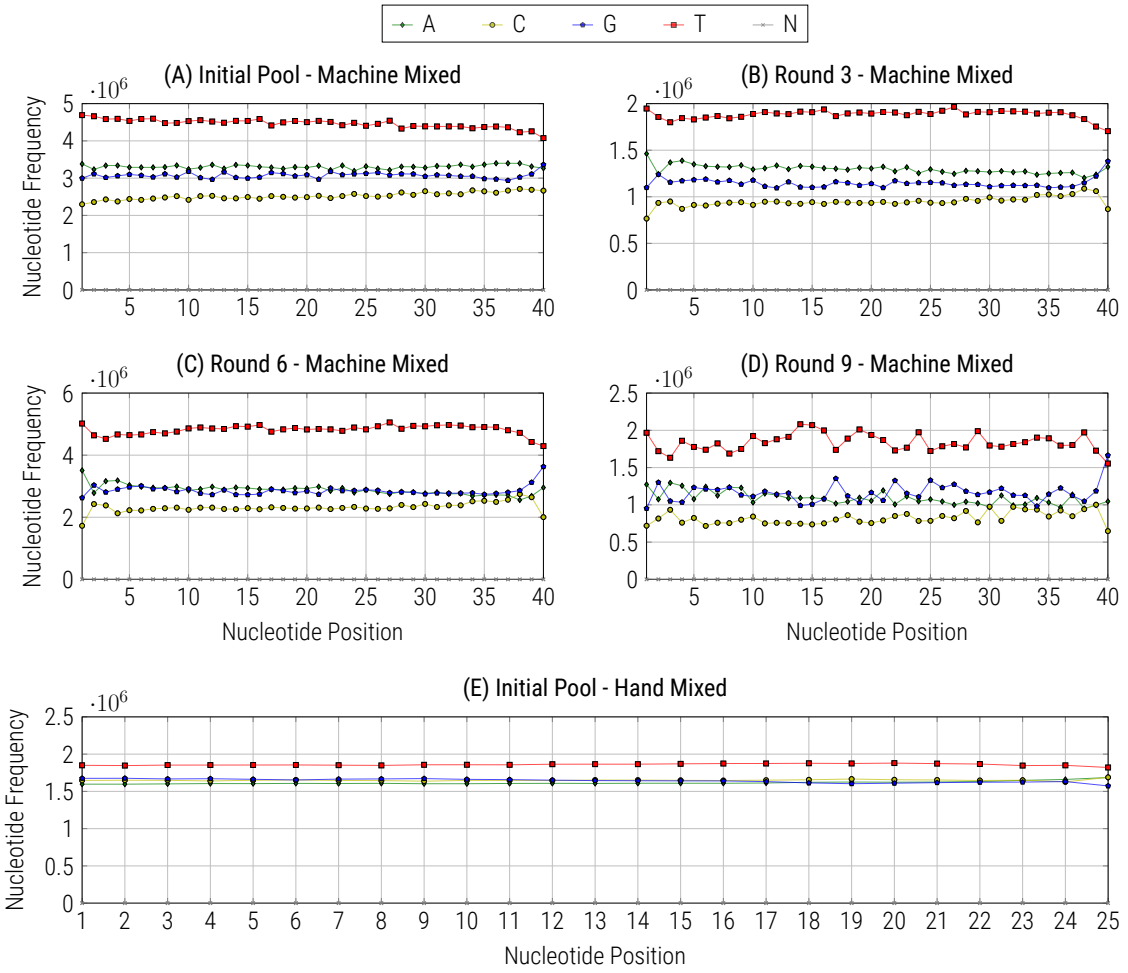


Figure 2.3: Nucleotide biases in the initial library. **(A)** Nucleotide distribution of the reads in the initial pool based on a 40nt library synthesized using a machine mixed procedure. **(B-D)** SELEX against the membrane associated protein Mucin 16. Shown are the nucleotide frequencies of rounds 3, 6, and 9 respectively. In each pool, the Thymine bias remains present. **(E)** Initial pool of size 25nt synthesized using the hand mixed procedure. While the Thymine bias is still present, the resulting nucleotide distribution is significantly closer to being uniform. Data Source: Rebecca Whelan Lab, Collaborator

manner. In addition, increasing the salt concentration is often used as a means of selecting aptamers with potentially higher binding affinity towards the target. By choosing an appropriate selection round in which most of the non-binders have been competed out of the pool, the salt concentration in each consecutive selection cycle is increased. Additional NaCl, which dissolves into Na^+ and Cl^-, increases the number of total ions in the solution. These ions consequently interact with both the target and the aptamer alike and must first be displaced in order to create a complex. Therefore, stronger binders are more likely of staying bound compared to their lower-affinity competitors. Of equal relevance, the ratio between target and aptamer concentration provides a crucial tool to regulate the overall selection pressure of a particular SELEX experiment. By iteratively reducing the amount of target molecules in the system, target-affine species are increasingly forced to compete against each other for the same binding site, adding to the probability for stronger binders to out-compete their opponents (see Figure 2.4 for an example of how a sudden change in target concentra-

tion affects the sequence composition of the pool). This concept of continued increase in selection pressure is also leveraged by the last factor, the incubation time. The longer the aptamers are incubated with the target, the more likely it becomes for lower-affinity species to bind the target surface. Hence, by reducing the incubation time with each consecutive cycle, selection towards high-affinity species can be achieved. Taken together, these selection parameters must be taken into account in subsequent *in silico* pipelines which include predicted secondary structure information in their model and which utilize data from multiple selection rounds.

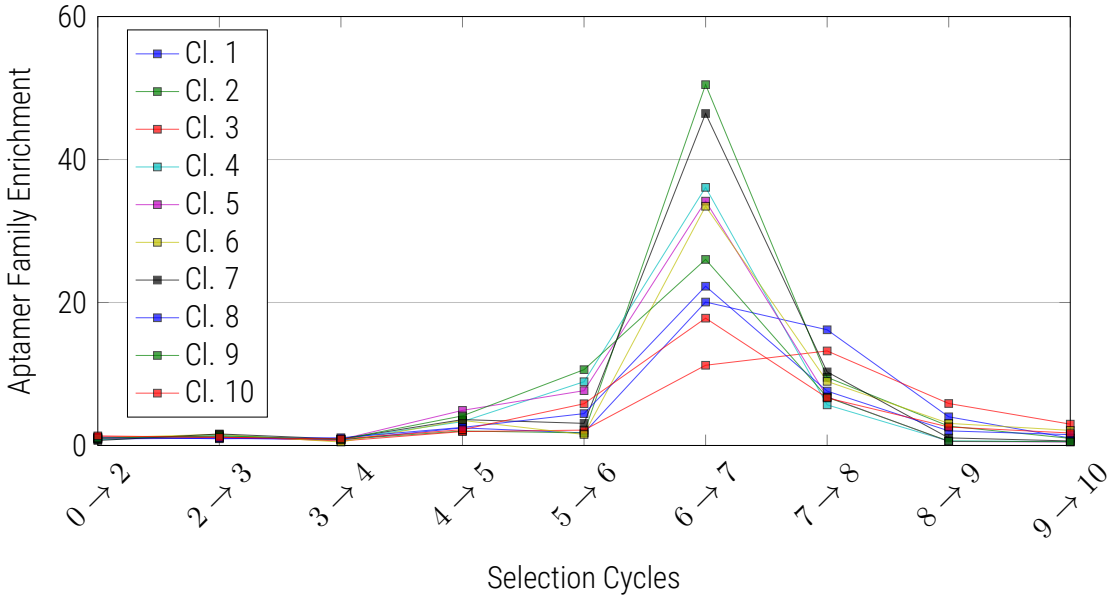


Figure 2.4: The effect of changing the target concentration during the selection on the example of a CELL-SELEX experiment performed against MUC16. Shown are the cycle-to-cycle enrichment values for the 10 most frequent aptamer families for 10 rounds of selection. The target concentration was kept constant for the first 5 selection rounds. In cycle 6, the target concentration was decreased by 50% resulting in a noticeable change of the aptamer landscape in the subsequent pool. Data Source: Rebecca Whelan Lab, Collaborator

2.4 Partitioning of Target-Bound Species

The incubation process as described in Section 2.3 results in a set of target-bound aptamers and a rather large fraction of unbound, non-affine species. In order to recover the target-bound fraction, the unbound species must first be removed from the solution. This is typically achieved by physically separating the two fractions. Depending on the target type, several experimental methods exist for the task, the most relevant of which include nitrocellulose filter binding, bead based partitioning, electrophoretic filtering, phosphate-buffered saline washing, and media based washing. A detailed description of the experimental proceeding regarding these methods can be found in Appendix A.1.

With the exception of electrophoretic filtering, the majority of these techniques allow for regulating the stringency of the partitioning process based on the washing time. In general, shorter washing times enable the inclusion of aptamer species with moderate binding affinities during the recovery whereas longer washing periods increase the likelihood of breaking the binding interaction with the target molecule, retaining only high-affinity aptamers. Analogous to the incubation procedure, this mechanism is routinely leveraged during the selection, by interactively increasing the washing time in subsequent selection cycles.

Notably, none of these procedures are capable of achieving a perfect separation of target-bound aptamers and non-binding species due to their experimental nature and currently available technologies. Consequently, a fraction of unbound aptamers remains in the resulting pool and is incubated with the target in the following selection round. In addition, due to the probabilistic nature of SELEX, some of the species might bind to the experimental equipment itself, such as beads or the surface of the plastic wells used during the incubation process. These non-specific binders can therefore accumulate in numbers throughout the selection and mask themselves as true binders. Taken together, this results in a considerable amount of low-frequency binders (non-affine species) and artificial signals (non-specific binders) when sequencing the pools for computational analysis and could be leveraged as background data for discriminatory approaches that attempt to predict high-affinity species *in silico*.

2.5 Elution of Target-Bound Species

After the recovery of the aptamer-target complexes from the incubation solution, the bound species are required be isolated (eluted) from the target in preparation of the subsequent amplification stage. Depending on the target, this can be accomplished through various methods including, but not limited to, washing the complex with a Sodium hydroxide (NaOH) buffer or, in case of proteins, by denaturation via heat treatment, adding the complex to deionized and RNase/DNase free water at up to $95°$C for 2-5 minutes.

Finally, the purified aptamer pool is recovered by separating it from the remaining molecules by mass via centrifugation.

2.6 Amplification of Aptamer Pools

The removal of non-binding species and consequent recovery of target-bound aptamers naturally results in a reduction of the overall number of species in the library. In order to prepare the pool for additional selection cycles, high-throughput sequencing, and storage, the sequences are subjected to a number of Polymerase chain reaction (PCR [49]) cycles in order to restore the library to its original size. In the case of RNA aptamers, these are first transcribed into cDNA.

Besides restoring the aptamer pool to its original size, the amplification step is also known to introduce variations of the original sequences into the selection process due to the error-prone nature of the polymerase chain reaction. These mutants represent a crucial element of SELEX as they not only increase the sampling of the sequence space, but might also exhibit enhanced binding properties to the target, hence opening the opportunity for the selection of higher-affinity aptamers. In fact, some variations of the SELEX protocol intentionally utilize hypermutagenic PCR with a mutation rate of up to 10 % per base in an effort to maximize the coverage of the sequence space [50]. Given a sufficient number of selection cycles, the repetitive amplification of the sequence pool consistently induces clusters of aptamers related to each other by primary structure and originating from a "seed" aptamer which was present in the initial library. In contrast to traditional SELEX, the data produced by HT-SELEX has enabled the detailed study and modeling of not only the mutational process itself, but also of the relationship between the mutation and its effect on aptamer properties such as binding affinity and specificity.

Notably, polymerases are also known to amplify certain sequences more efficiently than others in dependence of their nucleic content, structural properties, and the distribution of the PCR reagents in the solution buffer [51]. For example, sequences containing a high GC-content and consequently stable structure are less likely to be amplified compared to their thermodynamically less stable counterparts. This is caused due to GC-rich regions often forming stem-loop secondary structures that have been known to promote polymerase jumping during PCR amplification [52]. In the context of SELEX, this translates to a possible introduction of a bias throughout the selection in which the observed enrichment rate and frequency of target-affine aptamers in the pool do not correspond to their theoretical quantities under optimal amplification conditions. More importantly, aptamers with preferential properties for PCR amplification might be artificially enriched even though their binding strength to the target is weak. The mitigation of this issue is still an active field of research [47, 53], and innovative PCR technologies are being developed in order to counter this bias. One such approach relies on substituting the traditional PCR process with droplet (or digital) PCR [54]. Droplet PCR functions by compartmentalizing a sample of DNA or cDNA into many individual, parallel PCR reactions, each with their own reservoir of polymerase molecules and nucleotide reagents. The more uniform distribution of the latter therefore increases the probability of amplification for aptamer sequences despite their primary and secondary structure properties. In addition, the computational comparison of these technologies with traditional PCR in the context of SELEX can provide valuable insight into this bias and ultimately lead to improved *in silico* methods for the identification of high quality aptamers (see Figure 2.5 for a comparison of these methods).

2.7 Variations of the SELEX Protocol

The strategy of systematic evolution of ligands by exponential enrichment represents a general recipe for the selection of high-affinity aptamers against an arbitrary target of interest. In

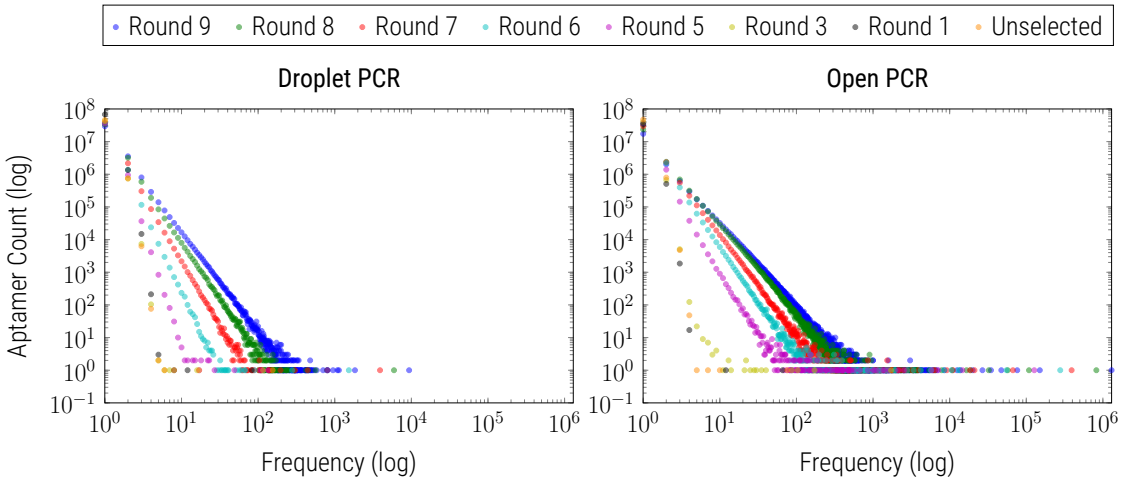


Figure 2.5: A comparison of droplet PCR (left) and open PCR (right) on the effect of the aptamer landscape based on a Cell-SELEX experiment against CCR7 over 9 selection cycles. Both experiments were performed in parallel based on identical initial pools, cell-lines, and selection conditions. Depicted are the number of aptamers (x-axis) with a particular count (y-axis) for the initial pool, and rounds 1, 3, 5, 6, 7, 8, and 9 respectively. The traditional PCR procedure produces an overall more exponential distribution of aptamer counts (long tail, x-axis) as compared to the droplet PCR. Note that this tail is dominated by a small number of highly amplified aptamer species. Data Source: John Burnett Lab, Collaborator

practice however, the implementation of that strategy is highly dependent on the properties of the target including, but not limited to, size, shape, charge, and type of the molecule. As a case in point, a selection against small compounds such as cocaine [55] is likely to require a differently fine tuned selection protocol as compared to designing aptamers against significantly larger proteins such as transcription factors [56] or even entire cell surfaces [57]. It is therefore not surprising that variations of the general protocol, adapted to the specific target properties, have been developed over time. These include peptide-SELEX, small-molecule SELEX, Genomic-SELEX and double stranded DNA SELEX (dsDNA-SELEX), Transcription factor SELEX (TF-SELEX), and Cell-SELEX, of which a detailed description of the most relevant protocols is outlined in Appendix A.2.

Naturally, the differences in the selection procedure are reflected in the resulting sequencing data and constitute a true challenge for the development of universal *in silico* approaches to aid the detection and refinement of aptamers that bind the target according to the scientists requirements. Prior to any computational analysis, it is therefore essential to adjust the raw output of modern HT-SELEX experiments and to filter possible contaminants, biases, and low quality aptamers from the sequencing data. For this purpose, we have developed an efficient computational preprocessing pipeline known as AptaPLEX, the details of which are outlined in Chapter 3.

3

Data Preprocessing

> *The nice thing about standards is that you have so many to choose from. Furthermore, if you do not like any of them, you can just wait for next year's model.*
>
> Andrew S. Tannenbaum, *Computer Networks*, 1981

3.1 Introduction

Chapter 2 illustrated the diversity and versatility of the SELEX protocol and highlighted the technological possibilities and limitations when coupled with high-throughput sequencing. Special emphasis was put on how different variations of the protocol as for instance TF-SELEX or CELL-SELEX affect the shape, enrichment properties, and background noise of the sequencing data. It is therefore required that prior to performing any computational analysis, the raw sequencing reads require extensive preprocessing such as contig assembly for paired-end data, aptamer extraction, and quality control in order to provide a consistent input for downstream analysis tools. More importantly, massively parallel sequencing technologies utilize a technique known as multiplexing in order to processes multiple samples from different experiments simultaneously by adding short barcode sequences (reference barcodes) to each sample such that these can be uniquely identified *in silico* (see Figure 3.1). Specific to HT-SELEX, the iterative nature of the protocol makes the protocol a natural candidate for this technology as it allows for multiplexing the different selection cycles from one or multiple SELEX experiments. In order to re-establish the membership of the aptamers to the selection cycles of origin, the individual reads are then demultiplexed by identifying which reference barcode best matches the sequenced barcode.

The task of demultiplexing is challenged by sequencing errors common in all currently available next-generation platforms. Traditionally, demultiplexers developed by Illumina, such as

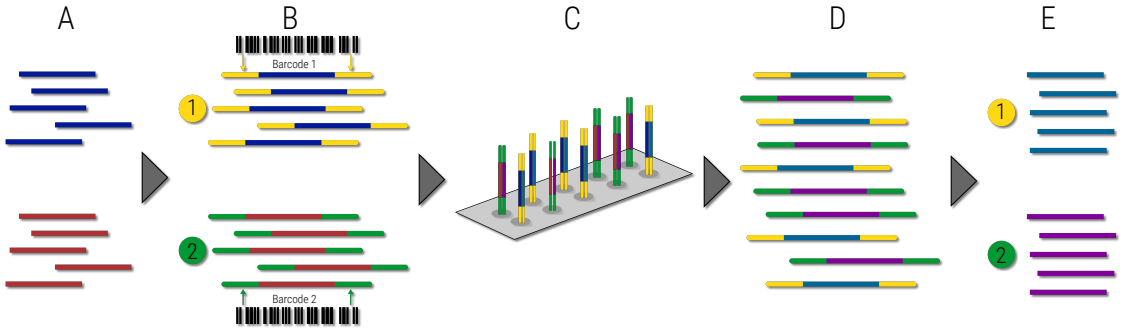


Figure 3.1: Conceptual overview of sample multiplexing. (A) Representative DNA fragments of two unique samples to be sequenced, e.g. belonging to different selection cycles of the same SELEX experiment. (B) Specific barcode sequences for each sample are ligated to the DNA fragments. (C) Libraries for each sample are pooled and sequenced in parallel. (D) Each new read contains both the fragment sequence and its sample-identifying barcode. (E) The barcode sequences are used to demultiplex, i.e. differentiate the reads from each sample *in silico*.

CASAVA and bcl2fastq2, allow for a user-defined number of nucleotide mismatches between the reference and sequenced barcodes as measured by the Hamming distance (see Section 3.2 for a formal definition). This concept was also adopted in form of exact overlap sequence alignment based techniques as implemented in FLEXBAR [58], which uses dynamic programming to calculate the optimal position and score of the reference barcodes with respect to the reads. In order to account for potential indels (insertion/deletion errors), this approach was further extended by deindexer [59] and Splitaake [60] to utilize the Levenshtein [61] distance. deML [62] on the other hand employs a maximum-likelihood technique based on the PHRED quality scores accompanying the reads in order to determine the probability of a reference barcode matching the sequenced barcode. Finally, Bayexer [63] leverages the information extracted directly from the error-containing sequences of the targeting reads for training a naïve Bayes classifier to assign barcodes. These procedures are frequently coupled into a single pipeline with the above mentioned, additional preprocessing steps.

While general-purpose demultiplexers yield satisfactory results in applications such as whole-genome sequencing or RNA-seq, none of these tools take the specific properties of HT-SELEX data into account: Besides the barcode, reads from a typical SELEX experiment are expected to contain a randomized region of predefined length flanked by constant primer regions required during the amplification stages of the protocol (see Fig. 3.2 B). Depending on the read length, additional nucleotides located after the read's 3' primer (predominantly sequencing adapters) might also be present and are required to be removed. Furthermore, contaminant sequences which do not contain an aptamer are considered noise and should consequently be excluded from downstream analysis. Finally, reads with above-average error rates in the primer regions might be indicative of error prone randomized regions and can therefore serve as an additional quality control mechanism. To date, only a limited number of solutions addressing these special cases have been published as integrated components of larger pipelines [64, 7, 3], making them unsuitable for independent, large-scale data processing and seamless integration into alternative software solutions.

To close this gap, we have developed AptaPLEX, a standalone and platform independent demultiplexer that is specifically designed for HT-SELEX data and other types of *in vitro* selections based on non-ligands such as DNAzymes and ribozymes. Given the multiplexed data from one or more selection experiments in either single-end or paired-end format, AptaPLEX partitions the reads into the individual selection cycles based on the barcodes used during sequencing. Simultaneously, AptaPLEX identifies the 5' primer and the 3' primer in each read and removes any additional nucleotides on either side that do not belong to the original aptamer. AptaPLEX is capable of fuzzy matching for both, the barcode and primers, allowing for a user-specified number of mismatches between the reference and the sequenced barcode and primer. Additionally, for paired-end data, AptaPLEX automatically corrects mismatches between forward and reverse reads up to a user-defined threshold. Our software provides a rich set of additional features such as the option to specify whether the primers should be trimmed from the reads (e.g. for motif searching applications), and to restrict the randomized region length to a constant size, hence discarding indel containing aptamers. AptaPLEX automatically recognizes and handles gzip compressed data and makes use of all available processing resources via its multi-threaded design while minimizing memory usage.

3.2 Methods

AptaPLEX takes as input a set of FASTQ files in either single-end or paired-end format, and partitions the identified aptamers into as many files as the number of specified barcodes. Our method uses a sliding window approach in combination with the Hamming distance as objective function for matching reference barcodes and primers to the reads. In what follows, we provide the definitions of the relevant concepts utilized by AptaPLEX and describe the general work flow of the algorithm for each read.

The FASTQ Format represents the *de-facto* standard format for representing sequencing data and its corresponding quality scores in a single file. The information for each read is coded into four lines, where the first line is reserved for a unique sequence identifier, the second line contains the raw sequence letters, followed by the third line with an additional but optional description of the read, and the last line encoding the quality values for the sequence in line 2. AptaPLEX is fully FASTQ compliant and outputs all sequences in this format such that these can serve as direct input to third party downstream analysis tools.

The Hamming Distance is an information theoretic measure for two strings of identical size defined as the number of positions at which these strings differ in their symbols. Similar to the approach first developed for CASAVA by Illumina, AptaPLEX utilizes this distance as an objective function to evaluate the quality of matching any of the reference barcodes and primers to a particular read position.

Paired-End Assembly: For paired-end data, our algorithm first assembles the full contig by generating the reverse compliment of the reverse read and by computing the optimal overlap between this read and the forward read (Fig 3.2 A). AptaPLEX automatically corrects up

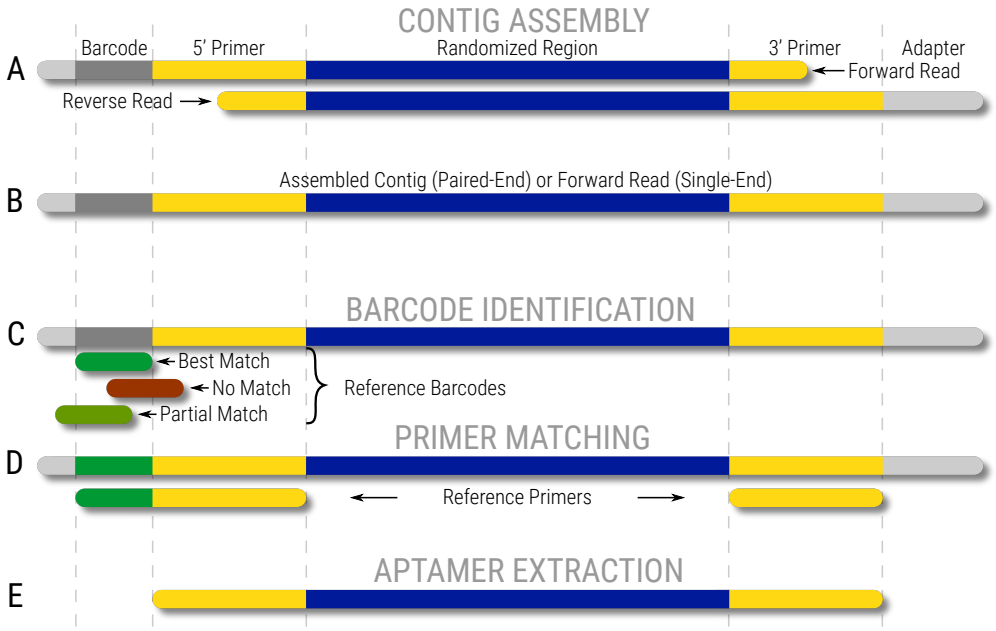


Figure 3.2: Schematic of the algorithmic workflow as employed by AptaPLEX. (A) For paired-end data, the forward and reverse reads are first assembled into a single contig (B). Next, the reference barcodes are matched to the contig (B) and the best correspondence is selected, followed by the identification of the primer regions (D). Finally, the aptamer is extracted from the contig (E).

to a user-defined number of mismatches between the two reads by choosing the nucleotide with higher PHRED quality score. Should additional mismatches be present, or in case no adequate overlap could be identified, the contig is discarded. When only single-end data is present, the forward read is used as the contig.

Barcode Identification: Next, our approach attempts to align each of the reference barcodes to the contig while tolerating up to m mismatches, and assigns the read to the selection cycle whose corresponding reference barcode retains the smallest Hamming distance to the sequenced barcode (Fig 3.2 C). This process makes no assumptions on the location of the barcode on the contig hence allowing it to be present on either side of the aptamer, even with the presence of additional nucleotides between the barcode and primer. To establish the validity of the identified barcode, AptaPLEX examines whether its location overlaps with the primer regions or the randomized region. Should this be the case, or equivalently if the number of mismatches exceeds the user-defined threshold m, the contig is discarded.

The optimally matching position is determined by sliding each reference barcode along the read in 5' to 3' direction in 1 nucleotide increments. For each position, the Hamming distance between the reference barcode and the corresponding subsequence of the read is computed, and the position with the smallest score is chosen as the best match.

Primer Matching: In order to extract only valid aptamers, our method proceeds to match both, the reference 5' primer, and the 3' primer to the contig. In analogy to the barcode matching procedure, a user-defined Hamming distance of up to p mismatches is allowed (Fig

3.2 D). Furthermore, AptaPLEX verifies the validity of these matches by ensuring that neither the barcode, nor the primers overlap in the contig.

Aptamer Extraction: Finally, the sub-sequence between the starting index of the 5' primer match and the end index of the 3' primer match, together with the corresponding quality scores, identifier, and comment line contained in the FASTQ format are extracted and written into the file corresponding to the identified barcode (Fig 3.2 E). Various applications, e.g. sequence based motif search, might not require the presence of the constant primer regions for their procedures. For these cases, AptaPLEX supports trimming the primer sequences to only extract the randomized region. Our method also has the option to discard those reads whose randomized region size does not coincide with the experimentally defined length.

3.3 Runtime Analysis

AptaPLEX is fully multithreaded and utilizes all available processing units on a given system. Parallelization is implemented on the read level such that paired-end assembly, barcode identification, primer matching, and aptamer extraction are performed concurrently for individual contigs. The expected speedup is hence dependent on the number of CPUs, as well as the total number of barcodes that are to be demultiplexed. In order to estimate the relationship between these parameters and the overall runtime, we benchmarked our approach on 2 million reads, utilizing 1 to 15 CPUs and 1, 3, 6, 9, 12, 15, and 18 barcodes respectively. We excluded the required time for disk operations as data I/O is known to be highly hardware dependent. Fig. 3.3 highlights the corresponding wall clock times of AptaPLEX averaged over 10 individual runs for each combination. On a modern system with 4 cores, our software is capable of processing at least 2 million reads per minute while demultiplexing up to 18 selection cycles, with throughput increasing even further when running AptaPLEX on more powerful systems.

3.4 Implementation Details and Availability

AptaPLEX is implemented as a multithreaded C++ library using openMP [65] and can easily be integrated into existing pipelines. Our software is platform independent and only requires the Boost libraries [66, 67] for its data I/O and compression operations. Our software is open-source, licensed under the GLP v.2 and can be downloaded along with manual and installation instructions at `http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#aptatools`.

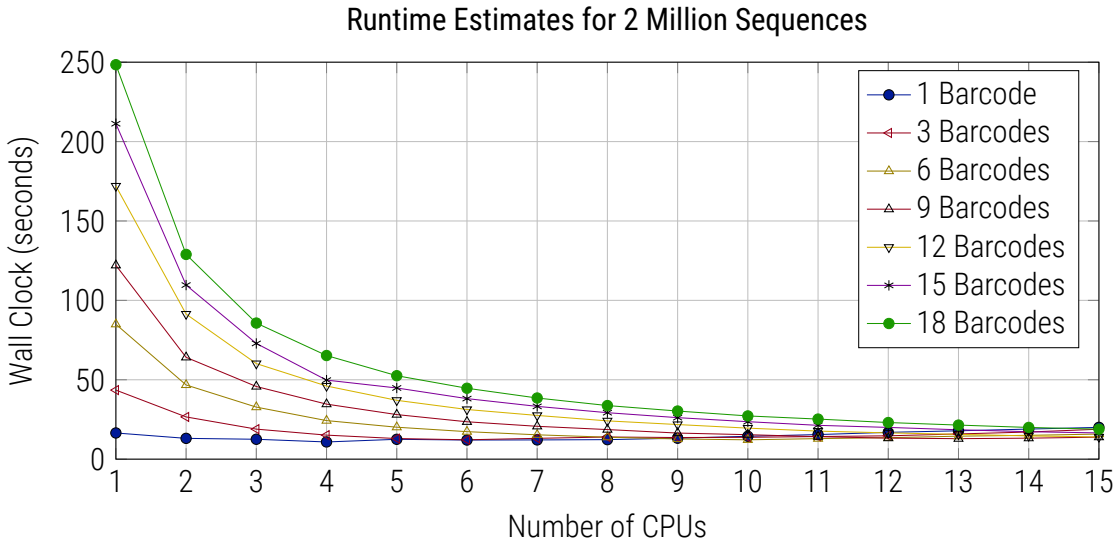


Figure 3.3: Runtime comparison of AptaPLEX at distinct hardware settings and increasing number of barcodes using 2 million reads as benchmark. Shown are the wall clock times averaged over 10 independent runs for each configuration, ranging from single core machines to 15 CPUs and extracting 1, 3, 6, 9, 12, 15, and 18 barcodes respectively.

3.5 Conclusion

With the integration of high throughput sequencing into *in vitro* selections being adopted as a de-facto component in modern pipelines, dedicated computational tools which streamline the analysis of the resulting data are becoming increasingly compulsory. This is especially true for the preprocessing stages of any computational pipeline, as the resulting data serve as the input for each consecutive analysis and therefore greatly impacts the quality and interpretation of these experiments. To date, only a limited number of tools for aptamer preparation, processing, and analysis have been published. These range from fully automated terminal based solutions [7] to web-based answers leveraging the Galaxy cloud platform for its operations [64]. While these tools contain demultiplexing and quality control capabilities as a subroutine, their design does not necessarily allow for automated preprocessing that can be integrated into larger pipelines without user intervention. AptaPLEX represents, to the best of our knowledge, the first stand alone demultiplexing software specifically designed for HT-SELEX data. Its well defined application programming interface (API) allows for full automation and trivial integration into arbitrary pipelines such as sequence based motif identification, structure based analysis, and exploration of the SELEX protocol as a whole. Alternatively, the resulting FASTQ files can be used directly as input for well established *in silico* tools in dependence on the researchers interest. Furthermore, its platform-independent and fully multi-threaded implementation makes AptaPLEX a suitable candidate for massive data processing on different architectures such as computer farms and cloud based solutions.

4

New Algorithms for HT-SELEX Data Analysis

> *Being abstract is something profoundly different from being vague. The purpose of abstraction is to create a new semantic level in which one can be absolutely precise.*
>
> Ew Dijkstra, *2012*

4.1 Introduction

Chapter 2 introduced the various types of Systematic Evolution of Ligands by Exponential Enrichment protocols that are at a scientists disposal in order to design aptamers with the desired properties suitable for their intended biotechnological application. It also highlighted a number of properties of the resulting high throughput sequencing reads, among others, the target-dependent complexity of these data in terms of their confounding factors, the estimated proportion of high affinity aptamers, and the expected enrichment rate of the species. Chapter 3 provided a partial solution to the first challenge by detailing a framework for the standardization of preprocessing aptamer reads while embracing strict quality control routines. Processing the data from a SELEX experiment in this matter hence results in a subset of sequences, typically between 1 to 50 million reads for each sequenced selection cycle, representing a tiny fraction of the true pool size of approximately 10^{15} individual RNA molecules.

In this chapter, we introduce several new computational methods designed to answer four central challenges in the field of aptamer research which have, until now, only been postulated but not rigorously studied. Specifically, we developed solutions to (a) how to study the effects of the selection pressures described in Chapter 2, (b) how to efficiently identify, cluster, and trace aptamer families throughout the selection to aid target-affine ligand discovery, (c) how to leverage the (hyper)mutagenic PCR during the amplification stages of the protocol in order to determine aptamers with improved binding properties, and (d) how to elucidate common sequence-structure motifs responsible for binding the target and shared by multi-

ple aptamer families. All our methods take advantage of the complete data set produced by modern HT-SELEX pipelines, consisting of sequencing data from all selection cycles of an experiment.

First, in order to study the effect of the different selection pressures on the pool composition, we developed AptaSIM. This software, capable of producing data sizes comparable to the output of next-generation sequencing technologies, is aimed at realistically recreating the general purpose SELEX protocol using error-prone PCR and can simulate many aspects of *in vitro* experiments such as sampling effects, the presence or absence of pool contaminants, and aptamer affinity (see Section 4.2). Understanding the relation between these pressures and the target complexity can ultimately lead to the design of target-optimized protocols capable of producing higher affinity aptamers while reducing the number of required selection cycles which tend to be very cost intensive and time consuming to perform. Using AptaSIM not only allows to quantify several of those parameters such as the PCR error rate in real data, but additionally serves as a benchmarking and validation system for the remaining approaches presented in this chapter.

Next, in Chapter 4.3, we introduce AptaCLUSTER, a novel technique to efficiently cluster aptamer families in each selection round and to trace their behavior throughout consecutive selection cycles. In contrast to traditional clustering techniques, AptaCLUSTER scales well with next generation sequencing data and incorporates a variety of quality control mechanisms designed to maximize data quality for downstream analysis. In addition, we show that AptaCLUSTER can reveal properties of the selection process which have previously not been appreciated and which can be utilized for an optimized aptamer discovery process. The development of such a method is of great relevance because to date, only a limited number of computational approaches centered around the identification of aptamer candidates from high throughput sequencing data have been published. Aptamer identification has traditionally been performed based on simplistic counting techniques in which sequences from the final round of selection are collapsed into the set of unique species along with their corresponding cardinality followed by *in vitro* binding assays of the top most enriched candidates in order to assess their target affinity [36, 68, 69] (see Figure 4.1). While these methods have proven successful in the generation of aptamers targeting transcription factors and other proteins evolutionarily designed to form DNA or RNA complexes (TF-SELEX), they suffer from various drawbacks when analyzing data generated by more complex SELEX experiments (i.e. CELL-SELEX) in which the number of non-specific binders, external contaminants, and other influences (see Chapter 2 for a comprehensive list) can dramatically bias the concentration of aptamer species in the later rounds [7, 4, 53, 47].

Building on the results of AptaCLUSTER we then provide an unprecedented in-depth analysis of the mutational landscape of HT-SELEX experiments and propose a mathematical model capable of discriminating favorable mutants from those which decrease the binding affinity to the target. We have validated our model experimentally and have implemented it in form of our AptaMUT algorithm as described in Section 4.4. Aptamer mutants are introduced into the pool, at each cycle, due to the error-prone nature of polymerases required during the amplification stage of a selection round. In select cases, the resulting mutant exhibits an increased target affinity as compared to the wild type and is consequently enriched and carried

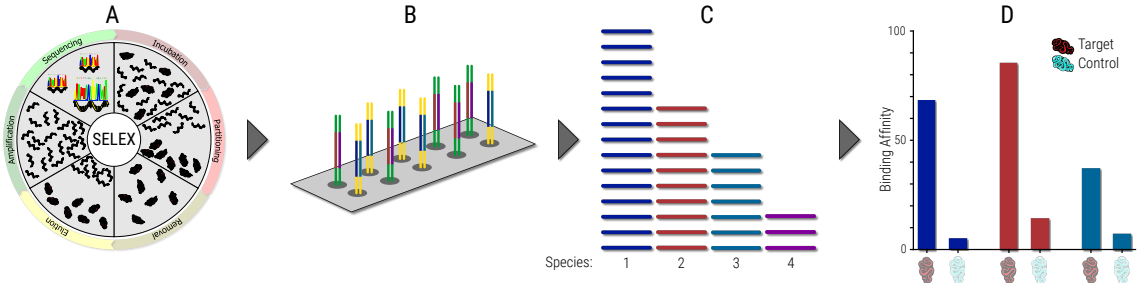


Figure 4.1: Schematic overview of primitive counting techniques. (A) During SELEX, only the last selection round is retained. (B) The species in the last round are sequenced using next-generation technology. (C) Individual aptamer sequences are consequently counted and sorted according to their frequencies. (D) A small number of high-frequency aptamers are selected and tested *in vitro* for binding.

over to consecutive rounds. Due to the iterative nature of the SELEX protocol, the continuous enrichment over time results in the creation of entire aptamer families in the pool. More importantly, because these mutants emerge at arbitrary selection cycles, they might be present in only very low copies in the final pool, posing a challenging task to discriminate these from background noise. The ability to identify beneficial mutants and well as discriminating these from their detrimental counterparts not only enables the generation of more powerful aptamers, but also provides valuable insight into the binding interaction mechanisms with the target for any given aptamer family.

Finally, we tackle the challenging task of sequence-structure motif identification in HT-SELEX data in form of our AptaTRACE algorithm. AptaTRACE is built on tracing the dynamics of the SELEX process itself to uncover motifs induced by (or emerging during) the selection procedure, is robust enough to be applicable to a broad spectrum of RNA/ssDNA HT-SELEX experiments independent of the target's properties, and capable of elucidating an arbitrary number of binding sites along with their corresponding structural preferences. Furthermore, our method is corroborated by *in vitro* binding assays and can be used to reveal previously underestimated properties of the selection process which could ultimately lead to time-and-cost optimized SELEX protocols. It has been empirically shown that the affinity of RNA to proteins is a function of both sequence and structure, typically localized to a specific region of the molecule [70, 71]. Understanding the properties of these binding motifs in aptamers, the importance and relationship of their sequence and structure components, as well as how these evolve throughout the selection is therefore crucial for the design of efficient binders and optimization of the protocol per-se. Furthermore, the application of well established motif identification methods such as MEME [72] , MEMERIS [73], DREME [74], or RNAContext [75] to HT-SELEX data are currently limited to utilizing the data from a single round of selection and, depending on the underlying model, the incorporation of an earlier cycle to use as background data. Moreover, the mathematical basis of most of these methods, and therefore their underlying assumptions regarding the expected properties and size of the input data, are mainly concerned with finding conserved regions within genomic sequences, limiting their transferability to general purpose SELEX experiments.

4.2 AptaSIM: Realistic, *in silico* Simulation of SELEX Experiments

HT-SELEX is revolutionizing the aptamer discovery field by providing researchers with unprecedented resolution of their selection experiments. At the same time, the amount of data, as well as the variability of its landscape induced by the type of SELEX and target properties pose a true barrier for the generation of *in vitro* benchmarking data containing high-throughout binding information for the majority of the pool species and/or near-complete elucidation of binding motifs. The prohibitively high cost of performing these experiments hence calls for alternative solutions which allow for fine grade control and supervision of as many aspects of the SELEX method as possible.

To close this gap, we have developed AptaSIM, an algorithm aimed at realistically recreating the selection process during SELEX while simulating the effect of a large array of user definable parameters on the pool. These include, but are not limited to, error-prone PCR, the target affinity of individual aptamer species, the effect of sampling the pool for sequencing and subsequent selection cycles, and the presence or absence of initial pool biases. In addition, AptaSIM supports the inclusion of sequence-structure motifs which can be implanted into arbitrary species in the initial library.

For our simulation, we represent a pool as a set of sequences in which each sequence is attributed with a count, representing its frequency, and a value between 0 and 100 simulating the binding affinity to a putative target. Given an initial pool, we then perform a user-defined number of iterations comprised of target affine selection, followed by error prone amplification. The remaining sequences after the selection stage represent the sequenced portion of HT-SELEX and are stored for further analysis.

4.2.1 Materials and Methods

Initial Pool Generation: To allow for the inclusion of existing biases such as the base composition and nucleotide dependencies of a pool originating from an in-vitro SELEX experiment, the input set of sequences for the simulation is generated based on a first order Markov Chain that captures the conditional probabilities of randomly selecting one nucleotide given the choice of the previous. Each sequence is then assembled by randomly selecting the first nucleotide with respect to the base composition of the training data and iteratively sampling the remaining nucleotides according to the conditional distributions of the model. In addition, each sequence is assigned a random initial count $c_i \in [0 \dots c]$, as well as a binding affinity $b_a \in [0 \dots b]$), where b and c correspond to user-definable values. Finally, we represent strong binders by selecting s arbitrary sequences for which the binding affinity is uniformly sampled between user defined values s_l and s_u where $b < s_l < s_u$.

Target Affine Sampling: The sampling step simulates incubation, binding, partitioning, and washing of a selection cycle during a SELEX experiment. Assuming enriched and target affine species to have a higher probability of selection, we sample, without replacement, p_s% of the current pool according to the distribution of the sequence counts and accept a sequence with the probability corresponding to its binding affinity. Hence, the probability of selecting a sequence from the pool is proportional to its frequency and affinity.

Amplification: In order to restore the pool to its original size, we simulate a number of PCR cycles in which the amplification efficiency $e \in [0, 1]$ as well as the mutation probability $p \in [0, 1]$ can be specified. The number of required PCR cycles c is computed as follows:

$$c = \left\lceil \frac{\log\left(\frac{i}{x}\right)}{\log(1+e)} \right\rceil$$

where i and x correspond to the sizes of the initial and current pool respectively. In each PCR cycle, every aptamer is then subject to amplification as many times as its current count and in dependency of the specified probability of amplification e. If accepted, and based on the mutation probability p, the sequence is either duplicated or a mutant, differing by one base from the original at a random position, is introduced into the pool.

Default Parameters and Implementation Details: An initial pool of $N = 10^7$ unique sequences of size $r = 40$ nt is generated containing approximately $s = 100$ high affinity binders. AptaSIM performs $k = 10$ rounds of selections by default during which, at each sampling step, $p_s = 20\%$ of the pool is retained. For amplification, a mutation probability of $p = 0.05$ and an amplification efficiency of $e = 0.95$ is used for realistically recreating the pool characteristics of a typical *in vitro* experiment. A complete list of all available parameters as well as their default values can be found in Supplementary Table B.1 in the Appendix. AptaSIM is implemented in a platform independent manner using the Java programming language.

4.2.2 Results and Discussion

AptaSIM aims at closing several gaps which are naturally enabled when coupling high-throughput sequencing technologies with Systematic Evolution of Ligand by Exponential Enrichment.

First, it allows for the analysis of nucleotide biases which might be present in the initial pool. By training a Markov model with aptamers from a sequenced initial round (or any other selection cycle of interest), the effect of these biases onto the pool composition in subsequent cycles can be studied in detail and under varying conditions.

Next, our model allows to study the emergence of mutants during the selection in high resolution via AptaSIMs mutagenesis driven PCR simulation. As a point in case, we leveraged

this feature to show that the species composition of the most dominant aptamer family belonging to a selection targeting IL-10 is indeed consistent with error prone amplification and subsequent selection (see Section 4.4.2 for a detailed description).

Finally, our approach makes for an appropriate candidate to generate benchmarking datasets with precise properties in order to aid the development of future algorithms and methods, or to compare existing approaches with each other under well defined conditions. We took advantage of this capability in order calibrate our motif identification approach AptaTRACE and to compare it to existing methods. Specifically, we generated an initial pool of approximately 4 Mio. sequences and enriched it with predefined sequence-structure motifs prior to performing 10 rounds of virtual selection. We then probed the resulting predictions of our method and comparable algorithms for their ability of extracting these implanted motifs. For the full application and methodology, we refer the reader to Section 4.5.

However, the affinity of the ligands, and consequently their probability of selection during the simulated incubation step, is currently based on the affinity parameter b alone and therefore independent of primary, secondary and/or tertiary structure information. In detail, the current simulation model assumes that any mutations introduced during the amplification stages do not affect the binding strength of the aptamer to the target. This assumption might not always reflect the experimental reality of mutagenesis-driven PCR as point-mutations in the sequence can have structural changes as a consequence which could in turn affect the binding affinity to the target. Incorporating such higher order information regarding the interaction between aptamer and target could potentially increase the accuracy of simulation systems such as AptaSIM, but at the cost of substantially higher runtime complexity. Furthermore, additional structural information regarding the target molecule must be known *a priori* and is often not available to the researcher.

Related to structure, our model is insensitive to temperature constrains which are known to affect the selection as a whole, but are currently not well understood. Despite these shortcomings, our results provide sufficient evidence that AptaSIM can serve as a valuable tool for aptamer-related research and to uncover and quantify selection pressures in HT-SELEX experiments. In summary, AptaSIM offers a sufficiently large array of parameters which control many of the known biases and selection pressures present when performing SELEX.

4.3 AptaCLUSTER: Efficient Clustering of HT-SELEX Data

4.3.1 Introduction

Computational processing of HT-SELEX data is currently largely based on simple counting of aptamer species in the final round of selection, frequently discarding low-frequency species from the analysis, and choosing the sequences that occur in high counts for further investigation [36]. In addition, a small number of most frequent sequences from the final selection round might be used as seeds for similarity searches. The underlying postulate of these methods is that the best predictor of binding affinity corresponds to the frequency at which a particular aptamer occurs in a pool. While these approaches might be suitable for candidate identification, they lack the ability of providing insight into the mechanisms governing the selection process itself. Uncovering these embedded characteristics requires more sophisticated models and perhaps an alternative, non-quantitative perspective of the selection process.

One approach providing a more detailed intuition of the SELEX protocol is motivated by a model of the selection process from the perspective of the binding energy between the target and the species in the pool and captures how the resulting landscape changes over time. For a SELEX experiment with a randomized region size of n nucleotides, this can be visualized by an arbitrary mapping of all 4^n possible species into a two dimensional space, where the distance between these aptamers on the surface is a function of their primary structure similarity. A third dimension (z-axis) can then be used to represent the binding affinity of any such sequence to a target molecule of interest (see Figure 4.2 A). *In vitro*, the total number of ssDNA/RNA molecules ranges between $10^{15} - 10^{18}$ and corresponds to a sparse and ideally equidistant sampling of the total sequence space in this model. As selection progresses, aptamer species with little to no binding affinity to the target (high z-axis) are competed out of the pool, whereas strong binders (low z-axis) are retained and amplified. More importantly, due to the error-prone nature of the amplification stage in each SELEX round, mutants are introduced into the pool that explore the topological neighborhood around their parent sequence (Figure 4.2 B). Repeating this process over multiple rounds of selection, in which aptamers harboring favorable mutations increase in frequency and are themselves subject to error prone amplification, therefore results in a configuration in which the species remaining in the final pool sample local minima of the energy landscape. These minima correspond to the distinct binding sites the target molecule exposes on its surface and which are accessible to the aptamer species (Figure 4.2 C).

In this context, the mutant-induced sampling of local binding energy minima naturally motivates the expectation that clustering of aptamers in consecutive cycles should provide valuable information about the selection process and should allow for the delineation of the entire aptamer landscape probed by the SELEX protocol. Hence, our primary objective is to cluster aptamers in all rounds of selection according to their sequence similarity. This task however could not be accomplished with previous clustering algorithms due to the enormous size (2-50 Million sequences per cycle) of the data set generated by high throughput

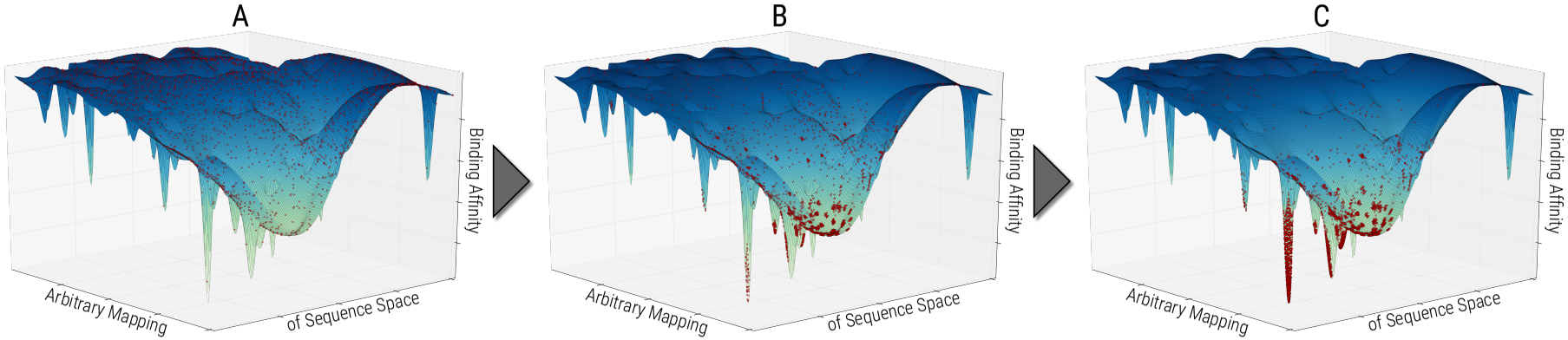


Figure 4.2: A visualization of the aptamer landscape probed by the SELEX protocol at different rounds of selection. The surface represents all possible aptamers of fixed length and the red dots represent the aptamers present in the pools during selection. The distance on the surface is a conceptual projection of sequence similarity. Multiple local minima correspond to groups of aptamers that bind to the different areas of the targets surface or to the same region but are related by structure rather than sequence similarity. The z-axis corresponds to the binding affinity of an aptamer to the target molecule. (A) The pool composition at the initial state of SELEX. Sequences in the initial pool are a uniform but scarce sample from all 4^n possible species, where n stands for the randomized region size. (B) As the selection progresses, non-binders (high z-values) are competed out of the pool, whereas target-affine species (low z-values) are amplified, introducing mutants into the pool that explore the topological neighborhood around their parent sequence. (C) Landscape at the final round of selection, showing the final aptamer families as the result of repeated selection and error-prone amplification.

sequencing, especially for early rounds of selection which feature a high degree of unique sequences ($\geq 90\%$). To address this challenge, we developed a novel approach, AptaCLUSTER, capable of efficiently clustering entire aptamer pools.

Several sequence similarity measures are commonly found in clustering methods, of which the Hamming and Levenshtein (edit) distances are most prominent. However, full-scale clustering approaches are computationally untrackable for HT-SELEX data. Therefore we use the randomized dimensionality reduction technique, known as locality-sensitive hashing (LSH) [76], to implicitly approximate an upper bound to the edit distance for each sequence pair without the need of exhaustive pairwise comparison. In the subsequent step, we eventually compute precise sequence distances based on k-mer counting between pairs of aptamers below this bound, while the remaining distances are not relevant and might be arbitrarily assumed to be infinity.

We applied AptaCluster to analyze the results of the HT-SELEX experiment that we preformed using Interleukin 10 receptor alpha chain (IL-10) as the target molecule. IL-10 is considered to be a master regulator of immunity to infection and is an important therapeutic molecular target [77]. We preformed 5 cycles of HT-SELEX with a 40nt variable region, sequencing the samples of pools 2-5. AptaCluster has enabled us to analyze the results of HT-SELEX, revealed interesting properties of the selection landscape, and allowed for a better understanding of the HT-SELEX experiment. AptaCluster scales very well with data size. While the sequenced pools in our IL-10RA HT-SELEX experiment varies between 2 and 4.5 Million aptamers, we have applied AptaCluster to much larger pools of more than 20 Million sequences in the context of whole-cell HT-SELEX (data not shown) without loss of noticeable performance.

4.3.2 Materials and Methods

We performed 4 rounds of selection and cDNA generated from round 5 bound fractions as well as RNA recovered from bound fractions at rounds 2, 3 and 4 was amplified and sequenced using Illumina's HiSeq 2500 device with the 100-cycle paired-end sequencing protocol. Aptamers were then demultiplexed and preprocessed using AptaPLEX as described in Chapter 3. For the entire experiment, a total of 12.895.554 sequences where retrieved of which 4.621.438 species belonged to round 5, 1.923.823 to round 4, 2.181.720 to round 3, and 4.168.573 to round 2. Out of these respectively 617.220, 1.021.668, 1.902.904, and 3.857.210 were unique.

The experimental details and supplementary information of the selection are described in the Appendix Chapter B.2.

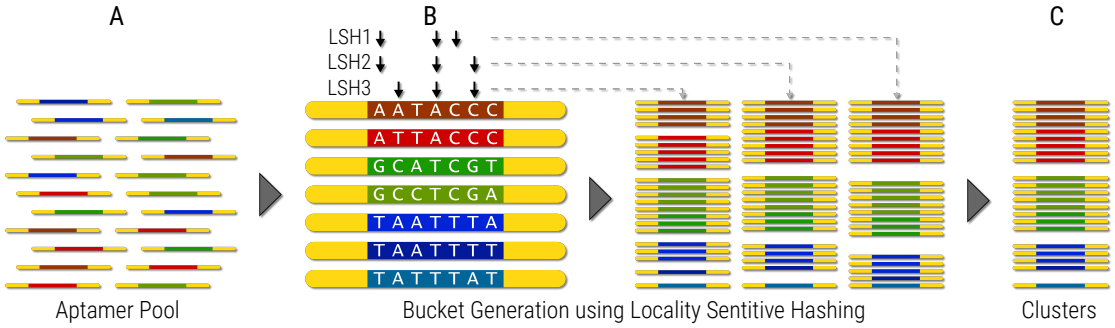


Figure 4.3: Algorithmic approach of AptaCLUSTER. Each colored sequence represents a distinct species in the pool while similar colors stand for aptamers related to each other in primary structure. The constant primer regions (yellow) are not considered during the computation of the clusters. **(A)** To better illustrate the concept, we included identical sequences (in the actual implementation identical aptamers are represented by their sequence and corresponding count). **(B)** AptaCLUSTER iteratively partitions the pool into sets of potentially similar sequences using the concept of locality sensitive hashing. In each iteration, an input to a hash function is generated by sampling a user defined number of nucleotide positions (black arrows). **(C)** Similar sets (e.g. green grouping on the left) from the different iterations are then combined such that sequences below a certain threshold are grouped together forming the desired clusters. This is accomplished by jointly clustering these groups using an exact distance measure based on the k-mer distance as show in Eq. 4.6

4.3.3 The AptaCluster Algorithm

Our approach is centered around a randomized dimensionality reduction technique, known as locality-sensitive hashing (LSH) [76]. First, a compressed representation of the data set is constructed by reducing the pool to non-redundant species and their corresponding frequency counts. We then apply a user-defined number of randomized locality-sensitive hash functions to the data set in order to distinguish sequence pairs that are potentially similar from those that are, with very high probability, not similar. Each function operates by selecting a small number of nucleotide positions from each aptamer and treats the substring, resulting from the concatenation of these bases, as input for the hashing procedure. Hence, aptamers with highly similar primary structure are likely to fall into the same group whereas dissimilar sequences rarely produce identical hash values. In the third step, the actual clustering step, we compute precise sequence distances between aptamers of identical hash value, while the distances between the aptamers never encountered in the same group are set to infinity. To accelerate the clustering, AptaCluster relies on a similarity measure based on k-mer counting. Thus, the algorithm preforms three main steps as outlined below and as visualized in Figure 4.3.

Dataset Compression

Data compression is achieved by using a hash map in which the keys correspond to the species in the pool and the values correspond to their respective frequency counts which can be done in $\mathcal{O}(N)$ time. In the following, let $s = (s_i)_{i=1}^{l}$ be an aptamer of of length l

defined by the sequence of nucleotides s_i over the alphabet $\Omega = \{A, C, G, T\}$ where the index i corresponds to the i^{th} position of the aptamer. Furthermore, we define $S = \{s^j \in P \parallel s^j \neq s^k \, \forall j, k \in [1, \ldots, |S|] \wedge \sum_{j=1}^{|S|} m(s^j) = N\}$, where $m(s^j)$ corresponds to the frequency of s^i, as the keys of the hash map, i.e. the set of unique aptamers for pool P.

Filtering using Locality Sensitive Hashing

LSH is based on the idea that data points that are close in high dimension, after applying a probabilistic dimensionality reduction and using the reduced representation as the input to a hash function, are likely to obtain the same hash value and hence fall into the same bucket [78].

AptaCluster exploits this property by treating each sequence $s^j \in S$ as an l-dimensional vector and reducing this vector into d dimensions ($d < l$). This is done by generating a set I_d of d randomly sampled indices $i \in [1, \ldots, l]$ and, for each sequence s^j, only selecting those nucleotides s_i for which $i \in I_d$ as input for the hashing procedure. Hence, the more similar the primary structure of a set of aptamers, the higher the probability that they will produce the same mapping. Similarly, the choice of d controls the minimal degree of similarity between the members of each partition since these are guaranteed to differ in at most $l - d$ positions. In other words, our approach implicitly computes an upper bound to the edit distance. We iteratively improve this upper bound by repeating this procedure a user defined number of times, each time using a different hash function. With sufficient number of of iterations, if two sequences never fall into the same bucket they are assumed to be dissimilar with very high probability. The iterative computation of the upper bound is performed as follows. Let $d_{lsh}^k(s^1, s^2)$ be the upper bound computed after the k^{th} iteration and let $L^k(s)$ be the value of the k^{th} hash function for sequence s. We assume that, by default, we have for all pairs $d_{lsh}^0(s^1, s^2) = \infty$. Then

$$d_{lsh}^k(s^1, s^2) = \begin{cases} l - d & L^k(s^1) = L^k(s^2) \\ d_{lsh}^{k-1}(s^1, s^2) & L^k(s^1) \neq L^k(s^2) \end{cases} \tag{4.1}$$

Clearly, only the assignment in the first line needs to be executed. To define $L^k(s)$, for each iteration k we randomly select a mapping h from a family of functions

$$F = \{h : \mathbb{N}^l \rightarrow \mathbb{N}^d \parallel h(I) = I_d\} \tag{4.2}$$

where $I = (1, \ldots, l)$ represents the nucleotide positions of an aptamer of size l, and apply the function

$$L = \{\Omega^l \rightarrow \Omega^d \parallel L(s) = (s_i) \, \forall \, i \in I_d\} \tag{4.3}$$

to each aptamer s, creating a sub-string $\hat{s}$ comprised of the concatenation of the nucleotides at the positions defined in I_d. Finally, traditional hashing is performed on the set $\hat{S} = \{\hat{s}^i\}, i = 1, \ldots, |S|$. $I_d = (i_0, \ldots, i_d)$ can be efficiently computed as follows: Let $i_0 \in [1, l]$ be a randomly selected index of I and define $x \in [2, l - 1]$ as a random number co-prime to l. Then, the remaining positions can be generated with

$$i_j = (i_{j-1} + x) \mod l, \quad j = 1, \ldots, d - 1 \tag{4.4}$$

and

$$I_d = (i_j)_{j=0}^{d-1}, \quad i_j < i_{j+1} \, \forall \, j \tag{4.5}$$

corresponds to the sequence of indices after sorting these in ascending order. Using this scheme guarantees that each index in I is selected exactly once and avoids scenarios in which only adjacent positions of the sequence are chosen.

Cluster Extraction

Based on the assumption that high-frequency of a sequence in a selection pool is related to its selective advantage due to its binding affinity, we build the clusters iteratively around these high frequency aptamers. We repeatedly choose the highest frequency sequence s not assigned to any cluster, making it a seed of the new cluster. We then we employ a k-mer based distance function [79] to compute the distance of the selected seeds to all other sequences for which the upper bound estimated with LSH was finite and include it in the cluster if d_{kmer} is smaller than a user defined cutoff. In particular,

$$d_{kmer}(s^x, s^y) = \sum_{i=1}^{4k} \left| \frac{X_i}{|s^x| - k + 1} - \frac{Y_i}{|s^y| - k + 1} \right|^2 \tag{4.6}$$

where X_i and Y_i denotes the number of times the i-th k-mer occurs in sequence s^x and s^y respectively and $|s^i|$ corresponds to the length of the aptamer. Since we compare only sequences that are in the same bucket in at least one iteration, this approach allows us to extract clusters in $\mathcal{O}(N*m*k)$ where m denotes to the maximum number of seed sequences in a bucket which is bounded by the size of the largest bucket generated during LSH.

Implementation Details

AptaCLUSTER is currently available as a multi-threaded implementation in C++ using the OpenMP and Boost libraries for its parallel programming operations and hashing procedures, respectively [65, 67]. It features a complete, highly modular pipeline from data input and parsing, over cluster extraction, to result visualization and database storage. We implemented threaded parsers for a number of file formats, including FASTA, FASTQ, and RAW sequence files by integrating the preprocessing library written for AptaPLEXinto this software. Depending on the number of available CPUs, clustering and distance calculations are performed in parallel for each pool. Cluster families and their evolution from cycle to cycle are currently visualized using our platform independent AptaGUI interface (See Chapter 5). Finally, the algorithms behavior can be controlled using a configuration file allowing for the assignment of most parameters used for parsing and clustering, among others. We have empirically determined a set of default values, of which the most relevant are discussed below.

4.3.4 Results of Application to HT-SELEX Experiment for IL-10RA

We performed 5 rounds of HT-SELEX against Interleukin 10 receptor alpha chain (IL-10RA) as the target molecule carrying out a total of $r = 10$ iterations of LSH sampling 60% of the randomized region (i.e. $l = 24$). The parameter $d = 4$ is set in terms of the maximal number of point mutations any pair of sequences should have and is converted into the k-mer distance cutoff by sampling a user defined number of aptamers from the pool (10000 by default), artificially mutating that sequence up to d times, and averaging over all d_{kmer} between these mutants and the wild-type. Furthermore we set $k = 3$ for the computation of d_{kmer} which has shown to give reasonable results for aptamer-sized sequences. Here, we summarize the insights obtained using AptaCluster.

Validating Clustering Results

The main advantage of AptaCLUSTER is that to cluster an aptamer pool it does not need to compute the distances between all pairs of sequences but instead uses locality-sensitive hashing to filter out pairs that do not need to be compared. However, the filtering step is heuristic and its outcome might depend on the number of LSH iterations and properties of the dataset. Therefore we started by confirming that the filtering step produces correct results, i.e. that sequences filtered out as not potentially similar are indeed remote from the seed sequences in terms of exact distance. Since the dataset size prohibits an exhaustive computation of all distances, we used 400 aptamers (the 20 most frequent species from the top 20 clusters) and computed their edit distances to all other aptamers. We then computed the distribution of the distances to the members of the same cluster to the distances to the rest of the aptamers. The former group sampled the sequences whose distances to the reference sequences has been computed and found to be below the clustering threshold. The latter group sampled two types of sequences: the sequences whose distance to the reference sequence has been computed but found to be above the threshold and the sequences filtered out without computing the distance based on our locality sensitive hashing function. The results for all selection cycles are summarized in Figure 4.4 for a set of default and relaxed parameters (see Parameters section). The results demonstrate that no sequence that was filtered out using locality sensitive hashing is close to the seed sequences of the clusters. In addition, it also demonstrates that SELEX derived aptamer clusters are well separated. Indeed, relaxing the locality-sensitive hashing based filtering and increasing clustering threshold did not change the clustering results appreciatively (Figure 4.4 (b)).

Cluster Accuracy and Reproducibility

We tested the accuracy and the reproducibility of our approach with respect to the distance computations using data from our IL-10RA experiment. Specifically, we used the 20 top clusters reported by AptaCLUSTER and determined the k-mer distances of all the cluster seeds to all other aptamers in the pool for different values of LSH iterations. We then calculated the

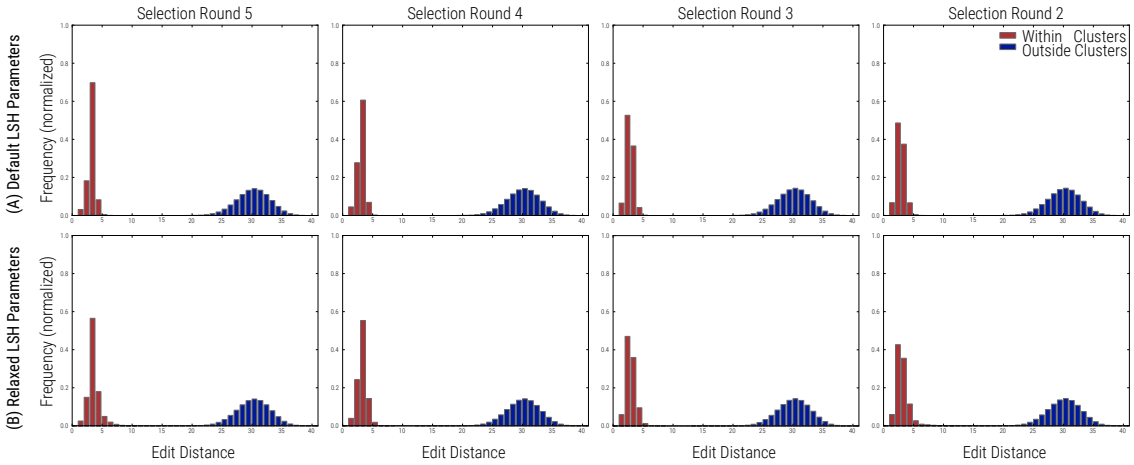


Figure 4.4: Distribution of the edit distances between aptamers belonging to a cluster (red) and distances between cluster members and all non-cluster sequences (blue) for selection rounds 2 to 5. Within each of the top 20 clusters, the 20 most frequent aptamers where compared against all other cluster members as well as the remaining aptamers of the pool. (A) Distributions using the defaults parameters of AptaCLUSTER as described in the Parameter section is shown in the top panel. (B) Relaxed parameters as depicted in the bottom panel in which only 40% of the randomized region was sampled during LSH.

false negatives rate (FNR) where a sequence pair considered a false negative refers to a sequence which has a distance to the seed below the specified clustering threshold but was assigned an "infinity" value. We found an on average overly low false negative rate varying between 10^{-6} to 10^{-4} as illustrated in Figure 4.5.

Distribution of Aptamers within Clusters

Next, we examined the distribution of aptamers within the clusters. Interestingly, we found that the distribution of these frequencies was very skewed (Fig. 4.6). Except for a handful of highly abundant aptamers, most of the species in a cluster had low frequencies. Such extreme differences in frequencies is consistent with a situation in which most of the cluster diversity can be attributed to mutations caused by polymerase errors. To test this hypothesis, we investigated whether aptamers with a maximal count of 5 from the top 20 clusters in cycle 5 were also present in the sequenced portion of the selection pool from cycle 2. Indeed, the vast majority of these sequences (99% of singletons, 97% for frequency 5) where absent in this pool (Table 4.1). Note that the sequences introduced by Polymerase errors can be subsequently selected and amplified providing an important source of cluster's diversity. However, due to the late introduction, their frequency count might not correctly reflect their binding affinity.

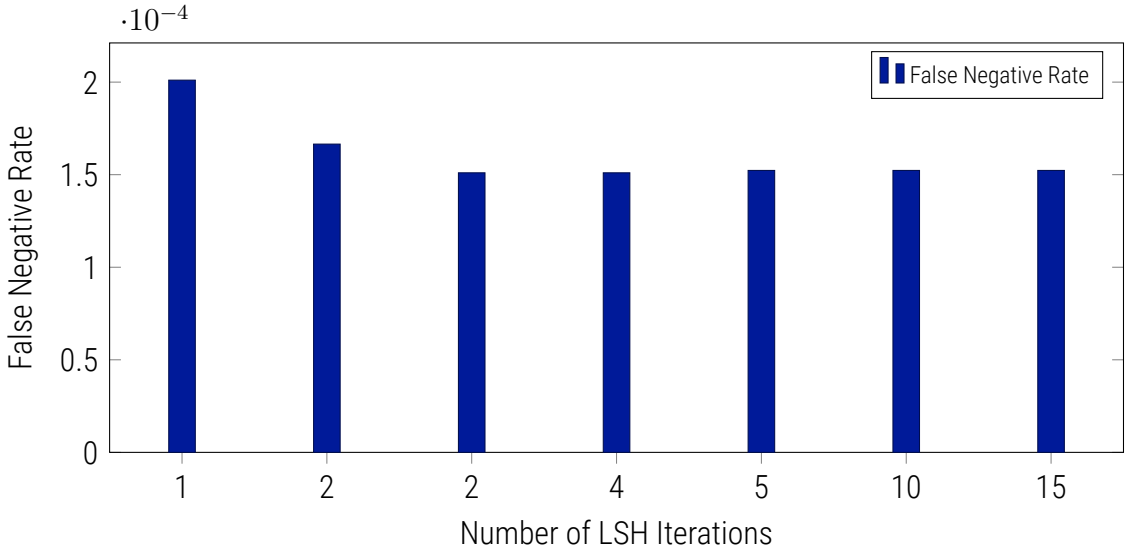


Figure 4.5: The false negative rates for the 20 largest clusters for different numbers of locality sensitive hashing iterations in selection cycle 5. The graph shows that LSH is stable for even a small number of iterations.

Frequency Counts Versus Binding Affinity

It is often assumed that an aptamer sequence's frequency in the pool later cycles provides a good predictor of its binding affinity. Indeed this would be a reasonable expectation under the assumption that the selection process is free of any artifacts, all aptamers are present in the initial pool with the same frequency, and there was no stochastic variability during the aliquotation process. However the realization that a large fraction of sequences in the final pool might have been absent from the initial pool but introduced in a later stage made us to reexamine this assumption. As can be appreciated in Table 4.2 b, we measured the dissociation constant K_d for 30 aptamers including the most frequent ones. We found that cycle-to-cycle enrichment of aptamer frequencies, i.e. their relative increase in multiplicity, from cycle 4 to 5 is a better predictor of binding than the frequency in the final pool. Specifically, taking 125 K_d as a reasonable threshold between binders and non-binders, sorting by cycle-to-cycle enrichment separates binders form non-binders while sorting by frequency leaves these two groups randomly mixed (Table 4.2 a).

In addition to the emergence of new sequences, another source of dissonance between aptamer frequency and its binding potential could also be the differences in their frequencies

Table 4.1: Number of species with counts 1 to 5 present in the top 20 clusters of selection round 5 compared to the frequency of their occurrence in selection round 2. The overwhelming majority of the sequences are not present in the latter.

	Nr. of aptamers with frequency 1-5				
	1	2	3	4	5
Top 20, cycle 5	8529	2202	1074	614	465
Found in cycle 2	61	36	27	18	16

Table 4.2: Cycle-to-cycle enrichment as a superior predictor for binding affinity and mutant analysis results. Top 20 clusters reported by AptaCLUSTER sorted by their enrichment from cycle 4 to 5. **(a)** Shown are the cluster id (ID), consensus sequence, cluster size in round 5 (Cluster Size), the percentage of a cluster with respect the remaining pool (Fraction), the cluster diversity as a function of the number of unique species in the cluster (Unique), cycle-to-cycle enrichment (Enrichment), and the KD values of the most frequent aptamer. Cycle-to-cycle enrichment successfully partitions the clusters into binders and non-binders. The black line at 125nM indicates the threshold used to visually separate strong binders from weak binders. In contrast **(b)** depicts the same aptamer families if sorted by cluster size only which cannot discriminate between target-affine binders and non-binders.

(a)

Row	Consensus Sequence	Cluster Size	Fraction	Unique	Enrichment	KD
6	ACTATAACGCGTCAAAGTGCTTATCGAACACTATTTGTAA	27842	0.006025	407	2318.03	50
2	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	401977	0.086981	2064	29.4554	65
30	ACAGACCAAGTGTTCAGAAAAACAGTTGCTCAATATACAT	2691	0.000582	102	4.29201	25
0	CCCCCGCATCACGCCGTGGTGCGATTGACACAATTGCAAT	2191739	0.474255	4883	4.02953	25
1	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	788907	0.170706	3199	3.99062	120
33	AAAGACCGTTTTTTAAAACGCTCAATATACACGACATAAA	2152	0.000466	95	3.96389	10
40	CCGCTAACACTCGATTCTGCGGAAATGCCCCCCTGAACCC	1695	0.000367	98	3.71368	120
4	AGCCATGACGATGTCGTTACGTAGATGCAGAGACTCCTAA	33473	0.007243	488	1.5169	18
8	AGCAAAGTCGTCAGCAATATAACACTCGATTCTTATTGGA	24982	0.005406	360	1.46659	80
12	GACTCAGCAGCGCGAGAAAGACGTGGTTAGCCTCAATATG	13018	0.002817	254	0.971524	20
3	AATCGCTCAGCCGGTCCGGAACTGGCAAAGTCAGGTGCTC	71213	0.015409	916	0.691825	60
18	TGTCGAGTACTTTTCTCAGTCTATTCCACAGTAGTGGAGA	6622	0.001433	273	0.448961	80
22	TTCTCAGTCAATAACTAGTCGGATCGTGGTCCAAGAGCAG	4547	0.000984	186	0.405318	20
7	CCCCTTCCAGCGATTACGATCATTGACTCTCAGTCCTGTG	25356	0.005487	427	0.391821	250
15	TGCAATGAGGACTTCTCTCAGTCTAACACAATGTTGTTTA	7519	0.001627	230	0.342604	500
14	ATCGACAGCTCTCAGTCCATTCGAGGAATGTTCATCGATA	9461	0.002047	245	0.300622	125
5	TGAGAACTTCTCTCAGTCGGTGGGAGAGTACATCCTAACA	33011	0.007143	519	0.25693	500
16	ACGATCACTTCTCTCAGTCGGACTAATATTACGGTTAGAA	7031	0.001521	223	0.21148	250
20	TATAGAGAACTTCTCTCAGTCGAAGCCAAAGAGCATTAAT	5901	0.001277	174	0.202396	250
9	TAGACGAGCACTTCTCTCAGTCGCATTCATTATTTAAATT	16788	0.003633	342	0.18987	250
24	TGCCCTGCTTCCCCAGTCTTGTCTCACAAGACTAATTTTA	4055	0.000877	152	0.153568	500
26	ATGAGGACTTCTCTCAGTCGTGACCATTAAATAAGAGAAA	3009	0.000651	163	0.13792	500

(b)

ID	KD
0	25
1	120
2	65
3	60
4	18
5	500
6	50
7	250
8	80
9	250
12	20
14	125
15	500
16	250
18	80
20	250
22	20
24	500
26	500
30	25
33	10
40	120

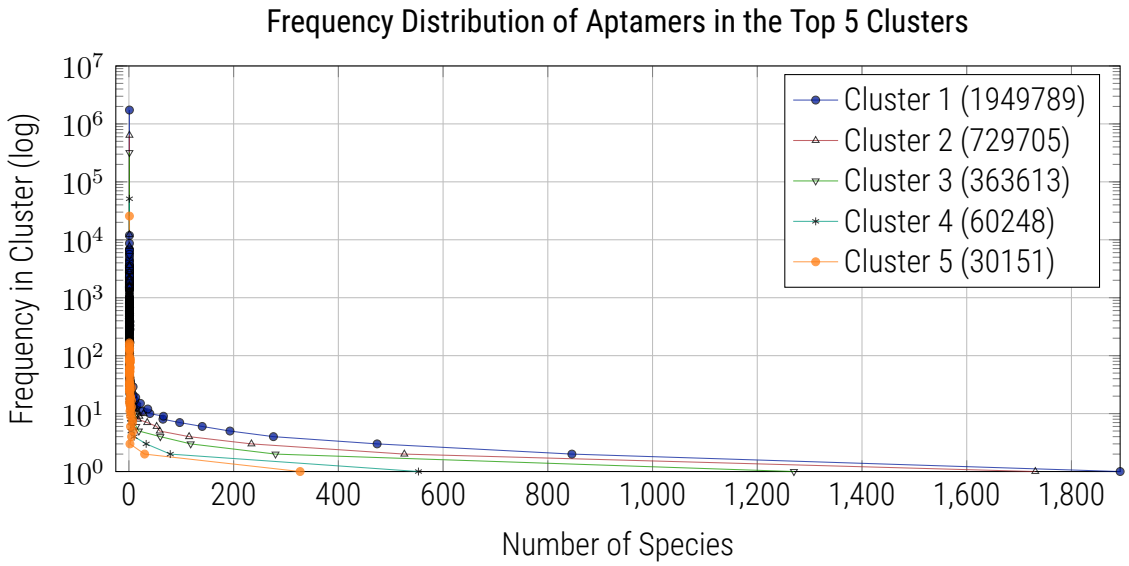


Figure 4.6: The frequency distribution of the members of the 5 largest clusters from the IL10 selection. The cluster sizes are given in the brackets.

in the initial pools due to the stochastic nature of partitioning the pool into groups to be used for storage/sequencing/next cycle. Looking at cycle-to-cycle sequence enrichment instead of counts permits a resolution of this problem. However, other artifacts exist that can affect aptamer frequencies as well. In particular, we also tested the K_d values for non IL-10RA specific binding using binding to IgG as proxy for such non-specificity (data not shown). We found for example that cluster with ID 3 has high frequency in cycle 5 but it is not IL-10RA specific.

Runtime Analysis and Performance Comparison

Finally, to appreciate the advantages of our approach, we have compared the performance of AptaCLUSTER to a member of the general class of clustering algorithms ($\Theta(N^2)$ computational time). We implemented a sample algorithm in this class which considers aptamers in decreasing order of their counts, computes their distance to all other aptamers in the pool, and assigns the aptamer to the seeds cluster if the distance is below a user defined threshold. While our approach can handle over 100 million sequences in a little more than one hour, this naive approach is unable to handle 1 million items within one day (Figure 4.7) on identical hardware.

4.3.5 Conclusions and Discussion

Given the great promise of the HT-SELEX approach and rapidly diminishing costs of next generation sequencing, the usage of this method is likely to increase rapidly. Therefore it is imperative that researchers are able to analyze and correctly interpret HT-SELEX results. We have developed a new approach, AptaCLUSTER that allows for clustering based on primary

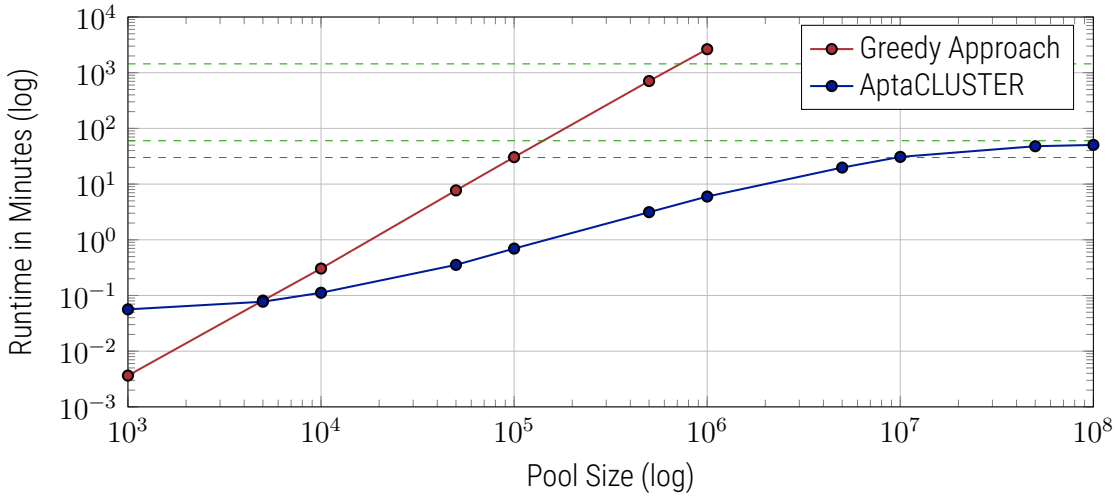


Figure 4.7: Runtime (wall clock) analysis of AptaCLUSTER (blue) and a greedy approach (red) for pool sizes varying between 1000 and 100 million (Mio) sequences. Horizontal green lines depict runtimes of 30 minutes, 60 minutes, and 24 hours respectively.

structure of pools of aptamers sequenced using Hi-Seq technology.

Until now, a typical HT-SELEX analysis was reduced to counting the frequency of each aptamer and using such counts as a predictor of binding affinity. However our results indicate that such counting is actually not as good of a predictor as it has been anticipated. Instead, a predictor that utilizes the dynamics of the cycle-to-cycle enrichment holds greater promise.

Our results of applying AptaCLUSTER to the outcome of the IL-10RA HT-SELEX experiment revealed important properties of the resulting clusters. We found the clusters to be well separated, and typically dominated by one or a few individuals. Relaxing the parameters to allow for larger intra-cluster distances did not change the results significantly. Consequently, sequence profiles of individual clusters were dominated by one or a few of the most abundant sequences. We have also implemented a procedure that enables the tracing of the clusters over consecutive selection cycles and, consistently with the observation above, we found that the clusters' sequence profile did not change much during consecutive selection steps.

The distribution of frequency counts within clusters suggests that cluster diversity is, in a large part, a result of Polymerase errors. The emergence of such Polymerase mutants creates an interesting opportunity to sample around local minima. This is strengthened by the observation that the number of mutations correlates with the frequency of the cluster seeds: the more frequent the seed, the more frequent the mutants. How to design the dynamics of the selection process to optimally utilize these emerging mutants is an open question. One possibility is to replace the typical selection procedure where selection pressure increases in each cycle by an approach that alternates between stronger and weaker selection.

AptaCLUSTER provides a valuable tool which will help us and others to analyze and to

optimize the HT-SELEX procedure. It has enabled us to analyze the results of HT-SELEX for IL-10 and allowed for a better understanding of the HT-SELEX experiment. We expect that the properties of the clusters obtained with AptaCLUSTER will vary depending on the experimental details of HT-SELEX protocol in use, the length of the variable region, error rate of Polymerase, and properties of the target. Independently of this expected variability, AptaCLUSTER can be used as the first step towards understanding the aptamer binding landscape, and for the identification of a broad spectrum of potential binders. We point out that AptaCLUSTER is not intended to elucidate complex, indel-containing motifs but rather to operate on sequences of equal length. It is designed to serve as a preprocessing step for approaches to uncover sequence-structure motifs such as the planned extension of our AptaMotif algorithm to high throughput sequencing data [8].

4.4 AptaMUT: Leveraging the Mutational Landscape in SELEX

Section 4.3 motivated the need for efficient clustering solutions for HT-SELEX data and provided insight into possible scenarios in which this information proves beneficial. Specifically, it was revealed that the vast majority of clusters correspond to aptamer families emerging through mutagenesis-driven amplification and, for RNA aptamers, transcription during SELEX. Due to the stochastic nature of the mutations introduced during PCR, in combination with the selection pressures exerted onto the species during incubation, the probability that some of these mutated species could exhibit a higher binding affinity and/or binding specificity to the target becomes likely enough to warrant further computational exploration.

Importantly, the principles of mutagenesis during traditional SELEX [80, 81, 82], and as a means of post-selection optimization of binding affinity [83], have previously been described. However, the lack of high-throughput sequencing of entire aptamer pools posed a natural limit to the resolution of the available data and consequent analysis. Particularly, understanding and quantifying the effect of these mutations on the binding affinity between the parent sequence and the mutants is of great interest as this information could be utilized as the basis for an *in silico* survey aimed at the identification of higher quality aptamers. The underlying postulate leverages the idea that any beneficial mutant introduced into the pool would be preferentially selected for and therefore, over time, become enriched at a higher rate as its parent sequence. In contrast, aptamers carrying non-beneficial mutations are expected to be competed out of the pool throughout the course of the selection.

The development of appropriate models however is confronted with a number of challenges emerging as a consequence of current technology limitations and the design of the HT-SELEX protocol itself. It has previously been established for non-mutant aptamers that cycle-to-cycle enrichment, i.e. the ratio of the aptamer's count normalized by the corresponding pool size between two consecutive cycles, represents a superior predictor of binding affinity as compared to the raw count of that species in a particular pool (see Section 4.3 for details). The accurate determination of enrichment values for these candidates is mainly enabled by deep sequencing each round of selection and therefore providing large enough copy numbers such that small variations do not significantly affect the enrichment computation. These variations can be caused by a variety of reasons. First, technological artifacts such as sequencing errors or low quality reads might affect the final count of a species. Next, these discrepancies can arise during the aliquotation of the PCR product during the SELEX protocol itself (see Figure 2.2 in Chapter 2 for a reminder). Here, aptamers already present in large quantities are not expected to be significantly affected by the sampling process and therefore retain their initial proportions in the aliquotes. In contrast, low frequency species are subject to much greater concentration fluctuations due to the mechanical nature of separating the samples. Finally, the significant difference in the sequenced pool sizes ($10^5 - 10^7$ reads) and the true pool sizes ($10^{13} - 10^{16}$ molecules) need to be acknowledged as an additional factor for the variation in observed aptamer cardinality.

It is therefore not surprising that these variations become increasingly pronounced when attempting to measure the enrichment of lower frequency species. This is especially true for

newly introduced mutants whose initial count is typically hundreds to thousands of orders of magnitude that of their parent sequence, leading to significantly different observed enrichment values when compared to their expected behavior. Hence, any computational model designed to leverage the enrichment rates of mutants as a predictor for binding affinity must take the above described artifacts into account.

With these challenges in mind, we specifically asked: (i) is the distribution of mutants consistent with the random mutation model, and (ii) is it possible to computationally identify mutations that improve binding affinity. Our study was informed by high-throughput sequencing data from five rounds of HT-SELEX developing aptamers against the Interleukin 10 receptor alpha chain (IL-10RA). IL-10 is considered to be a master regulator of immunity to infection and is an important therapeutic molecular target [77].

To address the first question, we utilized AptaCLUSTER to extract all aptamer families from the selection cycles, and we derived a mathematical estimator of the expected number of aptamers that originated from the initial pool as opposed to those arising by a mutation and that are above a specific similarity threshold with respect to an aptamer of interest. We used this estimator together with our AptaCLUSTER method to obtain families of aptamer sequences related to each other by mutations. Interestingly, we found that similar to a number of phenomena in life and social sciences, the distribution of aptamers in these families follows a scale-free distribution [84]. We obtained the same distribution using our *in-silico* aptamer evolution program AptaSIM and we discuss the practical implications of these findings for predicting binding affinity.

To address the second question, we developed AptaMUT - a method to identify mutations that improve or reduce binding affinity. AptaMUT leverages the cycle-to-cycle enrichment of all sequenced mutants in an aptamer family between two selection cycles and ranks these according to their likelihood of improved binding affinity compared to their parent sequence. Our method directly models the process of aliquotation embedded into the SELEX protocol as well as additional technological causes for noise in an attempt to detect and discriminate against these. Using this approach on our IL-10 data set, we successfully identified several affinity-improving mutations and we have confirmed these predictions experimentally. Furthermore, we discovered that in one particular cluster, mutations conferring the biggest change in the binding affinity stabilized a specific hairpin.

Our results demonstrate that new computational methods cannot only aid the elucidation of under-appreciated properties of the SELEX procedure but can ultimately lead to uncovering new practical predictive methods and aptamers of desired binding affinity. Taken together, we demonstrated that HT-SELEX data sets contain previously untapped information and provided methods for their utilization. We expect that these new methods along with our complete software suite will become indispensable for guiding aptamer selection and for uncovering additional general properties of the selection process, hence jointly contributing to a better utilization of HT-SELEX results.

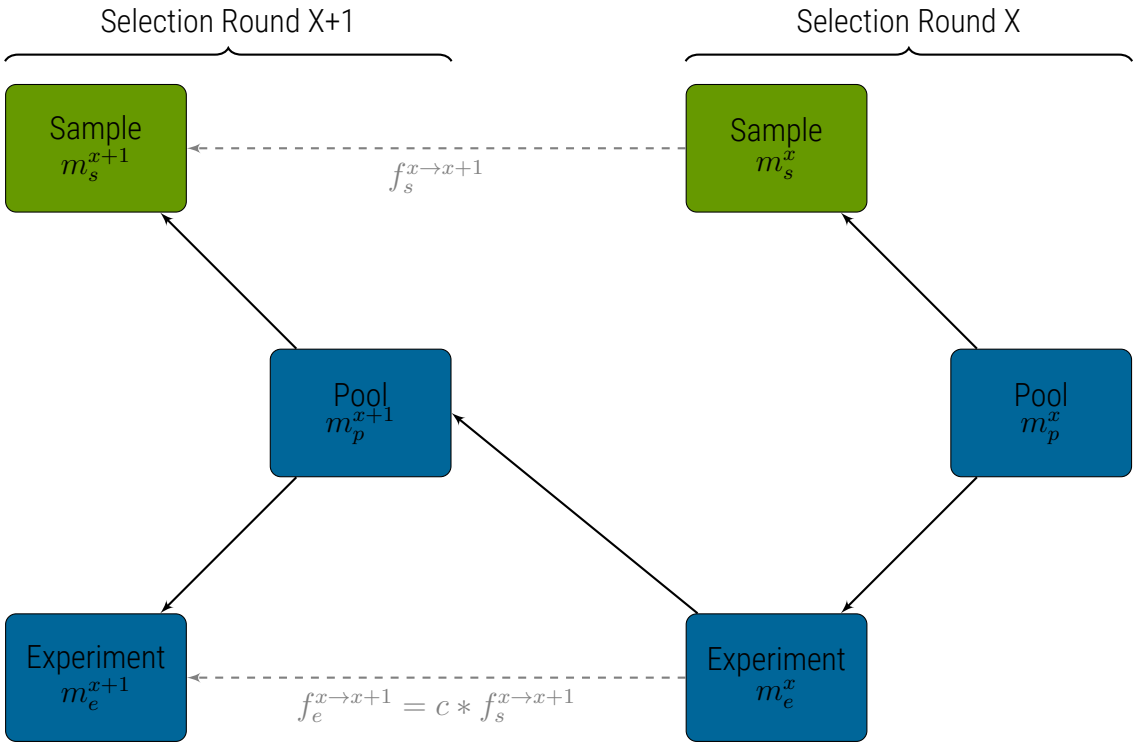


Figure 4.8: Visualization of the model used to estimate the significance of enrichment between the selection rounds. Here, only the sample sets (green) are observable quantities whereas the pool and experiment sets are hidden. Each selection round is partitioned into three sets denoted as pool, representing the remaining sequences after selection and amplification, sample, describing the established, sequenced portion of this pool and experiment, denoting the unknown species forming the input for the next cycle. m_s^x, m_e^x and $m_p^x = m_s^x = m_e^x$, stand for the frequency of a sequence in the sets sample, experiment and pool, respectively. The enrichment of the sequence between selection cycles is defined as $f_s^{x\to x+1}$ for the sample sets and as $f_e^{x\to x+1}$ for the experiment sets where c denotes the ratio of partitioning the pool into the sets sample and experiment.

4.4.1 Algorithmic Description of the Approach

AptaMUT aims at extracting favorable mutants by leveraging the fact that at each cycle, the sequenced aptamers represent a fraction of the true pool size (Figure 4.8). Starting with the null model that the binding affinity of the mutant sequence is the same as the parent's sequence (and thus the expected cycle-to-cycle enrichment the mutant is identical to that of the parent), we compute the probability that observed frequencies in the sequenced portion of the pool are higher than expected from this null model. This is archived in three main steps: sequence extraction that identifies favorable and unfavorable mutants in each cluster, scoring, in which we compute a p-value reflecting the statistical significance of a mutant with respect to their increase in binding affinity as compared to the seed sequence, and multiple hypothesis correction. In the following, we describe the details of each of these steps.

Sequence Extraction: We define a mutant as a sequence present in a particular cycle X and the next cycle $X + 1$, but that has never been encountered in any of the previous rounds. Consequently, a favorable mutant can be described as a mutant with significantly higher en-

richment as compared to its parent sequence. Here, enrichment refers to the ratio between the mutant's frequency in cycles $X+1$ and X normalized by their respective pool sizes. We therefore used the clustering results of the four largest aptamer families of selection cycle 5 and extracted potential favorable mutants.

Scoring: In order to compute a score for each mutant reflecting the significance of the fold-change in enrichment between cycles X and $X+1$, we developed a generative model mirroring the experimental design of the HT-SELEX protocol. The model is based on the notion that the sequenced aptamers at each cycle only represent a fraction of the true pool size and that the process of selecting these sequences from the pool can be described in terms of a Bernoulli experiment. In addition we assume that the enrichment and the amplification processes are subject to noise modeled by a normal distribution. The model is parameterized by the expected sequence enrichment so that different sequence enrichments correspond to different models. Given the model built using the enrichment equal to the enrichment of the seed, we compute the probability that the mutant's counts in X and $X+1$ could have been generated by this model. This probability is then normalized by the probability of the optimal counts, as described below.

We divide each selection round into three distinct sets denoted as *pool*, representing the remaining sequences after selection and amplification, *sample*, describing the established, sequenced portion of this pool, and *experiment*, standing for the unknown species forming the input for the next cycle (see Figure 4.8). Furthermore, let m_s^x, m_e^x, and $m_p^x = m_s^x + m_e^x$, be the frequency of a sequence in the sets *sample*, *experiment*, and *pool* respectively. We define the enrichment of a sequence between selection cycles X and $X+1$ as $f_s^{x \to x+1}$ for the sample sets, and as $f_e^{x \to x+1}$ for the experiment sets. Similarly, we define the enrichment of the parent of the sequence between selection cycles as $\hat{f}_s^{x \to x+1}$ for the sample sets and as $\hat{f}_e^{x \to x+1}$ for the experiment sets. Finally, for a mutant that is neutral w.r.t. its parent sequence, its expected frequency in the pool $X+1$ can be described as

$$m_p^{x+1} \approx c * \hat{f}_s^{x \to x+1} * \underbrace{\left(m_p^x - m_s^x\right)}_{=m_e^x} \tag{4.7}$$

for any unknown count m_p^x in pool X. Here, we use the constant c to model both, the amplification stage (PCR) after each selection round and for normalization purposes.

Our model then aims at comparing the probability of observing frequencies m_s^x, m_s^{x+1} in sample sets $X, X+1$ of a mutant with the probability of observing the expected frequencies m_s^x, $m_s^x * \hat{f}_s^{x \to x+1}$. Let $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ refer to the probability of simultaneously observing m_s^x in sample set X and m_s^{x+1} in sample set $X+1$ by chance given the expected abundance of the mutant in the experiment sets can be described as a function of the enrichment of the parent sequence between the sample sets. Similarly, $P(m_s^x, \hat{f}_s^{x \to x+1} * m_s^x, \hat{f}_s^{x \to x+1})$ refers to the probability of the mutant being neutral, i.e. observing m_s^x and $\hat{f}_s^{x \to x+1} * m_s^x$ in the sample sets of X and $X+1$ respectively and under the assumption that their actual enrichment is identical to the seed's $\hat{f}_s^{x \to x+1}$. Then, we aim to compare $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ and $P(m_s^x, \hat{f}_s^{x \to x+1} * m_s^x, \hat{f}_s^{x \to x+1})$. We therefore define a significance score $S(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ for a mutant as the probability of the mutant's observed enrichment being higher than its parent, normal-

ized by the probability of the mutant being neutral i.e. exhibiting an enrichment rate equal to its parent sequence:

$$S(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1}) = \frac{P\left(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1}\right)}{P\left(m_s^x, \hat{f}_s^{x\to x+1} * m_s^x, \hat{f}_s^{x\to x+1}\right)} \tag{4.8}$$

In what follows, we show how to compute $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1})$ and $P(m_s^x, \hat{f}_s^{x\to x+1} * m_s^x, \hat{f}_s^{x\to x+1})$. The observations m_s^x and m_s^{x+1} in the sample sets can be interpreted as the result of partitioning pools X and $X+1$ into *sample* and *experiment* sets and hence as random variables following binomial distributions, in which m_s^x and m_s^{x+1} correspond to a known number of successes out of m_p^x and m_p^{x+1} unknown trails respectively. For any frequency of a mutant m_p in each pool, the probability of observing exactly m_s mutants in the sample set is then given by the probability mass function (*pmf*)

$$f_B(m_s, m_p, p) = Pr(X = m_s) = \binom{m_p}{m_s} p^{m_s}(1-p)^{m_p - m_s} \tag{4.9}$$

of the Binomial distribution $B(m_p, p)$ and the probability of simultaneously observing both frequencies in the sampled pools corresponds to the product of their respective *pmfs*. Since the original number of mutants in pool X is an unknown quantity, we have to consider all possible pool sizes in order to estimate $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1})$:

$$P\left(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1}\right) = \sum_{m_p^x = m_s^x}^{\infty} f_B\left(m_p^x, m_s^x, p\right) * f_B\left(m_p^{x+1}, m_s^{x+1}, p\right) \tag{4.10}$$

So far, our model is only concerned with possible variations in the number of mutant sequences in the pool and does not take into account any biases that might affect the enrichment value of the seed sequence. These noises, such as artifacts during PCR and sequencing errors, might lead to an overestimation or underestimation of the true enrichment value. We therefore extend our approach with a continuous random variable f to model the observed seed enrichment $\hat{f}_s^{x\to x+1}$ in the sequenced portion of the pool. More specifically, we assume that f follows a normal distribution $\mathcal{N}(\hat{f}_s^{x\to x+1}, \hat{f}_s^{x+1}/3)$ with mean $\hat{f}_s^{x\to x+1}$ and standard deviation $\hat{f}_s^{x\to x+1}/3$. We then express the probability of observing frequencies of a mutant in sample sets X, $X+1$ and the probability of observing its expected frequencies as functions of f. It follows that the new significance score of a mutant, denoted as $\hat{S}$, corresponds to ratio of the expected values of these functions:

$$\hat{S}(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1}) = \frac{\int_0^\infty P\left(m_s^x, m_s^{x+1}, f\right) p(f) df}{\int_0^\infty P\left(m_s^x, \hat{f}_s^{x\to x+1} * m_s^x, f\right) p(f) df} \tag{4.11}$$

Here, $p(f)$ is the probability density function of the normal distribution $\mathcal{N}(\hat{f}_s^{x\to x+1}), \hat{f}_s^{x\to x+1}/3)$. Finally, we approximate each integral within three standard deviations from the mean by discretizing $p(f)$ into equidistant intervals of length d denoted as $k = \lfloor 2\hat{f}_s^{x\to x+1}/d \rfloor$-1. Below, we show how to approximate the integral on the example of the numerator and note that the

denominator is approximated in a similar manner.

$$\int_0^{\infty} P\left(m_s^x, m_s^{x+1}, f\right) p(f) df \approx \int_0^{2\hat{f}_s^{x \to x+1}} P\left(m_s^x, m_s^{x+1}, f\right) p(f) df$$
$$\approx \sum_{i=1}^{k} P\left(m_s^x, m_s^{x+1}, f\right) * P\left(i * d \leq f < (i+1) * d\right) \tag{4.12}$$

Setting $p = 0.5$, therefore assuming that each mutant has equal chance of being selected for sequencing, allows for the computation of the significance score $\hat{S}$ for all favorable mutants identified during the sequence extraction step. Analogous to p, we set $c = 2$, hence assuming that after selection, each pool is amplified back to its original size. For the discretization step, we found that a value of $d = 0.5$ as the width of intervals yielded the desired accuracy for our purposes.

Multiple Hypothesis Correction: In our model, each computed p-value $p_1 \dots p_x$, where x is the number of identified mutants, corresponds to a different hypothesis $H_1 \dots H_x$ making it necessary to correct for multiple hypothesis testing. We therefore apply a simple Bonferroni correction to control the p-value's family-wise error rate (FWER) at level α:

$$\tilde{p_i} = x * p_i < \alpha$$

4.4.2 Results

Our results combine the development of dedicated computational methods and insights into the HT-SELEX process gained using these methods. On the methodological side we developed AptaMUT and a mathematical estimator of the expected number of aptamers with at least K% similarity in a given pool. Alongside we report the results of the effort to identify aptamers targeting IL-10RA that was supported by our computational methods, as well as general insights into the selection process obtained with these analyses (see Appendix B.2 for selection details).

Identification and analysis of families of sequences related to each other by mutagenesis

The errors during the amplification step of the SELEX procedure can introduce new sequences into the selection pool. Importantly, the sequences that are selected for and thus appear in higher copy numbers in the selection pools are most likely to produce mutants. Since the randomized region is typically relatively long, the coverage of a randomized pool of $10^7 - 10^{15}$ initial sequences is sparse and such mutants might help to provide additional sampling of the sequence space around the sequences that are selected for. To better understand the initial sequence diversity, we started by estimating the expected number sequences with at least K% sequence identity in an initial pool of M random molecules and with a randomized

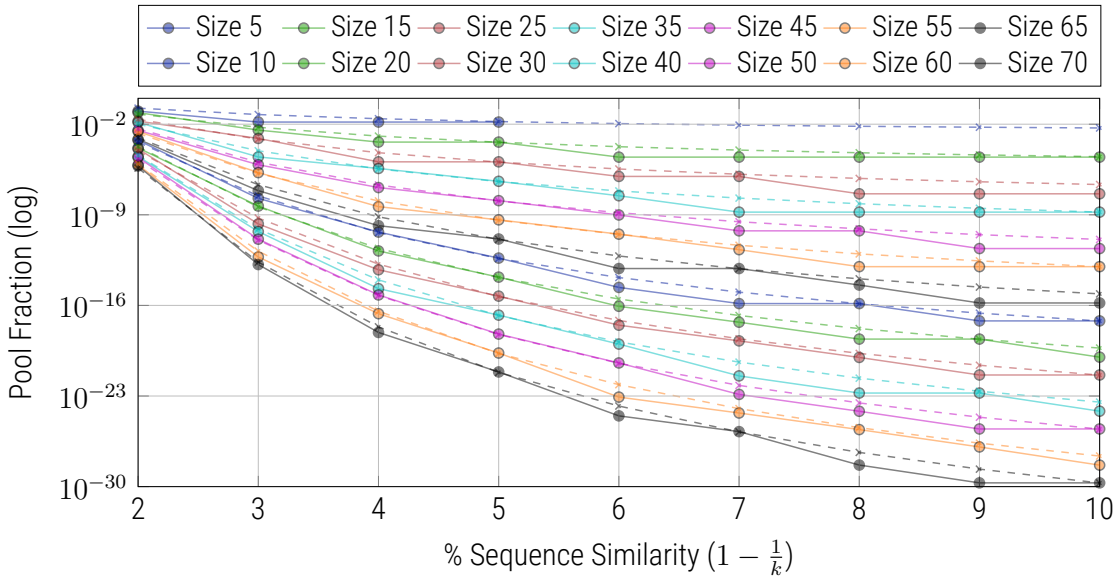


Figure 4.9: Comparison of the predicted pool fraction of sequences with an expected sequence similarity K between our estimator (dashed lines) and the exact formula (continuous lines). Our estimator provides a reasonable upper bound for the expected fraction of sequences in an initial pool with at most K% sequence similarity.

region of length n. This number is given by

$$F(n,k) = \sum_{i=1}^{\frac{n}{k}} \frac{\binom{n}{i} 3^i}{4^n}$$

which, as formally shown in Appendix Section B.3, decreases exponentially with n and where $\frac{1}{k}$ is the percent divergence, that is $K = 1 - \frac{1}{k}$ (See Figure 4.9).

Based on this formula, we set the clustering parameters for AptaCLUSTER such that, with high probability, all cluster members are obtained as a result of polymerase errors from a common "founder sequence" - the seed of the cluster. We confirmed that these putative mutants were indeed absent from early pools (see Table 4.1, Section 4.3). In addition, tracing the clusters over several selection rounds, we observed that not only the number of clusters decreases while the average number of the sequences per cluster increases, but also the variability of the sequences within the clusters increases, consistent with mutagenesis based cluster evolution (Figure 4.10). Importantly, similar to many evolutionary processes [84], the aptamer counts in each cluster followed approximately a scale-free distribution (Figure 4.11). To confirm that this distribution is indeed expected, we utilized AptaSIM and trained its first order Markov model with the earliest sequenced IL10 round in order to accurately represent the initial randomized pool in terms of base composition and dependencies within consecutive nucleotides, both possible technology-dependent artifacts. The results obtained with AptaSIM confirmed that the distribution of sequences related to a given "seed sequence" by mutations is consistent with distribution within the sequence families obtained with AptaCLUSTER (Figure 4.11).

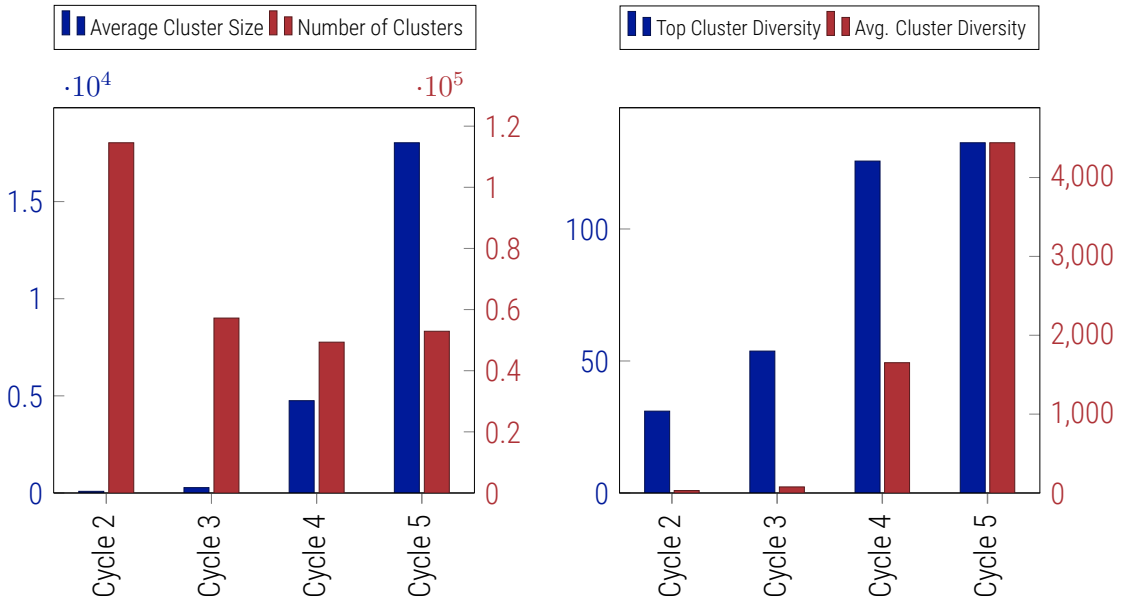


Figure 4.10: Changes in cluster size and diversity throughout the selection cycles. The labels of the y-axes correspond to the legend names matching the color of their axis ticks. (a) Average cluster size (blue) and the number of clusters (red) reported by AptaCLUSTER for each of the sequenced pools. (b) Average (blue) and top (red) cluster diversity for each selection round as measured by the number of unique sequences per cluster.

Combining cycle-to-cycle enrichment with a probabilistic model to identify mutations that influence binding

We note that some of the mutated sequences might be better binders than the sequence they are derived from. Such sequences would not only provide additional candidates for further *in-vitro* testing and refinement, but also reveal crucial information about the relevance of the nucleotide position as a function of binding affinity. However, due to the late introduction to the pool, their count is low excluding the use of an aptamer sequence's frequency as a predictor of their binding affinity. Since aptamers with advantageous binding properties are expected to be selected in consecutive SELEX rounds at a higher rate as compared to less affine species, one can therefore use cycle-to-cycle enrichment to predict the relative ordering of aptamers with respect to binding strength [85, 68, 69]. In Chapter 4.3, we tested the utility of cycle-to-cycle enrichment of aptamer frequencies, i.e. their relative increase in multiplicity as a predictor of binding affinity and found that it to be a better predictor of binding affinity than the simple aptamer count (see Table 4.2, Section 4.3.4).

Having confirmed the utility of cycle-to-cycle enrichment as a predictor of binding affinity, it might be tempting to apply this strategy to mutants as well. However the implicit assumption of the cycle-to-cycle enrichment strategy is that the fraction of the pool that is used for sequencing is a good representative of the fraction that is used as the input for the next cycle. This is a reasonable assumption for abundant sequences but can be incorrect for the less frequent mutants whose count in the sequenced pool is strongly affected by stochastic variations during pool partitioning and PCR amplification. Therefore, we developed an approach that directly models the fact that at each cycle, the sequenced aptamers represent

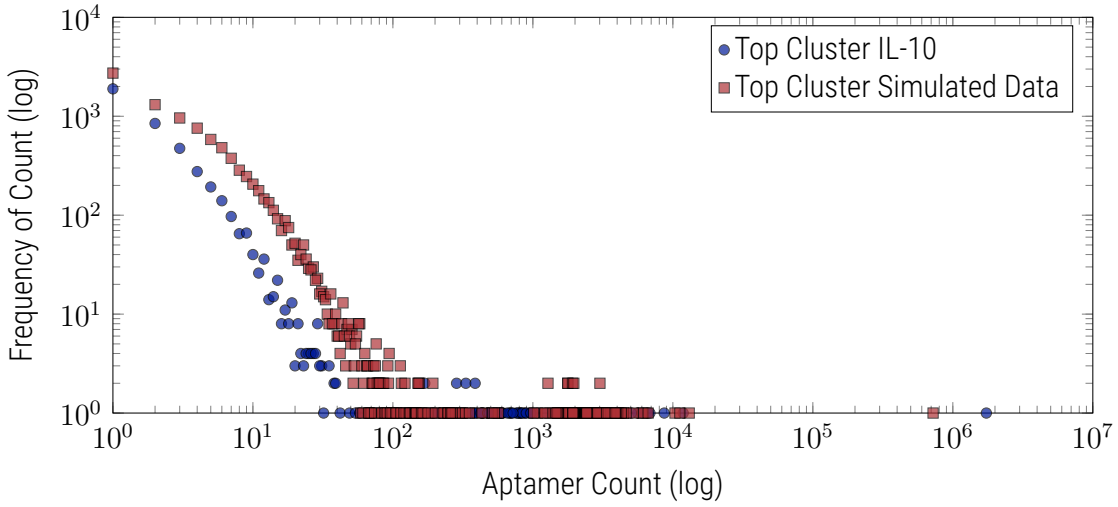


Figure 4.11: Scale free nature of the cluster composition. Shown are the distribution of aptamer frequencies as a function of their counts on the example of the top clusters from the IL-10 selection (blue) and the simulated data (red) produced by AptaSIM.

a fraction of the true pool size. Specifically, assuming that the partition into sequenced and experiment pool follows a Bernoulli process, we can compute the probability of observing a given number of sequence copies in the sequenced pools under the assumption of a particular enrichment value. After appropriate normalization, we obtain a score reflecting the likelihood of observing the counts of a mutant in consecutive cycles relative to the expected counts under the assumption of having the same enrichment as the parent sequence. In this model, a log of this score near zero indicates a neutral mutant while significantly positive (respectively negative) log scores indicate a possibility of beneficial (respectively detrimental) mutants. Note that all scores are computed relative to the parent sequence, and it is possible that a sequence with a detrimental mutation shows cycle-to-cycle enrichment. The mathematical details of the test are provided in Section 4.4.1.

Experimental and Computational Validation

We used AptaMUT to identify favorable mutants from three representative clusters identified by AptaCLUSTER of selection cycle 5 and whose binding affinities of their seed sequences had been determined to represent strong ($K_d = 27nM$), intermediate ($K_d = 65nM$) and weak target ($K_d = 120nM$) binding. We scored these mutants by their significance of enrichment, and experimentally tested a total of 8 candidates for their binding affinity. All but one mutant showed either comparable Kd values or an increased binding affinity with respect to their parent sequence (up to 3 fold when starting with a seed with intermediate K_d value). Interestingly, mutants from the strongest target-binding seed did not show improvement whereas significant better binders could be found in the two remaining categories (Table 4.4) suggesting that the sequence with intermediate Kd value was easiest to improve upon.

We noted that cluster 1 contained several mutants that were experimentally confirmed to improve binding affinity. Therefore we asked if these mutants collectively provide insight into the mechanism behind this improvement and if the analysis of mutants allows for the iden-

Table 4.4: Selection of mutants belonging to three clusters of interest reported by AptaMUT and tested for binding affinity (K_D). The last column displays the log score of the mutants' enrichment with respect to the seed sequence (grayed rows). All but one mutant show higher binding affinities compared to their parent sequence.

Aptamer Sequence	% Pool 5	Enrichment	% Pool 4	KD	P-Value
TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	0.14865	3.95120	0.03762	120	
TCACAGTCCCGGTGCCGCCCTAAAACCCATTGTTGTGCGA	0.00014	15.33306	9.36E-06	98	1.40E-09
TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCTA	0.00013	12.52811	1.09E-05	143	7.91E-08
TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGT	9.61E-05	8.03605	1.20E-05	78	1.23E-03
TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	0.07614	30.31212	0.00251	65	
TAACACTCGATTCTCCTAGCCCTCTAGAAATTCCCCTCCC	0.00067	429.74213	1.56E-06	46	1.34E-10
TAACACTCGATTCTCCTATCCCGCTAGAAATTCCCCTCCC	0.00030	45.72701	6.76E-06	27	3.00E-2
TAACACTCGATTCCCCTAGCCCGCTAGAAATTCCCCTCCC	0.00020	32.71285	6.24E-06	33	8.00E-2
AGCCATGACGATGTCGTTACGTAGATGCAGAGACTCCTAA	0.00629	1.59079	0.00395	18	
AGCCATTACGATGTCGTTACGTAGATGCAGAGACTCCTAA	2.49E-05	11.96811	2.08E-06	19.8	5.79E-06
AGCCATGACGATGTCGTTACGTAGATGTAGAGACTCCTAA	1.67E-05	8.01343	2.08E-06	21.2	7.9E-4

tification of the binding motifs approximate location. To see if this might be the case, we predicted the secondary structures of the seed and potentially beneficial mutants, selected based on having a log-score of less than -0.5 (30 in total), as well as the secondary structures of mutants with the highest depletion rate at the same cutoff (top degenerative mutants, 10 in total). We identified a hairpin loop that showed significantly less mutations in the set of beneficial mutants as compared to the set of degenerative mutants (p-value$= 0.025$, Fisher exact test) suggesting its importance for binding (Figure 4.13). Interestingly some of the predicted beneficial mutations, including one experimentally confirmed to improve binding affinity (Mutant 2), were found to induce a conformational change in the structure while still exposing the conserved loop region. Coincidentally, the mutant with the highest change in affinity (Mutant $1 - K_d = 27nM$) was also predicted to contain the most stable stem loop region. In addition, a global search for the hairpin loop in all sequenced pools (supported by the pattern search option implemented within AptaTOOLS) uncovered this motif in unstructured regions in at least two additional aptamer families which also showed cycle-to-cycle enrichment but with smaller values as compared to cluster 1. A detailed list of all analyzed mutants can be found in the Appendix, Table B.4.

We performed a similar analysis for Cluster 2 identifying a total of 46 mutants with log-score smaller than -0.5 for each, the beneficial, and degenerative sets respectively (Appendix, Table B.5). Out of the three putatively beneficial mutants that we have experimentally tested for binding, two confirmed our prediction of increased affinity to the target as compared to the seed sequence. In this case, manual analysis did not reveal any clearly conserved single stranded regions or other striking properties that would by eye distinguish beneficial and degenerative mutants. This prompted us to use additional computational analysis to confirm the consistency of our predictions. Interestingly, more than half $(24/46)$ of the sequences belonging to the degenerative set had two mutations per sequence, allowing us to construct a phylogenetic tree using PAUP [87]. Our null hypothesis was that if our predictions were random, the positive and negative sets would be arbitrarily mixed in the tree's branches. This however was not what observed. Rather, except for the sub-tree containing the false negative

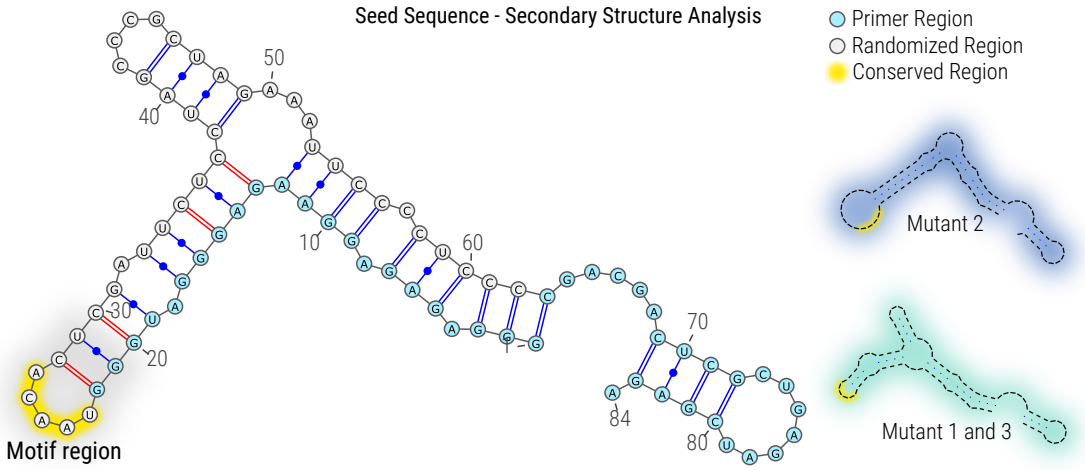


Figure 4.12: Structural analysis of the mutants of seed with ID 1 showing a conserved hairpin (indicated in yellow). The hairpin showed significantly less mutations in the set of top beneficial mutants compared to the mutation rates in the set of degenerative mutants. Structures were predicted using MFold [86] with standard parameters. Alternative structures induced by nucleotide substitutions are highlighted in blue and aquamarine respectively. The mutant with the highest improvement in binding affinity (Mutant 2) correlates with the most stable stem loop.

experimentally tested for binding affinity, the beneficial and degenerative mutants formed separated branches in the tree (see Figure 4.13). This clustering of mutants, either of the positive set or the negative set to the same evolutionary branches, validated that AptaMUTs scoring assignment is consistent with aptamer evolution and not a random pattern.

4.4.3 Discussion

In this section, we focused on the evolution of sequences in the context of an HT-SELEX experiment. Notably, the analysis of the sequence landscape has to start from an understanding of the properties of the initial pool. If all possible sequences where present in the pool and no polymerase errors or any type of biases occurred, we would expect that consecutive iterations of the SELEX procedure will converge to optimal binders (Figure 4.2 A). However, contrary to these theoretical predictions and consistent with other reports [68, 88], we found the most frequent aptamers are not necessarily the best binders. It is important to appreciate that if the randomized sequence region is relatively long, then the initial pool covers the universe of all possible sequences very sparsely. We provided a general formula that allows an estimate of how sparse the sampling is. This observation has two consequences. First, in the initial cycles, the process of partitioning into the pool that is sequenced and the pool that goes to the next cycle is expected to be very noisy and this noise is amplified in subsequent cycles. Therefore cycle-to-cycle enrichment, which is independent on the starting point but rather captures the enrichment of already abundant sequences, would be a better predictor of binding propensity than the current norm of using absolute counts. The second conse-

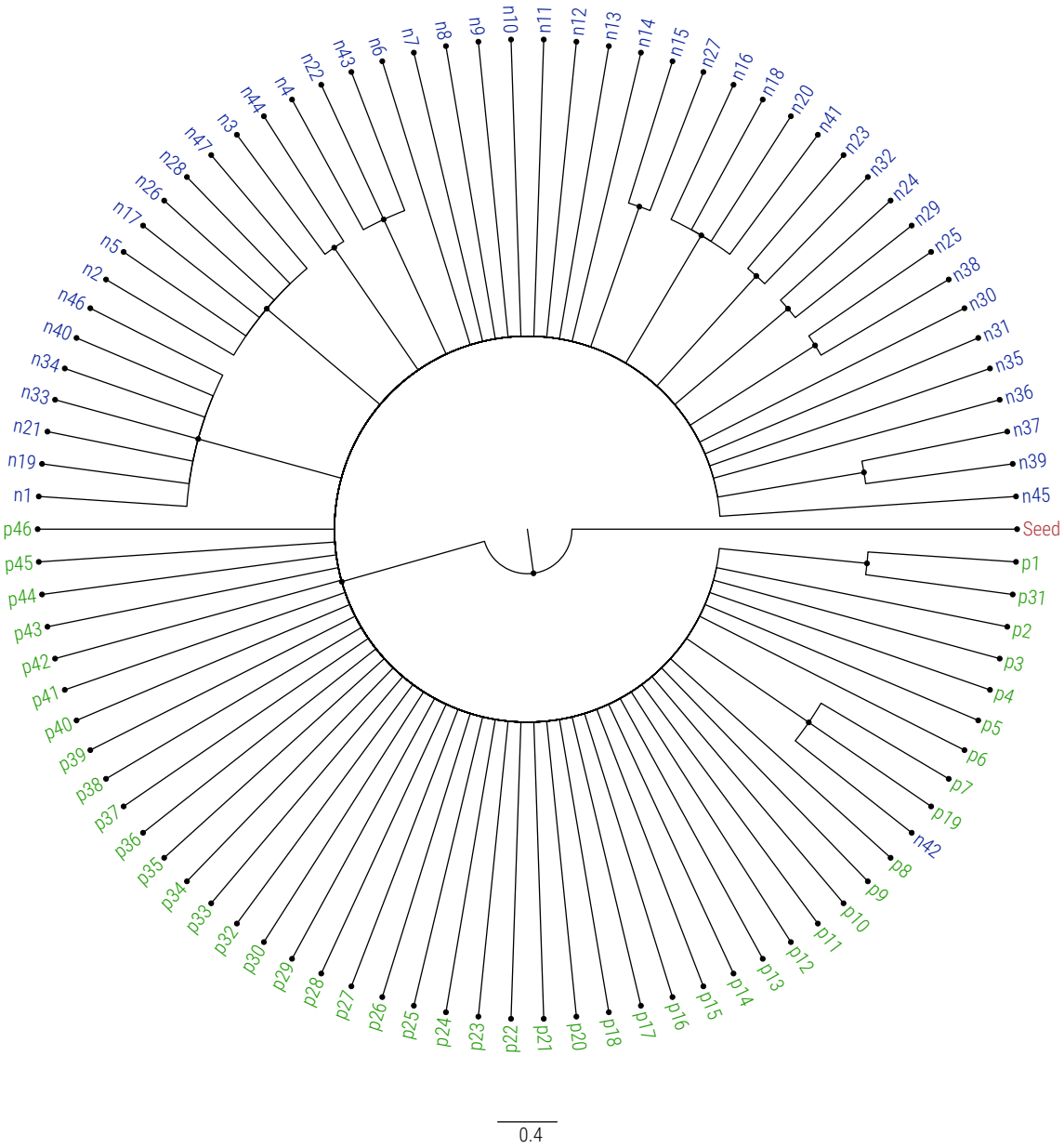


Figure 4.13: Phylogenetic tree of the mutants from cluster ID 2. Leafs labeled with p (green) correspond to the set of beneficial mutants in the order as depicted in Supplementary Table 3, where leafs starting with n (blue) stand for the degenerative species. The tree was constructed with PAUP* version 4.0 beta using the heuristic search option (1 Mio iterations) and an initial tree construction by adding species in the order of increasing log-score.

quence of the scarcity is that it is rather unlikely that any given sequence from the initial pool is "optimal" with respect to binding no matter how frequent it is (Figure 4.2 B). Instead, they merely mark the sequence neighborhood where the good binders might be. Incidentally, polymerase errors can help to explore these neighborhoods (Figure 4.2 C). Thus it is important to be able to predict which mutated sequences are likely to improve the binding.

Our AptaMUT procedure is designed, and experimentally and computationally validated, to serve this purpose. The consequences of identifying such beneficial mutations go beyond

identifying a better binder. We have demonstrated that the analysis of these mutants can help to identify important features related to binding, such as structural stability or sequence properties. Such subtleties are critical to many increasingly sophisticated applications for aptamers. Currently there are no methods, either experimental or computational, that address such requirements. Thus, with proper computational tools, polymerase errors can be leveraged to increase sampling density around the most important points of the sequence landscape and to provide valuable information about sequence properties that are important for binding.

4.5 AptaTRACE: Sequence-structure Motif Identification

Until now, our algorithm-driven data analysis approaches AptaSIM, AptaCLUSTER, and AptaMUT have tackled the challenge of aptamer discovery and protocol optimization from distinct, yet complementary angles. While AptaSIM is aimed at realistically reproducing SELEX experiments *in silico* in order to aid the discovery of global selection properties in real data and to provide a reasonable platform for the generation of benchmarking data sets, AptaCLUSTER is capable of efficiently extracting aptamer families from high-throughput sequencing data and tracing these families throughout the different selection rounds. This, among additional insights, consequently enabled a high-resolution survey of the mutational landscape of the HT-SELEX protocol through the application of AptaMUT on these aptamer families.

Despite these methods being highly distinct in their basic concepts, they all operate on full aptamer sequence in order to identify species with the desired properties with respect to the target and the researchers preference. These candidates are consequently synthesized and post-processed *in vitro*, often by reducing their overall size to the relevant target-binding regions while maintaining their structural stability. This refinement process is typically highly time consuming and cost intensive as it involves an iterative procedure of informed trail and error until all biochemical requirements for the aptamers intended use have been satisfied. At the same time, the vast amount of data produced by modern HT-SELEX experiments has opened the doors for a global and simultaneous exploration of aptamers and their overall properties. Notably, this poses the question whether the same data can also be utilized to directly uncover more localized features, responsible for characteristics such as target affinity and specificity, shared among multiple aptamer species .

Indeed, one of the most challenging properties to discover concerns the identification of aptamer regions that facilitate binding to the target. However, the development of universal methods for the discovery of binding motifs in HT-SELEX data is challenged by the vast diversity of selection conditions and technological varieties each selection cycle can be accomplished with. Even more importantly, the complexity of the target is also of great relevance. As a case in point, it has been shown that *in vitro* selection against transcription factors, and other molecules that are evolutionary optimized to efficiently recognize specific DNA/RNA targets, requires only a small number of selection rounds in order to produce high quality aptamers [36, 26]. On the other side of the spectrum, in the case of CELL-SELEX, the number of required selection cycles and the amount of non-specific binders that emerge during selection is significantly larger [29]. Such a target can in general accommodate a multitude of binding sites, each exposing different binding preferences and leading to a parallel selection towards unrelated binding motifs [30]. In addition, one of the very promising applications of aptamers concerns targeted drug delivery which builds on the principle of conjugating an aptamer selected against a specific cell surface marker (such as a cell surface receptor) with a drug of interest. This allows to deliver the drug to a distinct cell type and thus minimizes off-target effects [89, 90, 33]. In such applications, aptamers are typically optimized for high target specificity and minimal toxicity rather than maximized binding affinity. Hence, the exhaustive identification and enumeration of all possible motifs emerging from HT-SELEX experiments is crucial for the development of aptamers with tailored properties.

Current motif finding algorithms however have not been designed with these challenges in mind, and the need for the development of novel computational approaches that address the characteristics specific to the SELEX protocol has become highly relevant.

Traditionally, motif discovery has been defined as the problem of finding a set of common sub-sequences that are statistically enriched in a given collection of DNA, RNA, or protein sequences. To date, a large variety of computational methods in this area has been published (see [91, 92, 93] for a comprehensive review). In one of the earlier works, Lawrence and Reilly [94] introduced an Expectation Maximization (EM) based algorithm for finding motifs from protein sequences. This approach has been consequently adopted by various other methods [95, 96] including MEME [97] – one of the most widely used programs in this category. Lawrence *et al.* also introduced a Gibbs sampling approach for motif identification [98] which laid the grounds for other methods such as AlignACE [99], MotifSampler [100], and BioProspector [101] based on this general technique. In addition, numerous approaches have been designed based on efficient counting of all possible k-mers in a data set followed by a statistical analysis of their enrichment. Representatives for this category include Weeder [102], DREME [74], YMF [103], MDScan [104], and Amadeus [105]. Agius *et al.* designed a k-mer based Support Vector Machine discriminative framework for learning transcription factor binding preferences from high resolution *in vitro* and *in vivo* data [106]. Another group of algorithms that also allows for elucidation of motifs with mismatches is built on suffix tree techniques (Sagot [107], Pavesi *et al.* [108], and Leibovich [109]). Furthermore, regression based methods have been developed that take additional information, such as the affinity of the input sequences or the genomic regulatory contexts into account. These include, but are not limited to MatrixREDUCE [110], PREGO [111], ChIPMunk [112], and SeqGL [113]. For more information, we refer the reader to Weirauch *et al.* [114] for a comprehensive evaluation of many of the above techniques. Finally, a number of approaches for the identification of sequence motifs in HT-SELEX data targeting transcription factors (TF-SELEX) have been published. One representative of this category is BEEML [115], which is, to our knowledge, the first computational method for finding motifs on this type of high-throughput sequencing data. Assuming the existence of a single binding motif, the method aims at fitting a binding energy model to the data which combines independent attributes from each position in the motif with higher order dependencies. Another method by Jolma *et al*. approaches the problem by using k-mers to construct a position weight matrix in order to infer the binding models [26, 56]. Similarly, Orenstein *et al.* [116] also uses a k-mer approach based on frequencies from a single round of selection to identify binding motifs for transcription factor HT-SELEX data. Notably, despite of HT-SELEX' capability of generating data from multiple rounds of selection, all currently existent methods are based on the analysis of only a single selection cycle. However, choosing the round for optimal motif elucidation is not always trivial, and while some effort has been made to address this question (see for example Orenstein and Shamir [116], Jolma *et al.* [26]) this decision is ultimately left to the user.

The search for motifs in the context of RNA sequences faces another dimension in complexity as binding of ssDNA and RNA molecules is known to be sequence and structure dependent. In particular, it has been proposed that binding regions in those molecules tend to be predominantly single stranded [70, 71]. MEMERIS [73], for instance, leverages this assumption by weighting nucleotides according to their likelihood of being unpaired. These

positional weights then guide MEME to focus the motif search on loop regions. In contrast, RNAcontext [75] divides the single stranded contexts into known secondary substructures such as hairpins, bulge loops, inner loops, and stems. Consequently, RNAcontext is capable of reporting the relative preference of the structural context along with the primary structure of the potential motif. Recently, Hoinka *et al.* introduced AptaMotif [8] a method to discover sequence–structure motifs from SELEX derived aptamers. This method utilizes information about the structural ensemble of aptamers obtained by enumerating of all possible structures within a user-defined energy range from the Minimum Free Energy (MFE) structure. By representing each aptamer as the set of its unique substructures (i.e. hairpins, bulge-loops, inner-loops, and multi-branch loops), AptaMotif applies an iterative sampling approach combined with sequence-structure alignment techniques to identify high-scoring seeds which are consequently extended to motifs over the full data set. However, AptaMotif was designed for sequencing data obtained from traditional SELEX, under the assumption that this data predominantly consists of motif containing sequences. Subsequently, APTANI [117] extended AptaMotif to handle larger sequence collections via a set of parameter optimizations and sampling techniques, but it also expects a high ratio of motif occurrences.

Still, none of the above mentioned methods address the full spectrum of challenges related to analyzing data from HT-SELEX selections. First, none of these approaches, as currently implemented, scales well with the data sizes produced by modern high-throughput sequencing experiments. Next, only a few of the methods consider the existence of secondary motifs while the majority operates under the assumption that only a single primary motif is present in the data. This assumption might apply to TF-SELEX, but it cannot be generalized to common purpose HT-SELEX where many motifs of possibly similar binding strength or optimized for additional properties such as specificity and toxicity must be considered. Furthermore, secondary structure information, which has proven effective in guiding the motif search to biologically relevant binding sites, is not included in most of these methods. A notable exception is RNAContext which can handle relatively large data sets but suffers from the single motif assumption that cannot be easily removed. Finally, none of these approaches attempt to utilize the full scope of the information produced by modern HT-SELEX experiments that includes sequencing data from multiple rounds of selection.

In order to close this gap, we have developed AptaTRACE, a method for the identification of sequence-structure motifs for HT-SELEX that utilizes the available data from all sequenced selection rounds, and which is robust enough to be applicable to a broad spectrum of RNA/ssDNA HT-SELEX experiments, independent of the target's properties. Furthermore, AptaTRACE is not limited to the detection of a single motif but capable of elucidating an arbitrary number of binding sites along with their corresponding structural preferences. AptaTRACE approaches the sequence-structure motif finding problem in a novel and unique way. Unlike previous methods, it does not rely on aptamer frequency or its derivative - cycle-to-cycle enrichment. Aptamer frequency has been recently shown to be a poor predictor of aptamer affinity [4, 68, 69], and while cycle-to-cycle enrichment has shown a somewhat better performance, the choice of the cycles to compare is not obvious and does not always allow for extraction of sequence-structure motifs. In contrast, our method builds on tracing the dynamics of the SELEX process itself to uncover motif-induced selection trends.

We applied AptaTRACE to sequencing data obtained from realistically simulating SELEX over 10 rounds of selection (4 million sequences per round) with known binding motifs as well as to an *in vitro* CELL-SELEX experiment over 9 selection cycles (40 million sequences per cycle). In both cases, our method was successful in extracting highly significant sequence-structure motifs while scaling well with the 10-fold increase in data size. We verified the biological relevance of these motifs experimentally via a series of mutation studies in which either the primary or secondary structure of the motif was removed from a candidate aptamer. In both cases we observed a significant decrease in binding affinity as compared to the wild type. Our results furthermore indicate that the vast majority of motifs are residing in, and are selected for, single stranded regions, consistent with previous reports regarding RNA-target binding [70]. In addition, we observed that with sufficient sequencing depth, these motifs can be detected by AptaTRACE relatively early during the selection. Therefore the ability of AptaTRACE to handle very large input sets opens the possibility of reducing the required number of selection rounds which are typically expensive to perform in terms of time and cost.

4.5.1 Materials and Methods

SELEX Simulation Details

In order to create data sets with properties comparable to *in vitro* HT-SELEX experiments, we designed a simulation scheme capable of mimicking target-specific selection, error-prone amplification, as well as the stochastic nature of partitioning each cycle into sequencing set and selection set. Furthermore, our simulation allows for the introduction of an arbitrary number of sequence-structure motifs of different sizes and affinities, and provides control over the binding strength of background sequences. The simulation builds on our AptaSIM software [4] and extends it to take secondary structure into account. For each aptamer a, we use $|a|$, to denote its frequency and a_{bind} to denote its binding affinity.

For a user-defined pool size, a set of aptamers containing sequence-structure motifs of desired length is first produced, using a second order hidden Markov model (HMM) trained with sequences from an *in vitro* HT-SELEX experiment. To this end, we generate a sequence from the HMM, predict its secondary structure and identify all single-stranded regions larger or equal to the motif size. Next, we substitute the nucleotides in one of these unpaired regions with the motif and verify, using SFold, that the secondary structure profile of the motif region is predominantly single stranded. If the average probability of any of the single stranded secondary structure contexts in the motif region falls below 60%, we discard this sequence, and otherwise add it to the pool. This procedure is repeated as many times as required to create the desired number of sequences containing the motifs. The remaining aptamers are sampled directly from the training data in order to ensure realistic secondary structure properties that sequences generated directly from the HMM might not necessarily possess.

Using this set of sequences as input, we perform a weighted sampling without replacement according to the aptamers count and affinity, i.e. until the desired sample size is

reached, we compute a weight $w(a)$ for each aptamer a as

$$w(a) = \frac{|a| * a_{bind}}{\sum_{a \in R^x} |a| * a_{bind}}, \tag{4.13}$$

sample a sequence according to this distribution and adjust the weights to account for the removed aptamer (R^x refers to all unique sequences in pool x). The resulting pool is then amplified by means of virtual PCR in which the amplification efficiency of the polymerase, as well as the mutation rate is adjustable. In order to simulate the sampling effect, we inject a user-defined percentage of singleton aptamers with low affinity, sampled from the training data, into the pool, and reduce the pool size to its original size by weighted sampling without replacement where the weights are generated according to the aptamer counts.

After each cycle, the pool is stored in FASTQ format and used as input for our AptaTRACE algorithm.

CELL-SELEX and Flow Cytometry Experimental Details

The experimental details regarding the CELL-SELEX experiment as well as the flow cytometry assays are detailed in the Appendix, Chapter B.6.

4.5.2 Algorithmic Description of AptaTRACE

We start with a high-level outline of the method, followed by a more detailed description and refer the reader to the Appendix Chapter B.7 for a comprehensive list of the parameters used throughout the remainder of this section.

Top Level Description of the Algorithm

Our method builds on accepted assumptions regarding the general HT-SELEX procedure. First, we assume that the affinity and specificity of aptamers are mainly attributed to a combination of localized sequence and structural features that exhibit complementary biochemical properties to a target's binding site. Given a large number of molecules in the initial pool it is expected that such binding motifs are embedded in multiple, distinct aptamers. Consequently, during the selection process, aptamers containing these highly target-affine sequence-structure motifs will become enriched as compared to target non-specific sequences. Notably, under these assumptions, aptamers that contain only the sequence motif without the appropriate structural context are either not enriched at all, or enriched to a much lower degree. The second critical assumption we make is the existence of a multitude of sequence-structure binding motifs that either compete for the same binding site, or are binding to different surface regions of the target [32, 30].

Leveraging the above properties of the SELEX protocol, AptaTRACE detects sequence-structure motifs by identifying sequence motifs which undergo selection towards a particular

secondary structure context. Specifically, we expect that in the initial pool the structural contexts of each k-mer are distributed according to a background distribution that can be determined from the data. However, for sequence motifs involved in binding, in later selection cycles, this distribution becomes biased towards the structural context favored by the binding interaction with the target site. Consequently, AptaTRACE aims at identifying sequence motifs whose tendency of residing in a hairpin, bugle loop, inner loop, multiple loop, dangling end, or of being paired converges to a specific structural context throughout the selection. To achieve this, for each sequenced pool we compute the distribution of the structural contexts of all possible k-mers (all possible nucleotides sequences of length k) in all aptamers.

Next, we use the relative entropy (KL-divergence) to estimate, for every k-mer, the change in the distribution of its secondary structure contexts (K-context distribution, for short) between any cycle to a later cycle. The sum of these KL-divergence scores over all pairs of selection cycles defines the context shifting score for a given k-mer. The context shifting score is thus an estimate of the selection towards the preferred structure(s). Complementing the context shifting score is the K-context trace, which summarizes the dynamics of the changes in the K-context distribution over consecutive selection cycles.

In order to assess the statistical significance of these context shifting scores, we additionally compute a null distribution consisting of context shifting scores derived from k-mers of all low-affinity aptamers in the selection. This background is used to determine a p-value for the structural shift for each k-mer. Predicted motifs are then constructed by aggregating overlapping k-mers under the restriction that the structural preferences in the overlapped region are consistent. Finally, Position Specific Weight Matrices (PWM) of these motifs, specifically their sequence logos, along with their motif context traces (the average k-context traces of the k-mers used in the PWM construction), are reported to the user.

Detailed Description of AptaTRACE

AptaTRACE takes as input the sequencing results from all, or a subset of selection cycles from an HT-SELEX experiment and outputs a list of position specific weight matrices (PWMs) along with a visual representation of the motifs structural context shift throughout the selection.

K-context and K-context Distribution: Any individual occurrence of a k-mer in an aptamer has a specific secondary structure context called K-context that depends on the structure of that particular aptamer. In what follows, let K_i be the i-th k-mer (using an arbitrary indexing of all 4^k possible k-mers over the alphabet $\Omega_s = \{A, C, G, T\}$). In addition, let R^x be the set of unique aptamers sequenced in selection round x that have a frequency above a threshold α (this facilitates noise reduction - see computation of p-value below and Fig. 4.14, A).

First, for every aptamer a of fixed length n in R^x, we use SFold [118] to estimate the probability for each nucleotide in a of being part of a hairpin (H), an inner loop (I), a bulge loop (B), a multi-loop (M), a dangling end (D), or being paired (P) (Fig. 4.14, B). Each aptamer a is hence associated with a matrix of dimension $|\Omega_C| \times n$, where $\Omega_C = \{H, I, B, M, D, P\}$, in which rows correspond to a particular context C while each column contains the context

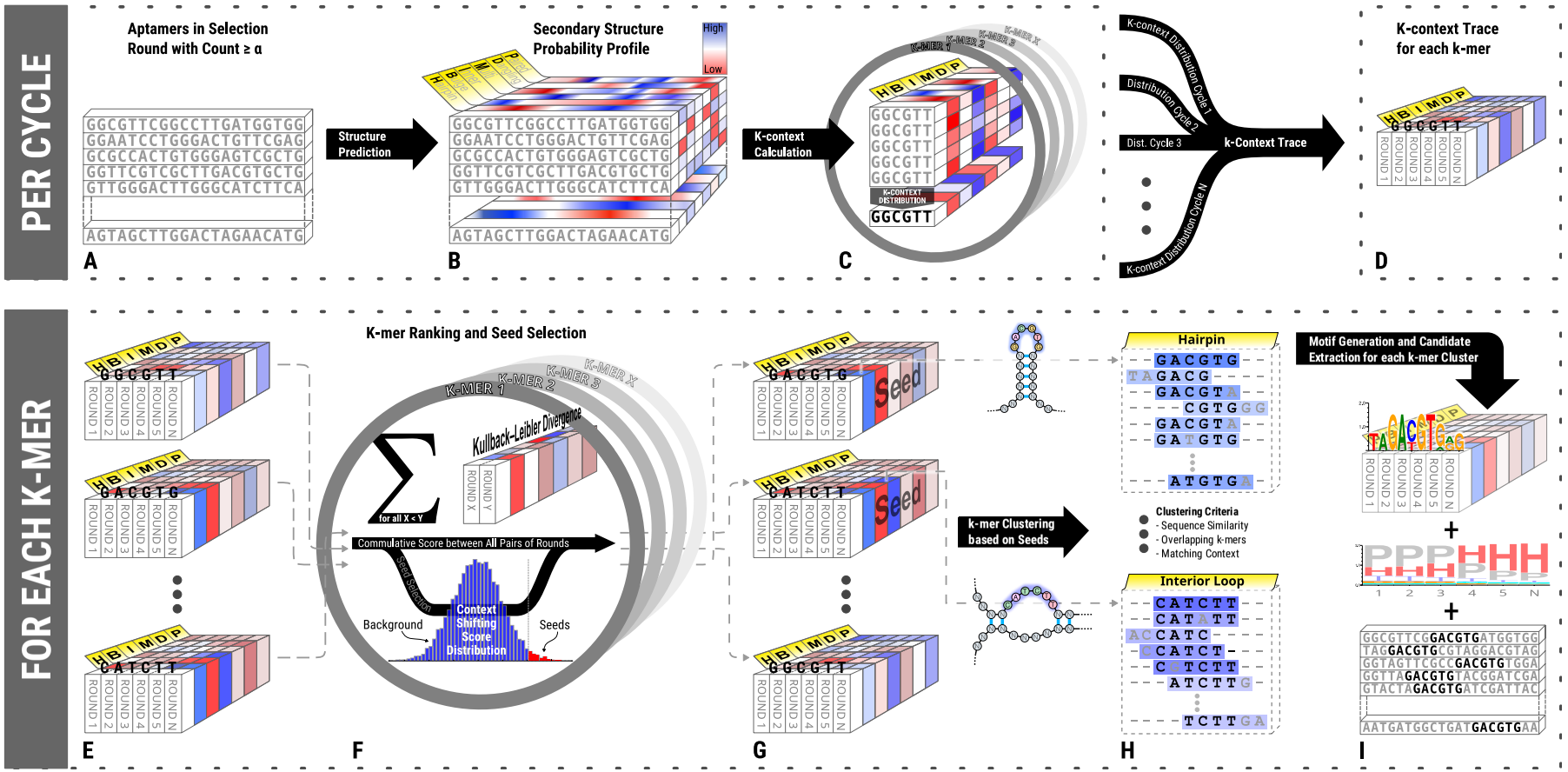


Figure 4.14: Schematic overview of our AptaTRACE method. **(A)** For each cycle, all sequences with frequency above a user defined threshold α are selected as input. **(B)** Computation of secondary structure probability profiles for each aptamer using SFOLD. For each nucleotide the profile describes the probability of residing in a hairpin, bugle loop, inner loop, multiple loop, dangling end, or of being paired. **(C)** K-context and K-context distribution calculation for each k-mer. **(D)** Generation of the K-context trace for each k-mer. **(E-G)** K-mer ranking and statistical significance estimation. Given any two selection cycles the relative entropy (KL-divergence) is used to estimate the change in the distribution of its K-context distribution. The sum of these KL-divergence scores over all pairs of selection cycles defines the context shifting score for a given k-mer. In order to assess the statistical significance of these context shifting scores, a null distribution is computed consisting of context shifting scores derived from k-mers of all low-affinity aptamers in the selection (frequency $\leq \alpha$). This background is used to determine a p-value for the structural shift for each k-mer. Top scoring k-mers are selected as seeds. (H) Predicted motifs are constructed by aggregating k-mers overlapping with the seed under the restriction that the structural preferences in the overlapped region are consistent. **(I)** Position Specific Weight Matrices representing these motifs, along with their K-context traces and corresponding aptamers are reported to the user.

probabilities of the corresponding nucleotide in a. Next, we define the K-context of a k-mer occurrence in aptamer a as the row-wise mean of the context probabilities over the matrix columns corresponding to the location of that k-mer in the aptamer sequence.

Recall, that the main idea behind AptaTRACE is to track the changes in secondary structure preferences of k-mers over the selection cycles. Capturing these secondary structure preferences should therefore take the entirety of K-contexts from all occurrences of a k-mer in a particular selection cycle into account. Thus, we define the K-context distribution of a k-mer K_i in round x as the averaged secondary structure profile of all K-contexts of K_i over all aptamers in R^x. Formally, let $\mathbb{P}_i^x(C)$, where $C \in \Omega_C$, be the average probability of the structural context C over all occurrences of the k-mer K_i in all aptamers that meet the threshold criteria in round x. Then, the K-context distribution of K_i in round x is the vector $\mathbb{P}_i^x = [\mathbb{P}_i^x(H), \mathbb{P}_i^x(B), \mathbb{P}_i^x(I), \mathbb{P}_i^x(M), \mathbb{P}_i^x(D), \mathbb{P}_i^x(P)]$, normalized such that all entries sum up to one. (Fig. 4.14 C).

Analysis of the Shift of K-context Distributions during Selection: If a k-mer forms part of a sequence-structure binding motif, its K-context distribution is expected to shift towards the context C that is preferred for the binding interaction throughout the selection. In contrast, if a k-mer is not affected by selection, we expect little to no change in its context distribution over consecutive rounds. We can capture this dynamics for any k-mer K_i by its so called K-context trace $\mathbb{K}_i = [\mathbb{P}_i^0, \mathbb{P}_i^1, \ldots, \mathbb{P}_i^m]$, defined as a vector tracking the K-context distribution over all m selection cycles (Fig 4.14 D). Our method consequently quantifies such shifts in the K-context distribution using the Kullback–Leibler divergence (relative entropy) – a measure for the difference between two probability distributions. Here, for any k-mer K_i the first distribution corresponds to the K-context distribution $\mathbb{P}_i^x$ of an earlier round x and the second to the K-context distribution $\mathbb{P}_i^y$ of a later selection cycle y.

The KL-divergence between two appropriately chosen selection cycles might suffice, at least for some scenarios such as TF-SELEX, to capture the shifts in K-context distributions. In practice however, for larger and more complex targets the selection landscape tends to be increasingly complicated, with various aptamers achieving peak enrichment at different selection cycles. This convolutes the task of confidently choosing such two presumably most informative cycles while ignoring the remaining information. Therefore, to capture the totality of changes during the entire selection, we compute the cumulative KL-divergence between all pairs of sequenced pools. In summary, we define the context shifting score $score(K_i)$ for k-mer K_i as

$$\begin{aligned} score(K_i) &= \sum_{x=1}^{m-1} \sum_{y=x+1}^{m} D_{KL}(\mathbb{P}_i^y \| \mathbb{P}_i^x) \\ &= \sum_{x=1}^{m-1} \sum_{y=x+1}^{m} \left[\sum_{C \in \Omega_C} \mathbb{P}_i^y(C) \times \log \frac{\mathbb{P}_i^y(C)}{\mathbb{P}_i^x(C)} \right], \end{aligned}$$

where $D_{KL}(P\|Q)$ is the Kullback–Leibler divergence between two discrete distributions P and Q. To ensure statistical accuracy, the context shifting scores are only calculated for all k-mers with a count of at least β individual occurrences in each pool (here, $\beta = 100$).

Significance Estimation and p-value Computation: While the context shifting score establishes a ranking of the k-mers in order of their overall change in secondary structure context, it does not provide any information over the statistical significance of that shift, i.e. it cannot distinguish between changes in response to the true selection pressure and changes associated with background noise such as non-binding species. These background species however are expected to occur in very low numbers throughout the selection. We leverage this property by using the context shifting scores of the k-mers from these low-count aptamers to construct a null distribution that is used to identify the significant context shifting scores for the full data set. In detail, we include all k-mer occurrences from aptamers that are not included in the previous generation of the context profiles, i.e. all aptamers below or equal to the user defined threshold α. α is chosen as the smallest integer T starting from 1 such that there is at least two thirds of aptamers in any selection round except for the initial pool that have frequencies less than or equal to T. We note that the resulting null follows a log-normal distribution in our *in vitro* experiment as well as for the simulation data presented in this study (see Section 4.5.3 for details).

The above described procedure hence allows for the computation of a p-value for each K-context trace and we only retain those K-context traces with p-value below a user specifiable threshold (default: 0.01, Fig.4.14 E-G).

Elucidating Sequence-Structure Motifs and Sequence Logos: In the last step, AptaTRACE proceeds to extract the final motifs by clustering similar and overlapping k-mers with correlating, statistically significant structural shifts together (Fig 4.14 H). This allows to uncover sequence-structure motifs that might extend over the chosen k-mer size and to build PWMs that summarize the motifs. The motif construction is accomplished iteratively. First, k-mers with significant context shifting scores are sorted in decreasing order according to their k-mer frequencies in the last selection cycle. We then select the top k-mer as a cluster seed, iterate through each of the remaining k-mers, and aggregate these to the cluster if they are sufficiently similar in primary structure and strongly overlap with the seed k-mer (see Section B.7 for a detailed description of this rather straight forward approach). Structurally, the K-context trace of the new cluster members must also display similar structural shifts that correlate with the seed in order to be included into the cluster. The similarities of the context shifts are quantified by identifying the specific structural context C' that changes the most over the selection rounds for each $\mathbb{K}_i$:

$$C' = \arg\max_{C} f(C) = \sum_{x=1}^{m-1} \sum_{y=x+1}^{m} \left[\mathbb{P}_i^y(C) - \mathbb{P}_i^x(C) \right]$$

If C' coincides with the dominant context of the seed, we include the K-context to the cluster.

The general notion of the SELEX protocol is that target affine motifs are simultaneously selected for across multiple aptamer families (an aptamer from the initial pool together with all its mutants emerging throughout the selection). In complex data landscapes such as those produced by CELL-SELEX experiments, aptamer families can arise which might contain additional conserved sequence-structure regions which co-occur with the main motif only in

a minority of family members. Due to their limited number of occurrence, these "passenger motifs" are not expected to be relevant for the aptamers biological activity and are therefore to be considered as noise. We solve this scenario by including an additional filtering step in order to retain only those clusters which represent motifs occurring independently across multiple aptamer families. Hence, for each additional cluster to be included in AptaTRACE's output, we perform the following validation scheme (additional information is provided in Section B.7): Let L be set of retained clusters so far, let $S(x)$ stand for the set of aptamers in the last cycle containing the motif x, and let $S(X)$ correspond to the set of aptamers that contains any motif in any motif set X. After obtaining the motif x, we only retain x if the set of aptamers containing motifs already present in L is less than γ ($\gamma \in [0, 1]$) of the cardinality of $S(x)$, i.e.

$$|S(x) \cap S(L)| \leq \gamma \, |S(x)|$$

Finally, the resulting motifs are reported to the user via their PWMs, sequence logos and their motif context traces, defined as the averaged K-context traces of those k-mers constituting the PWM. The set of aptamer candidates that satisfy both, the primary and secondary structure properties of the motifs, sorted by their statistical significance or frequency of occurrence is also included in the output (Fig. 4.14 I).

Choosing an Appropriate Value for k: The number of clusters that results from the application of our approach to a particular dataset, can also serve as a measure for an appropriate choice of the k-mer size for a particular data set. By choosing k such that it maximizes the number of the resulting clusters, we ensure to capture all motifs in the sequencing data.

Implementation Details and Runtime: Our AptaTRACE pipeline is implemented as a modular system in C++ and Java. RNA secondary structure profiles for all aptamers in the pool were predicted in parallel on the server farm at the National Center for Biotechnology Information, NIH, using 1000 Cores and 200MB of RAM per job. The simulated data required about 5 hours of wall clock time, whereas the *in vitro* selection data finished in roughly 4 days. The enumeration of k-mers and consequent extraction of k-contexts was implemented as a multi-threaded C++ library and is capable of processing the here presented results in approximately 10 hours for the combined choices of k between 5-8 on a large memory server with 100 CPUs. Finally, the context shifting scores, k-mer extraction, and clustering are implemented as a multi-threaded JAVA program requiring 1-4 hours, depending on the data size, for its completion.

4.5.3 Results

We start by using simulated data produced with a novel, extended version of our AptaSIM program [4] to compare the performance of AptaTRACE to other methods that can handle similar data sizes or incorporate secondary structure into their models. Finally, we show our results of applying AptaTRACE to an *in vitro* selection consisting of high-throughput data from 9 rounds of cell-SELEX (see Section B.6 in the Appendix for experimental details).

	Motif	B.A.	AptaTRACE (all cycles)	AptaTRACE (last cycle)	DREME (last cycle)
(a)	GTGTACTA	75%			
(b)	AGTCGG	70%			
(c)	GTTAA	60%			
(d)	GGAACTC	55%			
(e)	GGAGCG	50%			

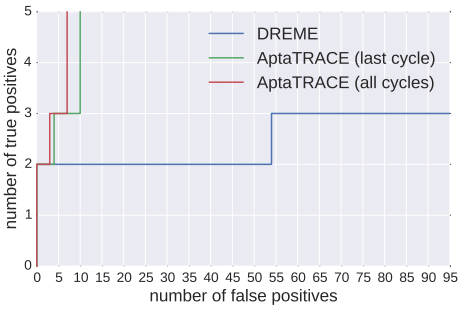


Figure 4.15: Left: Comparison of AptaTRACE against other methods based on simulated data. AptaTRACE was applied to the entire dataset as well as to the last selection cycle only. While our method successfully identified all implanted motifs, DREME was only able of extracting 3 out of 5. Shown in the first two columns are the implanted motifs and their binding affinity used throughout the selection (B.A.). The output PWMs produced by the tested methods that correspond to the implanted motifs are displayed in the remaining columns. **Right:** Plot depicting the number of false positive motifs reported by DREME (blue line) and AptaTRACE on the x-axis against the number of true positives recovered on the y-axis. For AptaTRACE, we utilized only the initial pool and the last cycle as input (green line), and the full simulated dataset (red line). The former yielded 15 motifs, while applying our method onto the full dataset a total of 12 motifs were identified.

Results on Simulated Data

To test our new approach, we applied AptaTRACE to a data set generated by means of *in silico* SELEX as no benchmarking set that could be used as a gold standard is currently available. To this end, we used an extension to our AptaSIM program [4] designed to realistically simulate target-specific selection including, among other factors, species affinity, polymerase amplification and polymerase errors, and the effects of sampling from the selection pools for sequencing. Our current extension additionally allows for implanting sequence-structure motifs with well defined properties. We generated a data set of 4 million sequences per round containing 5 motifs (denoted here as motifs (a)-(e)), 5-8 nucleotides in length, and located predominantly in unpaired regions. Note that the motifs' primary structures also occur randomly in the background aptamers, albeit in arbitrary structural contexts, and that the motifs are hence not over-represented in the initial pool. Each motif was initially present in 100 different target-affine aptamer species and consequently selected for over 10 rounds of SELEX. A complete description of the simulation as well as the parameters used during *in silico* SELEX are available in Supplementary Information, respectively.

We applied AptaTRACE, as well as DREME and RNAcontext to the data set to compare their capability of extracting these motifs. Since DREME and RNAcontext can only be applied to one selection round at a time, we provided these two approaches with data from the last selection cycle alone, choosing the initial pool as background when required. Since DREME and RNAcontext only utilizes the initial pool and the last selection round, AptaTRACE was applied to these two pools as well. Notably, RNAcontext was not capable of handling 4 million sequences in a reasonable time frame, prompting us to sample the 10000 most frequent and least frequent sequences of the last selection cycle as input. The full scope of parameters used for these methods during the comparison are detailed in Parameters used in this Study

in Supplementary Information.

Since RNAcontext's model assumes a single motif in the whole dataset, a direct comparison would not be fair for that software. Nonetheless, we examined the possibility of the method of identifying at least one binding site due to the large abundance of implanted motif (a) in the final selection round, however without success. Fig. 4.15 (left) summarizes the results of AptaTRACE when applied to the full dataset, as well as to the last selection cycle only, compared to DREME's performance. While DREME failed to identify the low-affinity motifs (d) and (e), AptaTRACE was able to recover all motifs in both test scenarios. In addition, AptaTRACE exhibits, by a large margin, the lowest false discovery rate compared to other methods (see Fig. 4.15 (right))

A more detailed summary of the sequence logos extracted by our approach on the full data set, including their motif context traces and statistical significance, is available in Fig. 4.16. Interestingly, a visual inspection of the motif context trace (last column, Fig. 4.16) points to the possibility of capturing most of these motifs at earlier cycles. Indeed, computing the selection round in which a motif was first detected by AptaTRACE (column C^*, Fig. 4.16), confirmed this expectation.

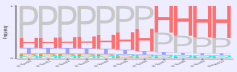
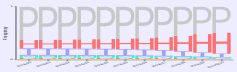
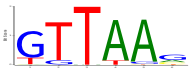
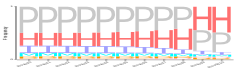
ID	Sequence Logos	Logo Seed	Seed p-value	Seed Freq.	C^*	Motif Context Trace
						0 1 2 3 4 5 6 7 8 9 10
1		GTGTAC	2.44E-33	11.30%	5	
2		AGTCGG	1.44E-37	6.02%	5	
3		GGAGCG	2.52E-23	1.45%	8	
4		GTTAAG	7.65E-24	1.09%	9	
5		GGAACT	7.90E-28	1.02%	10	

Figure 4.16: Sequence-structure motifs identified by AptaTRACE from virtual SELEX given all 10 selection cycles, including the initial pool, as input. AptaTRACE was able to recover all 5 motifs. Shown here are the identified sequence logos, the k-mer that scored highest in significance used for construction of each motif (seed) and its p-value, the abundance of seed of the motif in the final selection round (Frequency), the first cycle at which the motif was detected (C^*), as well as the motif context trace throughout the selection from the initial pool to round 10.

Results on CELL-SELEX Data

Next, we applied AptaTRACE to the results of an *in vitro* CELL-SELEX experiment targeting the C-C chemokine receptor type 7 (CCR7), where the initial pool as well as 7 of 9 selection rounds

have been sequenced, averaging 40 million aptamers per cycle (see Material and Methods Section B.6 for a detailed description of the experimental procedure). We did not challenge DREME with this task, since this data set is 10-fold larger in size compared to the simulated selection, and even in the latter case DREME required approximately 4 days to complete. AptaTRACE was able to successfully extract a total of 9 motifs, the five most frequent of which are shown in Fig. 4.17, and the full list is given in Supplementary Information.

The context trace of these motifs hints towards two properties of the selection process.

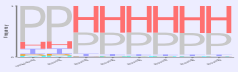
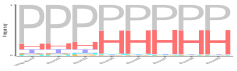
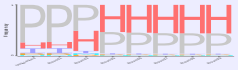
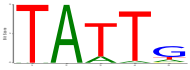
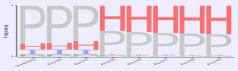
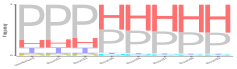
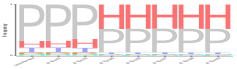
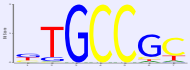
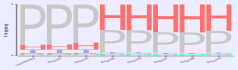
ID	Sequence Logo	Logo Seed	Seed p-value	Seed Freq.	Motif Freq.	Motif Context Trace
						0 1 3 5 6 7 8 9
1		CTGTG	3.43E-03	15.40%	33.00%	
2		CGCTG	2.47E-03	6.57%	11.30%	
3		TGCGC	2.59E-04	5.05%	11.13%	
4		TATTG	2.51E-03	4.83%	12.24%	
5		CTGGC	9.59E-4	4.82%	11.87%	
6		TGGTG	2.241E-3	4.67%	5.22%	
7		CTGCA	1.29E-3	4.54%	4.54%	
8		GCGTG	3.6E-3	3.46%	3.46%	
9		TGCCG	2.09E-4	1.37%	2.56%	

Figure 4.17: The full set of sequence-structure motifs as produced by AptaTRACE on CELL-SELEX data. The sequence logo as well as the most frequent k-mer constituting the logo (Logo Seed) and its p-value are depicted for each motif together with its seed frequency. The motif context trace for the sequenced cycles (0,1,3,5,6,7,8,9) is shown in the last column.

First, a clear selection towards single stranded regions for every extracted motif can be observed. It has always been postulated that ssDNA/RNA binding motifs are predominantly located in loop regions [71]. Indeed, this assumption was leveraged by MEMERIS, by imposing structural priors directing the motif search towards single stranded regions. In the case of AptaTRACE, no prior assumption of this type was made. The fact that despite a lack of such priors, motifs detected by AptaTRACE conform with the expected properties of RNA

sequence-structure binding sites support their relevance for binding.

Experimental Validation

We further substantiated our findings *in vitro* by performing a number of flow cytometry based binding assays using target expressing HeLa cells as positive controls, and HeLa cells in which the target is repressed as negative controls (see Section B.6 experimental for details). Using the most prevalent sequence-structure motif as reference, we selected two highly enriched aptamers denoted as C1-A and C2-A (see Figure 4.18A), which contain this motif in a hairpin located at the far 3' end of the randomized region. In order to verify that both sequence and structure are responsible for the binding interaction with the target, we additionally engineered four control experiments based on C2-A in which we either preserved the secondary structure of the aptamer but replaced the primary structure of the motif with an arbitrary sequence not related to any motifs identified by AptaTRACE (NEG1-DEL1 and NEG1-DEL2), or selected aptamers from the pool which retained the primary structure of the motif but in which the secondary structure is contained within a paired region of the species (NEG1-1 and NEG1-2). Our binding assays, as depicted in Figure 4.18B, show that while aptamer C1-A exhibits the highest affinity to the target, C2-A shows substantially greater specificity. More importantly, our control experiments demonstrate that eliminating either sequence or structure from the motif results in a significant decrease of binding ability to the target, strengthening our argument that AptaTRACE is capable of identifying biologically relevant sequence-structure motifs from complex HT-SELEX data. Notably, removing the secondary structure component from the motif resulted in the largest drop in affinity as compared to replacing the primary structure only, further validating the correlation between motifs located in unpaired regions and expected binding affinity.

Trade-off between the Number of Selection Cycles and Sequencing Depth

The ability of AptaTRACE to analyze very large data sets opens the possibility of reducing the number of selection cycles by increasing the sequencing depth. Such a reduction in number of cycles is desirable for two main reasons. First, with current technology, the cost savings from reducing the number of cycles outweighs the added cost due to deeper sequencing. Next, a decrease in the amount of selection rounds allows to reduce the number of potential artifacts that can accumulate during this multi-step procedure. Since the sequencing depth of our CELL-SELEX experiment significantly exceeds current practices, we were able to explore the relationship between (a) sequencing depth, (b) the number of required selection cycles, and (c) the number of identified motifs by AptaTRACE. For this purpose, we performed a series of sampling tests on the original data. Specifically, we iteratively reduced the number of selection cycles down to the first round and randomly selected 5%, 10%, 20%, and 100% of the sequences in each round. We then utilized AptaTRACE on the scaled down data sets and computed the ratio between the number of identified seed k-mers and significant k-mers compared to running AptaTRACE on the full dataset.

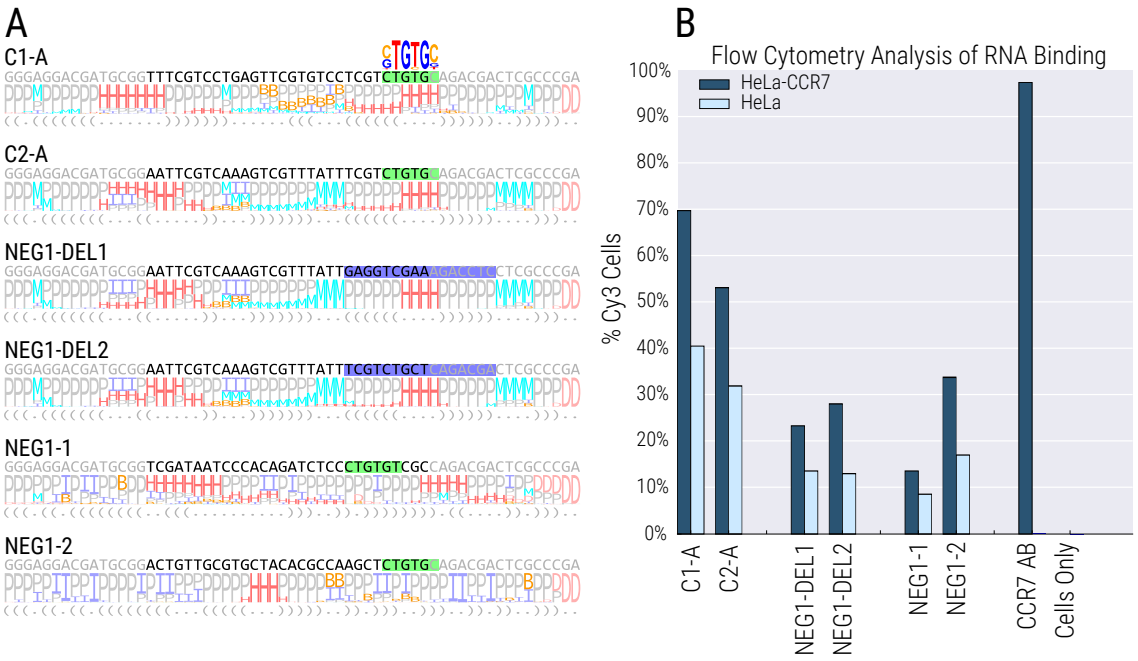


Figure 4.18: Experimental validation of sequence-structure motifs identified by AptaTRACE. **(A)** depicts the selected aptamers, their secondary structure probability profiles, as well as their MFE structure. For each species, the primer regions of the sequence are colored gray whereas the 30nt long randomized region is presented in black. The secondary structure probability profile reflects, for each nucleotide, the probability of residing in a **H**airpin, **I**nner loop, **B**ulge loop, **M**ultiple loop, or of being **P**aired. The position of the motif on the aptamer (as shown in the sequence logo) is highlighted in green and the primary structure region in which the motif was removed is highlighted in violet. **(B)** Flow cytometry analysis of the sequences as shown in (A). Depicted are the percentage of aptamer-bound cells for target-expressing HeLa cells (dark blue) and target-lacking HeLa cells (light blue) as compared to a high affinity antibody (CCR7 AB).

The results, depicted in Figure 4.19, suggest that AptaTRACE is capable of identifying motif signals from as early as selection cycle 3 and that with only 5 rounds the vast majority of motifs (80%) can be recovered. These findings therefore strongly indicate towards the possibility of trading off additional (and expensive) selection cycles in favor of deeper (and more economic) high throughput sequencing even when analyzing complex landscapes as those generated by CELL-SELEX experiments.

4.5.4 Discussion

The importance of the ability to identify sequence-structure binding motifs in HT-SELEX data, and thus to predict the features behind binding affinity and specificity, cannot be overestimated. First, for RNA binding proteins, elucidating these motifs is important for understanding their function. In the context of targeted drug therapy, where the selected aptamers are consequently conjugated with a therapeutic RNA/DNA sequence, it is crucial to know which part of the aptamer is most relevant for biding so the sequence-structure properties are preserved in the conjugate construct. Finally, in the scope of CELL-SELEX, the current practice relies on interleaving positive and negative selection steps [57]. The purpose of the negative

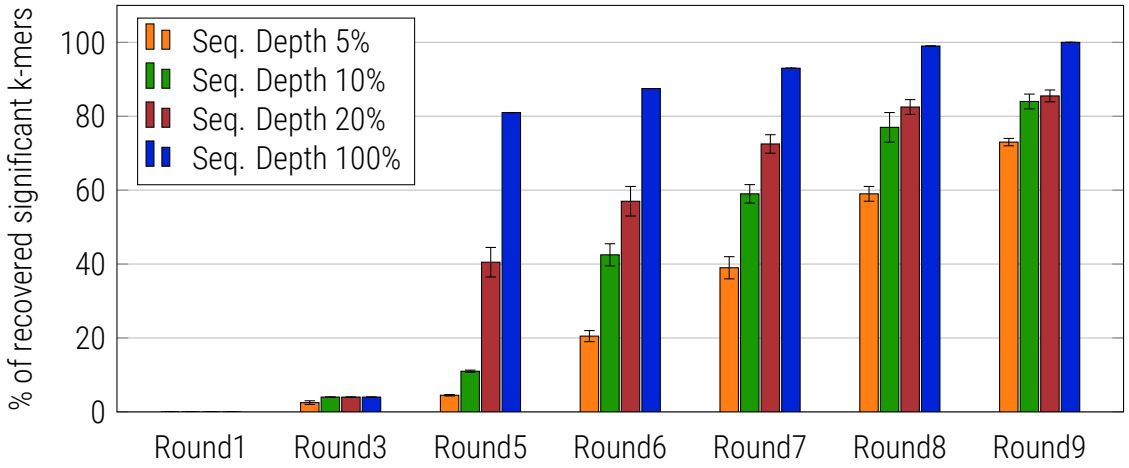


Figure 4.19: Percentage of significant k-mers identified by AptaTRACE when reducing the number of cycles and sequencing depth. Each bar corresponds to the application of our approach onto a reduced dataset containing all selection cycles up to round x while utilizing a random subset of y% reads of each cycle. The height of each bar stands for the percentage of the number of retrieved significant k-mers compared to running AptaTRACE on the full Cell-SELEX data. The standard deviations correspond to the sampling effects caused by repeating each experiment 20 times.

selection is to remove aptamers that interact with non-target molecules or non-target cells such as for instance healthy cells. However, as it can be appreciated from Figure 4.18 B, with this strategy some residual binding to the negative target persists. The development of AptaTRACE opens alternative strategies. In particular, it can be used to computationally detect the motifs responsible for binding to the negative and positive targets and thus enable rational aptamer design.

An important feature of AptaTRACE is that, rather than using quantitative information, it directly leverages the experimental design of the SELEX protocol and identifies motifs that are under selection through appropriately composed scoring functions. By focusing on local motifs that are selected for, AptaTRACE bypasses global biases such as the PCR bias, which is typically related to more universal sequence properties such as the CG content. In addition, because AptaTRACE measures selection towards a sequence-structure motif by its shift in the distribution of the structural context and not based on abundance, it can uncover statistically significant motifs that are selected for even when these only form a small fraction of the pool. This is an important property that can ultimately help to shorten the number of cycles needed for selection and thus to reduce the overall cost of the procedure.

In testing on simulated data, AptaTRACE outperformed other methods, in part because these methods were not specifically designed to handle this type of data. Furthermore, no competitors exist that could be tested on *in vitro* data as none of the current programs scale to the amount of data points produced by HT-SELEX. We successfully validated the biological relevance of the motifs identified in this study via a series of mutation studies coupled with flow cytometry experiments. Removing either the primary or secondary structure of the motif resulted in a significant decrease in target affinity. These tests, in combination with the observation that the motifs converge structurally to loop regions, provide reassuring evidence

for the correctness of our approach and are consistent with the accepted view where such binding sites reside.

Sequence logos provide a convenient visualization of the selected motifs. For TF binding, information used to derive these logos also serve to estimate their binding energy [119, 120]. However, for general HT-SELEX this connection is less immediate as one has to take into account the energy contribution from the structure component [121]. Perhaps even more importantly, aptamers binding to large cell surfaces are likely to be exposed to more binding opportunities than aptamers binding to single receptors and thus the number of resulting motifs can be expressed as a function of interaction probability and binding affinity. We hypothesize that the here presented K-context trace will be helpful in untangling some of these contributions.

Finally, analysis of the K-context trace indicates that the selection signal can be identified at very early cycles. This suggested that, with deep enough sequencing, only a limited number of selection cycles might be required and we have confirmed this expectation. In addition, our analysis also shows that the dynamics of K-context traces is not the same for all sequence-structure motifs. While most trends essentially stabilize at a relatively early cycle, some continue to grow. We hypothesize that this type of information can aid the identification of the most promising binders. Note that while we defined the K-context shifting score on all pairs of sequenced selection pools, it can also be used to focus the analysis to any part of the selection as long as it includes at least two cycles. Other variants of the K-context shifting score (e.g. always using the initial pool as background/reference in the summation) can also prove informative in discovering of additional details of the selection dynamics. AptaTRACE is not only a powerful method to detect emerging sequence-structure motifs, but also a flexible tool that can be readily adopted to interrogate such selection dynamics.

5

Visualization

> *By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you're lost in information, an information map is kind of useful.*
>
> David McCandless, *2010*

Chapter 4 provided an insight into the advancements of algorithm driven aptamer analysis enabled by the massive amount of data produced through modern HT-SELEX experiments. While approaches such as AptaTRACE aim at significantly reducing the total amount of candidate aptamers via the identification of biologically relevant sequence-structure motifs responsible for binding the target, methods such as AptaCLUSTER are more concerned with rearranging the complete dataset into meaningful sub-units and tracing these throughout the selection. This poses the question as to how to efficiently represent millions of data points and their relation between each other in a clear, concise and easily accessible manner. Given the size of the data, it might not be surprising that purely text-based tools to visualize and manipulate aptamers have proven impractical and time consuming for many researchers in this field.

To close this gap, we developed AptaGUI, a platform independent graphical user interface (GUI) for the dynamic visualization of HT-SELEX data. In particular, AptaGUI includes support for applications of our AptaCLUSTERpackage, including but not limited to data preprocessing, tracking the changes of aptamer families throughout selection cycles, computing cycle-to-cycle enrichment, aptamer candidate identification, and candidate refinement through the analysis of mutagenesis driven selection.

AptaGUI currently features four main sections focusing on (a) data quality control, (b) experimental details such as enrichment statistics throughout selection cycles, (c) sequence based analysis, and (d) aptamer family (cluster) based analysis. In addition, AptaGUI provides secure, global, multi-user access to the data-sets via our RESTful API denoted as AptaREST hence enabling real-time, collaborative studies in the field of aptamer discovery.

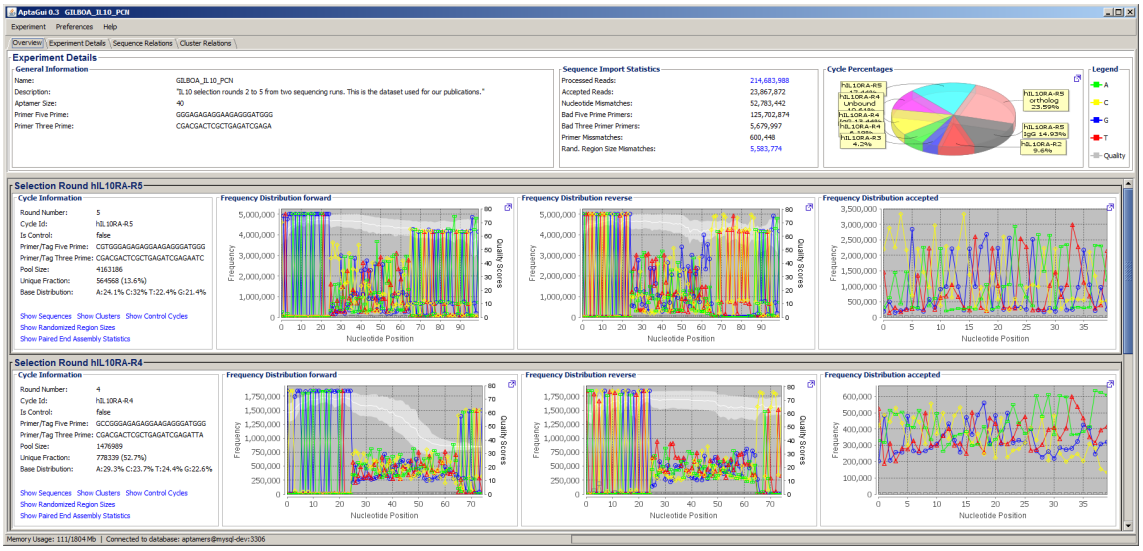


Figure 5.1: Screen capture of AptaGUI showing the *Overview* tab on the example of our IL-10 selection utilized in Section 4.3.

5.1 Program Description

AptaGUI is written in Java and uses a MySQL database containing all results produced by our software package AptaTOOLS. AptaTOOLS includes, in addition to basic analysis tools, the clustering algorithm AptaCLUSTER and the mutant analysis utility AptaMUT [4, 3]. The interface is tabulated into four sections denoted as *Overview*, *Experiment Details*, *Sequence Relations*, and *Cluster Relations*, each of which corresponding to a fundamental aspect of aptamer related data analysis.

The *Overview* tab displays general information about a selected experiment, including but not limited to quality control reports and selection round details (see Figure 5.1). In terms of quality control, it provides several statistics in the *Sequence Import Statistics* panel regarding the consistency of paired-end sequence reads, demultiplexing the reads into the different selection rounds, and extracting the randomized region. Furthermore, by clicking on the value of *Processed Reads*, the user has the option to view additional details regarding the distribution of the assembled contig sizes if paired-end data was used. This information can prove valuable for quality control purposes, especially when several selections have been multiplexed and jointly sequenced on the same lane. In a similar fashion, AptaGUI is capable of producing a bar chart depicting the distribution of the identified randomized region sizes during the preprocessing stages of the input data. This plot can be leveraged to reveal possible secondary aptamer families which differ in the expected randomized region size from the original pool that were created by PCR induced insertions or deletions but that remained in the pool due to their putative target affinity.

In terms of selection round details, for each sequenced cycle, tag-primer combinations used during selection and sequencing, the pool size, the pools nucleotide distribution, and detailed information regarding the cycles nucleotide composition for the forward reads, reverse reads

Figure 5.2: Screen capture of AptaGUI showing the *Experiment Details* tab on the example of our IL-10 selection utilized in Section 4.3.

(if paired-end data was used), and the randomized regions that passed the demultiplexing and quality control filters is displayed. Finally, for each round, the option to view all aptamers as well as aptamer families and control cycles (should these be available) is also available. Taken together, the data summarized in this tab serves as a first step to quickly asses the overall quality of not only the raw sequencing data, but also the quality of the selection process itself.

The *Experiment Details* tab (see Figure 5.2) provides information regarding global selection properties of the SELEX experiment for both sequences, and aptamer families. For sequences, it summarizes the frequency of singletons (aptamers with a count of 1), the frequency of enriched species, as well as the percentage of distinct aptamers that make up each pool. Equivalently, for aptamer families (clusters), the average cluster size and diversity (i.e. the number of unique aptamers per family), as well as the number of clusters per pool is displayed. These pool properties allow for progress assessment of the SELEX experiment and provide insight into the exponential enrichment aspects of target-affine species in the pool.

The *Sequence Relations* tab retrieves detailed information regarding sequence count (the frequency of the aptamer), fraction (the percent of this aptamer w.r.t. the pool size), and enrichment (fold change in fraction between consecutive cycles) for every aptamer in the experiment and for every selection cycle present in the experiment (Figure 5.3). The enrichment values between consecutive selection cycles can also be plotted graphically as a function of the cycle number allowing for capturing cycle specific enrichment dynamics. In addition, aptamer sequences can be queried for occurrences of a particular motif (including wild-card searches). The results are displayed according to the users specifications including sorting aptamers of a specific cycle by count, fraction, or enrichment, allowing for easy accesses

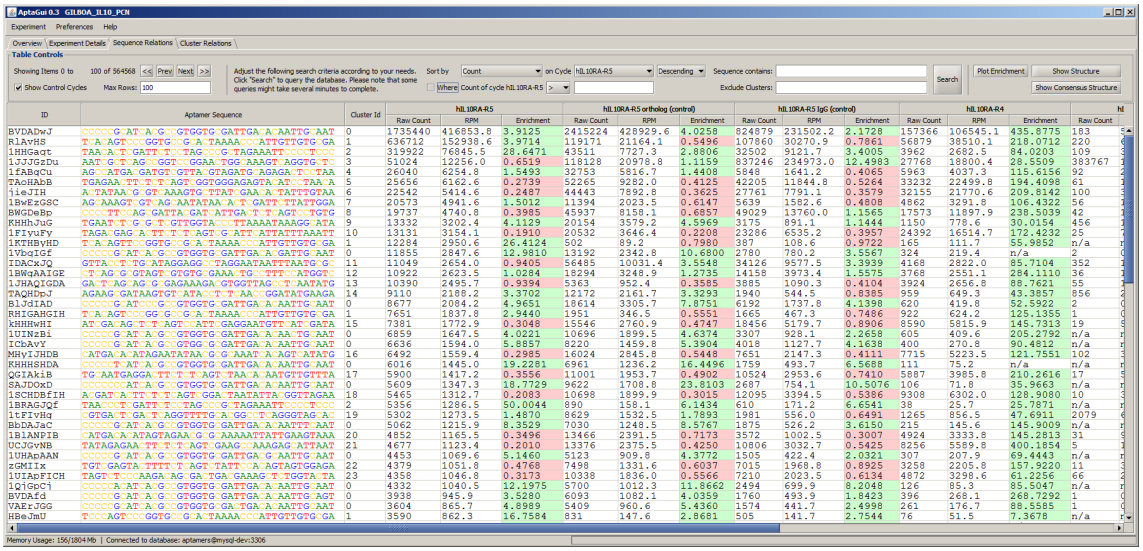


Figure 5.3: Screen capture of AptaGUI showing the *Sequence Relations* tab on the example of our IL-10 selection utilized in Section 4.3.

to aptamers of interest. AptaGUI also supports the prediction and graphical presentation of secondary structures for single aptamers as well as aptamer families (consensus structures) via the integration of PPFold and VaRNA [122, 123] into the software. Furthermore, each sequence can be annotated with custom information in real time for collaborative editing and analysis of a dataset across multiple platforms. Finally, candidate sequences can be exported (with or without primers) should any kind of post-processing by third party software be required.

The *Cluster Relations* tab provides detailed information about aptamer families identified by AptaCLUSTER and their relation throughout the selection cycles (Figure 5.4). For each family, the cluster identified in the last selection round and the corresponding clusters from all previous selection cycles are shown. Each cluster is summarized by its sequence logo and, in analogy to the properties shown in the *Sequence Relations* tab, displays its size (total number of sequences), pool fraction (the percentage the cluster occupies in the corresponding selection round), the cluster diversity (number of unique sequences in the cluster), as well as the clusters enrichment. The cluster enrichment values can be analyzed graphically on order to identify aptamer families of potential interest. Additional information for each cluster, including a list of its comprising sequences, the distribution of their counts, and a comprehensive analysis of the single nucleotide variations present in the cluster is available by clicking the *Show Sequences* button. Additionally, the user has the option to sort the clusters based on size, unique sequences, or enrichment and to search for particular clusters containing a specific, possibly gaped, sequence motif. Finally, a global overview of the distribution of cluster sizes (cluster size vs. their frequency of occurrence) for all selection rounds is also provided.

A large number of additional features, such as the analysis of mutants within each cluster

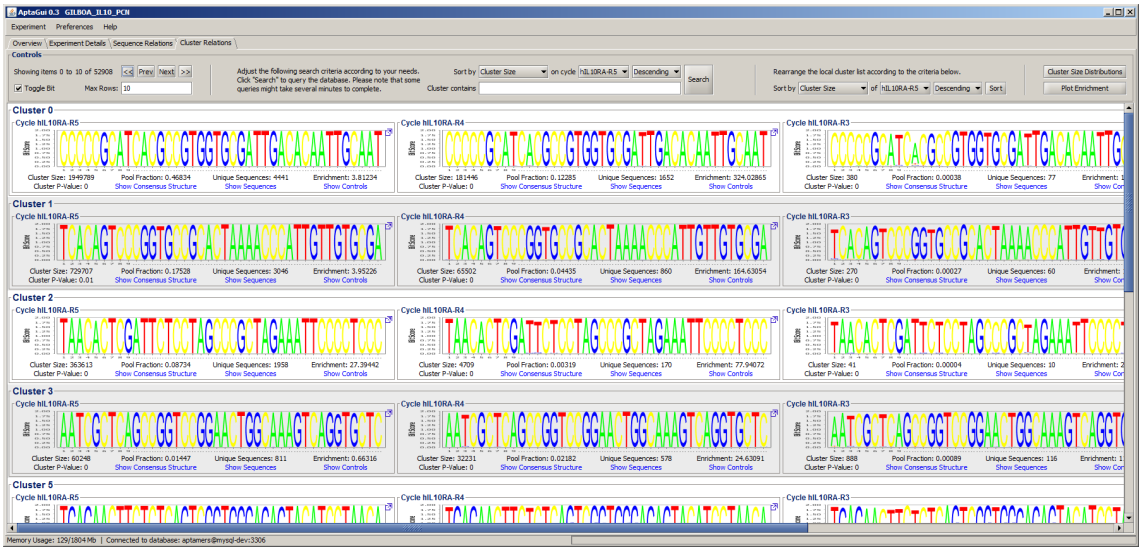


Figure 5.4: Screen capture of AptaGUI showing the *Cluster Relations* tab on the example of our IL-10 selection utilized in Section 4.3.

via AptaMUT [4], and the ability to export all aptamer sequences from the different selection cycles in FASTQ format, are described in the user manual available online.

5.2 AptaREST: Secure and Global Database Communication

As previously described, AptaGUI retrieves and visualizes data stored in a MySQL database for efficient querying and combining different data sources. At the same time, the ability of realizing secure data communication between local and remote systems is becoming increasingly important, and in many research institutions, mandatory in today's globalized IT infrastructure. This is especially crucial for systems such as AptaCLUSTER in combination with AptaGUI, as the computer systems running the former and the MySQL database might not be physically located in the same network as the machine on which AptaGUI is installed. Similarly, if data is shared, e.g. with collaborators, it is equally important to provide access to the local database without compromising security.

In order to allow the different components presented in this work to communicate securely across any platform and network, we additionally developed a RESTful interface called AptaREST. AptaREST can be installed on a Tomcat Server running in an IT-departments DMZ (demilitarized zone) which manages communication between the database and the end user (AptaGUI), hence effectively shielding the database from direct exposure to the outside (see Figure 5.5). AptaREST is implemented in JAVA and based on the Jersey libraries.

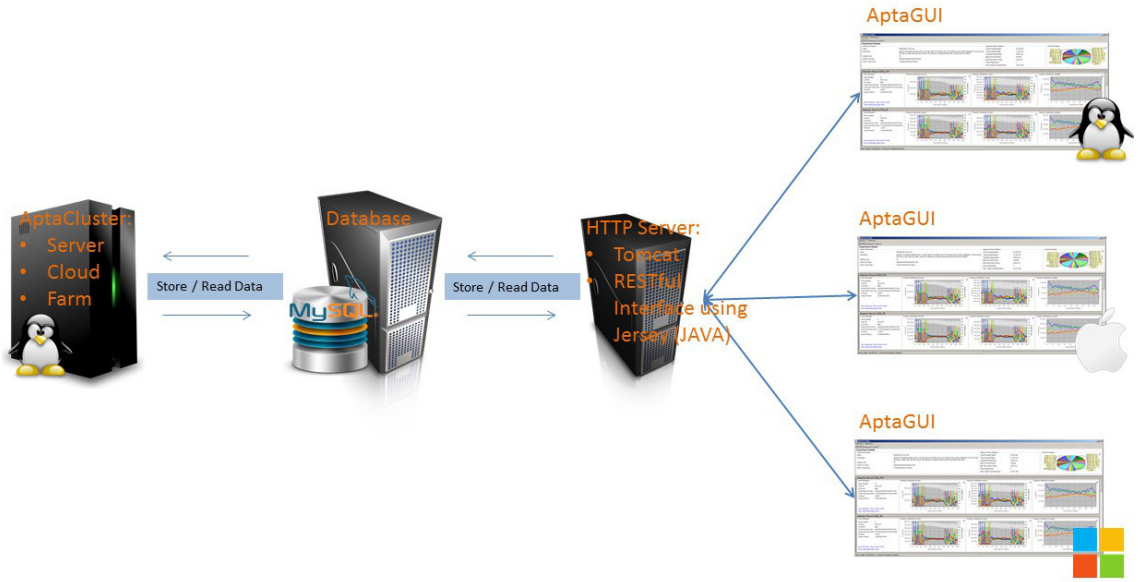


Figure 5.5: Schematic representation of utilizing AptaREST in a typical research oriented IT infrastructure.

5.3 Conclusion

AptaGUI provides, to the best of our knowledge, the first graphical, multi-user, and platform independent user interface for navigating high throughput sequencing data from HT-SELEX experiments and facilitates the identification and analysis of aptamers through its interactive capabilities. At the same time, secure data communication is assured via our AptaREST interface.

6

Conclusion and Outlook

> *If I finish a book a week, I will read only a few thousand books in my lifetime, about a tenth of a percent of the contents of the greatest libraries of our time. The trick is to know which books to read.*

Carl Sagan, *Cosmos*, 1980

The field of aptamer research in combination with the traditional SELEX protocol has until recently taken a purely experimental approach to the discovery of species with the desired properties depending on the intended application. After a target-dependent number of selection cycles, a limited quantity of highly enriched aptamers were cloned and and tested for their binding properties. This procedure frequently resulted in the selection of aptamers with lower than desired target affinity due to the large amount of parameters and selection conditions, ranging from experimental equipment, over temperature, salt concentration, and incubation time, to the washing procedure, amplification techniques, and ultimately the target itself, the SELEX protocol offers. Because of the black-box approach, very little techniques existed for monitoring the actual success of the process throughout the selection cycles .

The introduction, and subsequent coupling, of affordable next-generation sequencing technologies with Systematic Evolution of Ligands by Exponential Enrichment revolutionized the selection process by enabling an unprecedented resolution into every stage of the protocol. The full power of this novel protocol known as HT-SELEX however can only be leveraged in conjunction with suitable computational tools to guide aptamer discovery. As the same time, initial *in silico* approaches typically consisted of ad-hoc counting techniques based on the equally naive assumption that the multiplicity of a species in a pool directly correlates with its affinity to the target. While this assumption might hold true in a very limited number of scenarios such as transcription factor SELEX and selections against other targets evolutionarily adapted to bind (ribo)nucleic acids, artifacts such as non-specific binding, contaminants in the initial pool, and PCR biases significantly skew the landscape of the individual aptamer pools.

True appreciation of the sequencing data produced by modern HT-SELEX experiments is

therefore only possible with the development of innovative mathematical models which take the extensive variability of the protocol into account. The implementation of these models into computational tools is additionally challenged by the immense size of the data, ranging from 2-50 million individual sequences per cycle, and must put additional emphasis onto the scalability of the approach in order to remain computationally tractable.

With the exception of specialized *in silico* techniques for the analysis of transcription factor SELEX [124], the work presented in this dissertation constitutes, to the best of our knowledge, the first attempt of comprehensively bridging the gap between the experimental realm of aptamer discovery and the theory-driven, computational power of modern data processing. Our AptaTOOLS suite therefore engages at the earliest possible stage and provides an efficient, robust, and scalable preprocessing routine for raw aptamer sequencing data via our AptaPLEX software. AptaPLEX is capable of extracting aptamers and demultiplexing jointly sequenced selection cycles while filtering out low quality reads, providing a de-facto standard for downstream processing. Next, we offer a wide array of analysis pipelines for aptamer family identification and tracing (AptaCLUSTER), aptamer refinement by means of favorable mutation recognition (AptaMUT), sequence-structure motif identification (AptaTRACE), and *in silico* SELEX approaches for the study of the protocol itself in a controlled environment (AptaSIM). In recognition of the vast amount of information produced by these methods we developed AptaSIM, a fully interactive, graphical user interface able to visualize many aspects of our aptamer related research and provide the wet-lab scientist with meaningful information regarding the quality, enrichment properties, and candidate sequences of the selection. Naturally, a number of open questions and possibilities of improvement for the existing methods remain despite these advances.

Until now, clustering of aptamers related to each other by sequence was realized using classical hierarchical clustering techniques in conjunction with appropriate similarity functions based on the number of mismatches between two species or more advanced methods such the edit distance. While these approaches are suitable to cluster data sizes from technologies such as Illumina's MiSeq sequencer, next-generation HiSeq systems, producing at least 10-fold the throughput, make such naive clustering techniques computationally untrackable. AptaCLUSTER provides a solution to this problem by closely approximating the exact solution of hierarchical clustering approaches on modern desktop hardware. The key innovation combines the concept of locality sensitive hashing and the fact that, at least in theory, the length of each aptamer is well defined by its randomized region size and flanking primer regions. The considerable speedup however comes at the cost of certain limitations within AptaCLUSTER. While our technique has proven effective for the identification of aptamer families and tracing their behavior throughout the selection, it cannot cope with aptamers containing insertions or deletions introduced into the selection due to error-prone polymerase. Extending our method to be sensitive towards indels therefore remains a challenge yet to be solved. This is important because these species are often of equal interest to the experimentalist as, analogous to aptamers containing nucleotide mutations, they might exhibit improved binding affinities or other desired properties towards the target.

Until now, and despite its potential, the mutational landscape of aptamer pools has typically been regarded as a nuance and has largely been ignored in traditional *in silico* SELEX analy-

sis. Two main factors contributed to this development. First, ad-hoc methods were largely based on simple counting techniques and focused their attention on a small subset of high-frequency species, true to the assumption that these would correspond to the best binders. In contrast, the introduction of mutants into the selection can occur at any SELEX round and, compared to already enriched species, these are only present in a minute initial quantity. Second, the realization of cycle-to-cycle enrichment as a superior predictor of binding affinity had not yet been fully established. With AptaMUT, we introduced a first mathematical approach which not only takes advantage of cycle-to-cycle enrichment but directly incorporates the experimental design of the SELEX protocol itself into the model. This allowed us to discriminate between mutants with improved binding affinity compared to their parent sequence and between those species whose mutation deteriorates (or does no affect) the interaction with the target. Notably, our approach is currently based on primary structure properties only and we hypothesize that extending AptaMUT to include the effect of these mutations onto the secondary structure might further increase its predictive power. This is of relevance as it is well known that aptamer-target binding is a function of both sequence and structural features. Understanding the mutational landscape on both levels could therefore significantly aid the discovery of the conformational characteristics that facilitate binding to the target.

The inclusion of secondary structure information into the model driving AptaSIM is also expected to improve the accuracy of the simulation. While our simulations allow, among many other parameters, for a fine-grained control of error-prone amplification, any mutations introduced into the pool currently get assigned the affinity value of their parent sequence. As a consequence, any structural changes that might arise due to these mutations, and can potentially influence the binding affinity to the target, are currently not taken into account. The inclusion of this feature was mainly limited by the high computational burden of predicting the secondary structure of millions of sequences. In future releases however, and as a result of recent advances in more efficient secondary structure prediction algorithms, we plan to incorporate this feature into our approach. Specifically, we plan to extend the simulated amplification with a probabilistic approach which, for each mutant, predicts its secondary structure and assesses the degree of change to the parent aptamer through an appropriate similarity function. Small structural variations then contribute an at random either positive or negative small value to the binding affinity, whereas large changes in secondary structure are assumed to always be detrimental and more pronounced. By leveraging the insights on mutation rate and binding affinity obtained via AptaMUT, an appropriate parameterization of this method can be achieved. A complete simulation of an experiment that includes structural information in its model for aptamer-target binding therefore has the potential to elucidate properties of the SELEX protocol which are currently inaccessible due to the high cost of performing these *in vitro* and could lead to optimized selections in terms of aptamer quality, time, and cost.

The importance and relation of sequence and structure of aptamers to their corresponding binding affinity cannot be understated. Specifically, the biomolecular recognition mechanism by which individual species interact with the target is typically localized to a precise region of the aptamers exhibiting structural and biochemical properties which are complementary to the surface of the target molecule. Depending on the complexity and size of the target, a number of binding moieties can exist on its surface, leading to the selection of multiple

affine aptamer families. These families either compete for the same apatope or interact with a distinct region of the target. Notably, the selection of numerous species belonging to either the same or to distinct aptamer families which share conserved sequence-structure regions responsible for target-affinity naturally allows to capture these features as binding motifs. In the scope of HT-SELEX, we have shown that sequence-structure motifs are highly suitable for providing crucial information that is not limited to understanding aptamer features which facilitate target interaction alone. In addition, by tracing the evolution of these motifs throughout the selection, they enable the elucidation of selection pressures in unprecedented resolution and clarity.

Tracing motifs throughout consecutive rounds allowed to uncover the possibility of significantly reducing the number of cost-intensive selection cycles of a SELEX experiment in favor of more economical ultra-deep sequencing of the pools in combination with AptaTRACE-based *in silico* analysis. This is feasible in part because AptaTRACE does not rely on quantitative information such as sequence frequency or cycle-to-cycle enrichment. Instead, it directly measures the convergence of secondary structure features towards a common substructure shared by multiple aptamers throughout the selection and compares these to a background model consisting of low frequency aptamers assumed not to bind the target. The assumption that these regions indeed correspond to the binding regions was validated experimentally.

Currently, the approaches presented within this work focus on utilizing the information from positive selection cycles only. However, in an effort to increase the specificity of aptamer candidates, most experimental setups include a series of counter selections (negative selections) after some (or all) selection cycles. A negative selection round entails incubating the aptamer pool with a molecule that is related, but not identical to the target. In the case of protein based SELEX, these typically correspond to homologs from the same protein family whereas in CELL-SELEX, cells which to not express the target on their surface are used. In contrast to positive selections, at the end of each incubation step, only the species which did not bind the negative target are recovered, amplified and used as input for the next cycle. Notably, the properties and utility of counter-selection data is still largely unexplored mainly because experimentalists have not considered these pools as valuable aspects of the selection. However, inclusion of such data into AptaCLUSTER for instance might provide researchers with additional insights into the target-specificity of the aptamer families that have been selected for. In such an application one would expect a target-specific aptamer family to decrease in cardinality when comparing positive and negative selections from the same round. In contrast, background binders are not expected to exhibit this behavior. Information from negative selections could also benefit the model behind AptaTRACE in order to provide more accurate background data. Currently, AptaTRACE leverages low-count species in the sequenced pools to construct a null model that is used to compute the statistical significance of potential motifs. Determining the threshold at which a sequence is considered background (non-binders) or foreground (binders) however is not always a trivial task, especially during the early stages of the selection. Substituting the usage of low-count sequences with data from the negative selections therefore has the potential to circumvent this issue.

In summary, the approaches presented throughout this dissertation have contributed significantly to the advancement of aptamer research in general and are being utilized exten-

sively, not only by our collaborators, but by the SELEX community as a whole. Our software is currently available as individual, but compatible tools which can be combined into larger *in silico* pipelines. However, in recognition of the specific requirements of this mainly wet-lab based community, we plan to consolidate our approaches into a single, purely graphical, and platform independent software package in the near future. By doing so, we hope to achieve a lasting impact and valued contribution to the amazing field of aptamer discovery.

Bibliography

[1] Jan Hoinka, Phuong Dao, Yijie Wang, Mayumi Takahashi, Jiehua Zhou, Fabrizio Costa, John Rossi, John Burnett, Rolf Backofen, and Teresa M Przytycka. AptaTRACE Elucidates Aptamer Sequence-Structure Motifs in HT-SELEX Experiments. *Cell Systems*, (Manuscript Accepted for Publication), 2016.

[2] Jan Hoinka and Teresa Przytycka. AptaPLEX – A dedicated, multithreaded demultiplexer for HT-SELEX data. *Methods*, 2016.

[3] Jan Hoinka, Phuong Dao, and Teresa M Przytycka. AptaGUI—A Graphical User Interface for the Efficient Analysis of HT-SELEX Data. *Molecular Therapy—Nucleic Acids*, 4(10):e257, 2015.

[4] Jan Hoinka, Alexey Berezhnoy, Phuong Dao, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Research*, 43(10):5699–5707, 2015.

[5] Jan Hoinka, Phuong Dao, Yijie Wang, Mayumi Takahashi, Jiehua Zhou, Fabrizio Costa, John Rossi, John Burnett, Rolf Backofen, and Teresa M Przytycka. AptaTRACE: Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments. *Research in Computational Molecular Biology : Annual International Conference, RECOMB: Proceedings*, 2016.

[6] Agata Levay, Randall Brenneman, Jan Hoinka, David Sant, Marco Cardone, Giorgio Trinchieri, Teresa M Przytycka, and Alexey Berezhnoy. Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment. *Nucleic Acids Research*, 43(12), 2015.

[7] Jan Hoinka, Alexey Berezhnoy, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. *Research in Computational Molecular Biology : Annual International Conference, RECOMB: Proceedings*, 8394:115–128, 2014.

[8] Jan Hoinka, Elena Zotenko, Adam Friedman, Zuben E. Sauna, and Teresa M. Przytycka. Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics*, 28(12), 2012.

[9] Matthijs M Jore, Magnus Lundgren, Esther van Duijn, Jelle B Bultema, Edze R Westra, Sakharam P Waghmare, Blake Wiedenheft, Umit Pul, Reinhild Wurm, Rolf Wagner, Marieke R Beijer, Arjan Barendregt, Kaihong Zhou, Ambrosius P L Snijders, Mark J Dickman, Jennifer A Doudna, Egbert J Boekema, Albert J R Heck, John van der Oost, and Stan J J Brouns. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular Biology*, 18(5):529–536, 2011.

[10] John van der Oost, Edze R Westra, Ryan N Jackson, and Blake Wiedenheft. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature reviews. Microbiology*, 12(7):479–92, 2014.

[11] G Canziani, W Zhang, D Cines, A Rux, S Willis, G Cohen, R Eisenberg, and I Chaiken. Exploring biomolecular recognition using optical biosensors. *Methods*, 19(2):253–69, 1999.

[12] Robert E Babine and Steven L Bender. Molecular Recognition of Protein − Ligand Complexes : Applications to Drug Design. *Chemical Reviews*, 97(5):1359–1472, 1997.

[13] Philip J Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews. Drug discovery*, 6(3):211–9, 2007.

[14] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe. Computational methods in drug discovery. *Pharmacological reviews*, 66(1):334–95, 2014.

[15] C D Floyd, C Leblanc, and M Whittaker. Combinatorial chemistry as a tool for drug discovery. *Progress in medicinal chemistry*, 36:91–168, 1999.

[16] B. E H Maden and J. M X Hughes. Eukaryotic ribosomal RNA: The recent excitement in the nucleotide modification problem, 1997.

[17] Tsanev RG Hadjiolov AA, Venkov PV. Ribonucleic acids fractionation by density-gradient centrifugation and by agar gel electrophoresis: a comparison. *Anal Biochem*, 17(2):263–267, 1966.

[18] A Gregory Matera and Zefeng Wang. A day in the life of the spliceosome. *Nature reviews. Molecular cell biology*, 15(2):108–21, 2014.

[19] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

[20] Bruce Wightman, Ilho Ha, and Gary Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862, 1993.

[21] Lin He and Gregory J Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7):522–531, 2004.

[22] S M Hammond, E Bernstein, D Beach, and G J Hannon. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404(6775):293–296, 2000.

[23] Alexander Serganov and Evgeny Nudler. A decade of riboswitches, 2013.

[24] A D Ellington and J W Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818–22, 1990.

[25] Yeon Seok Kim and Man Bock Gu. Advances in aptamer screening and small molecule aptasensors. *Advances in Biochemical Engineering/Biotechnology*, 140:29–67, 2014.

[26] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M. Vaquerizas, Jian Yan, Mikko J. Sillanpää, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R. Hughes, Nicholas M. Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, 2010.

[27] Alexey Berezhnoy, C. Andrew Stewart, James O. Mcnamara II, William Thiel, Paloma Giangrande, Giorgio Trinchieri, and Eli Gilboa. Isolation and Optimization of Murine IL-10 Receptor Blocking Oligonucleotide Aptamers Using High-throughput Sequencing. *Molecular Therapy*, 20(6):1242–1250, 2012.

[28] Jennifer M. Binning, Tianjiao Wang, Priya Luthra, Reed S. Shabman, Dominika M. Borek, Gai Liu, Wei Xu, Daisy W. Leung, Christopher F. Basler, and Gaya K. Amarasinghe. Development of RNA aptamers targeting Ebola virus VP35. *Biochemistry*, 52(47):8406–8419, 2013.

[29] Dion a Daniels, Hang Chen, Brian J Hicke, Kristine M Swiderek, and Larry Gold. A tenascin-C aptamer identified by tumor cell SELEX: systematic evolution of ligands by exponential enrichment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15416–15421, 2003.

[30] Kevin N Morris, Kirk B Jensen, Carol M Julin, Michael Weil, and Larry Gold. High affinity ligands from in vitro selection : Complex targets. *RNA*, 95(March):2902–2907, 1998.

[31] Hui Shi, Wensi Cui, Xiaoxiao He, Qiuping Guo, Kemin Wang, Xiaosheng Ye, and Jinlu Tang. Whole Cell-SELEX Aptamers for Highly Specific Fluorescence Molecular Imaging of Carcinomas In Vivo. *PLoS ONE*, 8(8), 2013.

[32] Ran Zichel, Wanida Chearwae, Gouri Shankar Pandey, Basil Golding, and Zuben E. Sauna. Aptamers as a sensitive tool to detect subtle modifications in therapeutic proteins. *PLoS ONE*, 7(2), 2012.

[33] Dongxi Xiang, Sarah Shigdar, Greg Qiao, Tao Wang, Abbas Z. Kouzani, Shu Feng Zhou, Lingxue Kong, Yong Li, Chunwen Pu, and Wei Duan. Nucleic acid aptamer-guided cancer therapeutics and diagnostics: The next generation of cancer medicine, 2015.

[34] Kathleen A. Tobin. Macugen treatment for wet age-related macular degeneration. *Insight - Journal of the American Society of Ophthalmic Registered Nurses*, 31(1):11–14, 2006.

[35] Khalid K Alam, Jonathan L Chang, and Donald H Burke. FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Molecular Therapy—Nucleic Acids*, 4(August 2014):e230, 2015.

[36] Gillian V. Kupakuwana, James E. Crill, Mark P. McPike, and Philip N. Borer. Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. *PLoS ONE*, 6(5), 2011.

[37] C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, N.Y.)*, 249(4968):505–510, 1990.

[38] Brian R Baker, Rebecca Y Lai, McCall S Wood, Elaine H Doctor, Alan J Heeger, and Kevin W Plaxco. An electronic, aptamer-based small-molecule sensor for the rapid, label-free detection of cocaine in adulterated samples and biological fluids. *Journal of the American Chemical Society*, 128(10):3138–9, 3 2006.

[39] Xiaolei Zuo, Yi Xiao, and Kevin W Plaxco. High specificity, electrochemical sandwich assays based on single aptamer sequences and suitable for the direct detection of small-molecule targets in blood and other complex matrices. *Journal of the American Chemical Society*, 131(20):6944–5, 5 2009.

[40] Todd R Riley, Matthew Slattery, Namiko Abe, Chaitanya Rastogi, Richard Mann, and Harmen Bussemaker. SELEX-seq, a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes.

[41] Akira Ishihama, Ayako Kori, Etsuko Koshio, Kayoko Yamada, Hiroto Maeda, Tomohiro Shimada, Hideki Makinoshima, Akira Iwata, and Nobuyuki Fujitac. Intracellular concentrations of 65 species of transcription factors with known regulatory functions in Escherichia coli. *Journal of Bacteriology*, 196(15):2718–2727, 2014.

[42] Giomar Rivera-Cancel, Laura B. Motta-Mena, and Kevin H. Gardner. Identification of natural and artificial DNA substrates for light-activated LOV-HTH transcription factor EL222. *Biochemistry*, 51(50):10024–10034, 2012.

[43] Jee Woong Park, Su Jin Lee, Eun Jin Choi, Jaejo Kim, Jae Young Song, and Man Bock Gu. An ultra-sensitive detection of a whole virus using dual aptamers developed by immobilization-free screening. *Biosensors and Bioelectronics*, 51:324–329, 2014.

[44] Laura Cerchia, Jörg Hamm, Domenico Libri, Bertrand Tavitian, and Vittorio De Franciscis. Nucleic acid aptamers in cancer medicine, 2002.

[45] Nicole Lambert, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54(5):887–900, 2014.

[46] L C Bock, L C Griffin, J a Latham, E H Vermaas, and J J Toole. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature*, 355:564–566, 1992.

[47] Michael U. Musheev and Sergey N. Krylov. Selection of aptamers by systematic evolution of ligands by exponential enrichment: Addressing the polymerase chain reaction issue. *Analytica Chimica Acta*, 564(1):91–96, 2006.

[48] David E Draper. A guide to ions and RNA structure. *RNA (New York, N.Y.)*, 10(3):335–43, 2004.

[49] Randall K Saiki, David H Gelfand, Susanne Stoffel, Stephen J Scharf, Russell Higuchi, Glenn T Horn, Kary B Mullis, and Henry A Erlich. Primer-directed enzymatic amplification of DNA with a thermostable polymerase. *Science*, 239(4839):487–491, 1988.

[50] Joshua a Bittker, Brian V Le, and David R Liu. Nucleic acid evolution and minimization by nonhomologous random recombination. *Nature biotechnology*, 20(10):1024–1029, 2002.

[51] Silvia G. Acinas, Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj, and Martin F. Polz. PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, 71(12):8966–8969, 2005.

[52] Jonghoon Kang, Myung Soog Lee, and David G Gorenstein. The enhancement of PCR amplification of a random sequence DNA library by DMSO and betaine: Application to in vitro combinatorial selection of aptamers. *Journal of Biochemical and Biophysical Methods*, 64(2):147–151, 2005.

[53] Roman Yufa, Svetlana M. Krylova, Christine Bruce, Eleanor A. Bagg, Christopher J. Schofield, and Sergey N. Krylov. Emulsion PCR significantly improves nonequilibrium capillary electrophoresis of equilibrium mixtures-based aptamer selection: Allowing for efficient and rapid selection of aptamer to unmodified ABH2 protein. *Analytical Chemistry*, 87(2):1411–1419, 2015.

[54] P. J. Sykes, S. H. Neoh, M. J. Brisco, E. Hughes, J. Condon, and A. A. Morley. Quantitation of targets for PCR by use of limiting dilution. *BioTechniques*, 13(3):444–449, 1992.

[55] Milan N Stojanovic, Paloma de Prada, and Donald W Landry. Fluorescent Sensors Based on Aptamer Self-Assembly. *J. Am. Chem. Soc.*, 122(10):11547–11548, 2000.

[56] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[57] Kwame Sefah, Dihua Shangguan, Xiangling Xiong, Meghan B O'Donoghue, and Weihong Tan. Development of DNA aptamers using Cell-SELEX. *Nature protocols*, 5(6):1169–1185, 2010.

[58] Matthias Dodt, Johannes Roehr, Rina Ahmed, and Christoph Dieterich. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, 1:895–905, 2012.

[59] Jingtao Liu. deindexer - an easy deindex tool for illumina multiple barcodes sequencing, 2015.

[60] Brant Faircloth. Splitaake - demultiplex massively parallel sequencing data, 2013.

[61] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[62] Gabriel Renaud, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. Sequence analysis deML : robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(October 2014):770–772, 2015.

[63] Haisi Yi, Zhe Li, Tao Li, and Jindong Zhao. Bayexer: an accurate and fast Bayesian Demultiplexer for Illumina sequences. *Bioinformatics (Oxford, England)*, 31(24):4000–4002, 2015.

[64] William H. Thiel and Paloma H. Giangrande. Analyzing HT-SELEX data with the Galaxy Project tools - a web based bioinformatics platform for biomedical research. *Methods*, 2015.

[65] L. Dagum and R. Menon. OpenMP: an industry standard API for shared-memory programming. *IEEE Computational Science and Engineering*, 5(1):46–55, 1998.

[66] From Wikipedia and Boost Software License. Boost C ++ Libraries, 2010.

[67] Björn Karlsson. *Beyond the C++ Standard Library: An Introduction to Boost*. 2005.

[68] M. Cho, Y. Xiao, J. Nie, R. Stewart, A. T. Csordas, S. S. Oh, J. A. Thomson, and H. T. Soh. Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 107(35):15373–15378, 2010.

[69] William H. Thiel, Thomas Bair, Andrew S. Peek, Xiuying Liu, Justin Dassie, Katie R. Stockdale, Mark A. Behlke, Francis J. Miller, and Paloma H. Giangrande. Rapid Identification of Cell-Specific, Internalizing RNA Aptamers with Bioinformatics Analyses of a Cell-Based Aptamer Selection. *PLoS ONE*, 7(9), 2012.

[70] Philip E Johnson and Logan W Donaldson. RNA recognition by the Vts1p SAM domain. *Nature structural & molecular biology*, 13(2):177–178, 2006.

[71] Christian Schudoma, Patrick May, Viktoria Nikiforova, and Dirk Walther. Sequence-structure relationships in RNA loops: Establishing the basis for loop homology modeling. *Nucleic Acids Research*, 38(3):970–980, 2009.

[72] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

[73] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34(17), 2006.

[74] Timothy L. Bailey. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[75] Hilal Kazan, Debashish Ray, Esther T. Chan, Timothy R. Hughes, and Quaid Morris. RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Computational Biology*, 6(7):28, 2010.

[76] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity Search in High Dimensions via Hashing. *Search*, 99(1):518–529, 1999.

[77] KN Couper, DG Blount, and EM. Riley. IL-10: the master regulator of immunity to infection. *Journal of immunology*, 180(9):5771–5777, 2008.

[78] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 459–468, 2006.

[79] Kuan Yang and Liqing Zhang. Performance comparison between k -tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Research*, 36(5), 2008.

[80] Gerald F. Joyce. Amplification, mutation and selection of catalytic RNA. *Gene*, 82(1):83–87, 1989.

[81] D Nieuwlandt, M Wecker, and L Gold. In vitro selection of RNA ligands to substance P. *Biochemistry*, 34(16):5651–5659, 1995.

[82] Regina Stoltenburg, Christine Reinemann, and Beate Strehlitz. SELEX-A (r)evolutionary method to generate high-affinity nucleic acid ligands, 2007.

[83] Bruce E. Eaton, Larry Gold, Brian J. Hicke, Neboj??a Janji??, Fiona M. Jucker, David P. Sebesta, Theodore M. Tarasow, Michael C. Willis, and Dominic A. Zichi. Post-SELEX combinatorial optimization of aptamers. In *Bioorganic and Medicinal Chemistry*, volume 5, pages 1087–1096, 1997.

[84] Albert-László Barabási and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):11, 1999.

[85] Tatjana Schütze, Barbara Wilhelm, Nicole Greiner, Hannsjörg Braun, Franziska Peter, Mario Mörl, Volker A. Erdmann, Hans Lehrach, Zoltán Konthur, Marcus Menger, Peter F. Arndt, and Jörn Glökler. Probing the SELEX process with next-generation sequencing. *PLoS ONE*, 6(12), 2011.

[86] J A Jaeger, D H Turner, and M Zuker. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci U S A*, 86(20):7706–7710, 1989.

[87] D L Swofford. PAUP* phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. *Sinauer Associates*, 2002.

[88] Abdullah Ozer, John M Pagano, and John T Lis. New Technologies Provide Quantum Changes in the Scale, Speed, and Success of SELEX Methods and Aptamer Characterization. *Molecular therapy. Nucleic acids*, 3(August):e183, 2014.

[89] Kristina W Thiel and Paloma H Giangrande. Intracellular delivery of RNA-based therapeutics using aptamers. *Therapeutic Delivery*, 1(6):849–861, 2010.

[90] Jiehua Zhou and John J Rossi. Aptamer-targeted cell-specific RNA interference. *Silence*, 1(1):4, 2010.

[91] M Tompa, N Li, T L Bailey, G M Church, B De Moor, E Eskin, A V Favorov, M C Frith, Y T Fu, W J Kent, V J Makeev, A A Mironov, W S Noble, G Pavesi, G Pesole, M Regnier, N Simonis, S Sinha, G Thijs, J van Helden, M Vandenbogaert, Z P Weng, C Workman, C Ye, and Z Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, 2005.

[92] Modan K Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. *BMC bioinformatics*, 8 Suppl 7:S21, 2007.

[93] Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–237, 2013.

[94] C E Lawrence and a a Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990.

[95] Saurabh Sinha. PhyME: a software tool for finding motifs in sets of orthologous sequences. *Methods in molecular biology (Clifton, N.J.)*, 395:309–318, 2007.

[96] John E. Reid and Lorenz Wernisch. STEME: Efficient EM to find motifs in large data sets. *Nucleic Acids Research*, 39(18), 2011.

[97] Timothy L. Bailey and Charles Elkan. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Machine Learning*, 21(1):51–80, 1995.

[98] C E Lawrence, S F Altschul, M S Boguski, J S Liu, a F Neuwald, and J C Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.)*, 262(5131):208–214, 1993.

[99] F P Roth, J D Hughes, P W Estep, and G M Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

[100] G Thijs, K Marchal, M Lescot, S Rombauts, B De Moor, P Rouzé, and Y Moreau. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of computational biology : a journal of computational molecular cell biology*, 9(2):447–464, 2002.

[101] X Liu, D L Brutlag, and J S Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–138, 2001.

[102] Giulio Pavesi, Paolo Mereghetti, Giancarlo Mauri, and Graziano Pesole. Weeder web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, 32(WEB SERVER ISS.), 2004.

[103] Saurabh Sinha and Martin Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation, 2002.

[104] X S Liu, D L Brutlag, and J S Liu. An algorithm for finding protein DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, 20(8):835–9, 2002.

[105] Chaim Linhart, Yonit Halperin, and Ron Shamir. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Research*, 18(7):1180–1189, 2008.

[106] Phaedra Agius, Aaron Arvey, William Chang, William Stafford Noble, and Christina Leslie. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Computational Biology*, 6(9), 2010.

[107] M F Sagot. Spelling approximate repeated or common motifs using a suffix tree. *Lecture notes in computer science*, pages 374–390, 2009.

[108] G Pavesi, G Mauri, and G Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics (Oxford, England)*, 17 Suppl 1:207–14, 2001.

[109] Limor Leibovich, Inbal Paz, Zohar Yakhini, and Yael Mandel-Gutfreund. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic acids research*, 41(Web Server issue), 2013.

[110] Barrett C. Foat, Alexandre V. Morozov, and Harmen J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. In *Bioinformatics*, volume 22, 2006.

[111] Amos Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research*, 16(8):962–972, 2006.

[112] I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov, and V. J. Makeev. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, 26(20):2622–2623, 2010.

[113] Manu Setty and Christina S. Leslie. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLOS Computational Biology*, 11(5):e1004271, 2015.

[114] Matthew T Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, Harmen J Bussemaker, Quaid D Morris, Martha L Bulyk, Gustavo Stolovitzky, and Timothy R Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–34, 2013.

[115] Yue Zhao, David Granas, and Gary D. Stormo. Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5(12), 2009.

[116] Yaron Orenstein and Ron Shamir. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, 42(8), 2014.

[117] J Caroli, C Taccioli, A De La Fuente, P Serafini, and S Bicciato. APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. *Bioinformatics*, 32(2):btv545, 2015.

[118] Ye Ding, Chi Yu Chan, and Charles E Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA (New York, N.Y.)*, 11(8):1157–66, 2005.

[119] Panayiotis V. Benos, Alan S. Lapedes, and Gary D. Stormo. Probabilistic code for DNA

recognition by proteins of the EGR family. *Journal of Molecular Biology*, 323(4):701–727, 2002.

[120] Panayiotis V Benos, Martha L Bulyk, and Gary D Stormo. Additivity in protein - DNA interactions: how good an approximation is it? *Nucleic Acids Research*, 30(20):4442–51, 2002.

[121] Teresa M Przytycka and David Levens. Shapely DNA attracts the right partner. *Proceedings of the National Academy of Sciences of the United States of America*, 112(15):4516–7, 4 2015.

[122] Zsuzsanna Sükösd, Bjarne Knudsen, Jorgen Kjems, and Christian N S Pedersen. PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, 28(20):2691–2692, 2012.

[123] Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–1975, 2009.

[124] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotech*, 33(8):831–838, 8 2015.

[125] Irina Gitlin, Jeffrey D. Carbeck, and George M. Whitesides. Why are proteins charged? Networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis, 2006.

[126] A. Yu Grosberg, T. T. Nguyen, and B. I. Shklovskii. Colloquium: The physics of charge inversion in chemical and biological systems, 2002.

[127] Kuangwen Hsieh, B. Scott Ferguson, Michael Eisenstein, Kevin W. Plaxco, and H. Tom Soh. Integrated electrochemical microsystems for genetic detection of pathogens at the point of care. *Accounts of chemical research*, 48(4):911–920, 2015.

A

Appendix: SELEX - Supplementary Materials and Methods

This chapter contains supplementary materials and information regarding the systematic evolution of ligands by exponential enrichment protocol.

A.1 Experimental Details for Partitioning Target-Bound Species during SELEX

In what follows, a detailed description of the experimental protocols commonly used for partitioning the target-bound species from the sequences which do not interact with the target are provided.

Nitrocellulose filter binding is a partitioning technique commonly employed in selections against proteins and leverages the differences in electrostatic charge between proteins, which typically have a net positive charge [125], and DNA/RNA sequences, whose net charge is overall negative [126]. When applying the solution onto the negatively charged nitrocellulose paper, aptamer-bound proteins are retained on the filter, while the negatively charged unbound-species repel the filters surface and pass through. The molecules retained on the filter are consequently recovered by washing these from the cellulose paper using an appropriate buffer.

Bead based partitioning represents another accepted technique in order to separate-target bound species from non-binders. Prior to incubation, the target molecules are immobilized on a magnetic solid support (beads) which consequently facilitates the partitioning process when applying an electric field onto the solution. After recovering the beads, the aptamer-target complexes are purified and removed from the solid support.

Electrophoretic filtering, especially capillary electrophoresis, is also a prevalent technique in which aptamer complexes and non-interacting species are embedded into a buffer onto which a high voltage current is applied. Depending on the size and charge of the particles,

unbound aptamers migrate to the opposite pole at a faster rate than target-bound complexes yielding the desired separation of the two.

Phosphate-buffered saline and media based washing is typically used for complex targets such as entire cells (see Section 2.7 for details). Here, cells are immobilized on a support platform and unbound species are removed by either washing the cells with PBS or the media in which they are grown in.

A number of additional techniques, including microfluidic separation, microarray-based methods, and microscopic approaches are also part of the repertoire for partitioning the pool and vary in popularity depending on the complexity of the target. These however, are more specialized methods when applied to SELEX and we refer the reader to the excellent review by Ozer *et al.* [88] for details.

A.2 Variations of the SELEX Protocol

In what follows, we provide a brief overview of the currently those SELEX protocols which require a high degree of adaptation as compared to the general approach described above.

Small-Molecule SELEX is mainly concerned with the selection of aptamers binding small organic or non-organic particles which consequently form the central component in sensoring technologies such as drug detection kits in forensic science [55] or for monitoring metabolites in organic systems [127], among others.

Transcription Factor SELEX (TF-SELEX) is one of the most widely known and effective selection protocols for the determination of binding site motifs. Here, either a synthesized initial library, or a library consisting of shotgunned genomic sequences is incubated with the target. Notably, the specific property of transcription factors of being evolutionarily adapted to bind (ribo)nucleid acids results in a rapid selection of target affine binders, typically requiring a very low number of selection cycles [36, 45]. Computational analysis of these binders then enables the generation of accurate binding motifs for the intended target.

Cell SELEX, is a novel and rapidly developing technique and allows for the selection of aptamers which bind a cell type with high specificity, typically by interacting with a surface protein uniquely expressed in that particular cell lineage. By conjugating these aptamers with additional payloads such as pharmaceuticals, targeted drug delivery systems, e.g. against certain cancer types, can be designed. Specificity of aptamers is commonly achieved by alternating positive selection cycles against the target cell with negative selections against related cell types which do not express the surface protein in question. Due to the size and complexity of these targets, a large number of selection rounds are typically required and the number of non-specific binders and non-binding sequences in the pools is considerable larger as compared to selections against purified proteins or smaller molecules.

B

Appendix: Analysis - Supplementary Materials and Methods

This chapter contains supplementary materials and methods regarding the methods described in Chapter 4.

B.1 AptaSIM - List of Full Parameters

The list of options for AptaSIM detailing the name of the corresponding parameters, their default values, as well as a description of how these influence the selection simulation are shown in Table B.1 below.

B.2 Selection Protocol against Interleukin Receptor 10

Here, we describe the experimental details of the selection against Interleukin receptor 10, the data of which was utilized to benchmark our AptaCLUSTER and AptaMUT algorithm.

Selection Details: A DNA template for the selection library was ordered from IDT (Coralville, IA). 1 nM of each N_{40} template `(5'-TCTCGATCTCAGCGAGTCGTCG-`N_{40}`-CCCATCCCTCTTCCTCTCTCCC-3')` and 5' primer `(5'-GGGGGAATTCTAATACGACTCACTATAGGGAGAGAGGAAGAGGGATGGG-3')` were annealed together, extended with Taq polymerase (Life Science), and transcribed in vitro using Durascribe (in-vitro transcription) IVT kit (Illumina). The random R0 RNA was purified by denaturing PAGE and, after preclearing with human IgG-coated (Sigma) beads (GE Healthcare), used for in-vitro selection. 1 nM of R0 RNA was used in a first round of selection to coincubate with 0.3 nM of bead-bound human IL-10RA-Fc fusion protein (Novus Biologicals) in 100 mM NaCl selection buffer. After washes, a recovered bound RNA fraction was reverse transcribed using the cloned AMV RT kit (Life Science). cDNA was amplified by either emulsion or open PCR using Platinum Taq PCR kit (Life Science) as described be-

Table B.1: List of parameters for AptaSIM. Shown are the name of the corresponding parameter, their default value, as well as a description of how this parameters influences the selection simulation.

Parameter	Default Value	Description
hmm_file	none	FASTQ file containing training sequences
hmm_degree	2	Degree of the Markov model
randomized_region_size (r)	40	Length of the aptamers
number_of_sequences (N)	10000000	Number of sequences in the initial pool
number_of_seeds (s)	100	Number of seeds in the initial pool
min_seed_affinity (s_l)	80	Minimal affinity for seed sequences (range: 0-100)
max_sequence_count (c)	10	Maximal count of remaining sequences
max_sequence_affinity (b)	25	Maximal sequence affinity for non-seeds (range: 0-100)
nucleotide_distribution	A:0.25, C:0.25, G:0.25, T:0.25	If no training data is specified, create pool based on this nucleotide distribution
selection_percentage (p_s)	0.20	The percentage of sequences that remain after selection (range: 0-1)
base_mutation_rates	A:0.25, C:0.25, G:0.25, T:0.25	Mutation rates for individual nucleotides
mutation_probability	0.05	Mutation probability during PCR (range: 0-1)
amplification_efficiency	0.995	PCR amplification efficiency (range: 0-1)
number_of_cycles	10	Number of selection cycles to perform

low. The DNA template was used to IVT RNA for the next round. During subsequent rounds, amount of protein was reduced 25% each time, while concentration of NaCl was gradually increased to 150 mM.

Emulsion PCR: cDNA was amplified using Platinum Taq PCR kit with addition of 10% PCRx enhancer solution and following primers: `5'-GGGGGAATTCTAATACGACTCACTATAGGGAGAGAGGAAGAGGGATGGG-3'` and `5'-TCTCGATCTCAGCGAGTCGTCG-3'`. After preparing the master mix PCR reaction solution, it was separated to 100 μL aliquots and each aliquot was mixed with 600 μL ice-cold oil fraction assembled from components supplied with emulsion PCR kit (EURx) according to manufacturer's instructions. Water and oil mixture was emulsified by 5' vortexing at +4C and amplified in standard PCR machine for 25 cycles. Control open PCR reaction was carried with aqueous phase only for 16 cycles.

Preparing Libraries for HTS: After 4 rounds of selection, 3 nM of RNA was prepared for round 5. The RNA was precleared using IgG-coated beads and separated into three identical aliquots. Each aliquot was incubated with either human IL10RA protein, murine IL10RA

protein or human IgG. After standard washes, bound RNA fraction was extracted from beads and reverse transcribed as described previously. A cDNA generated from round 5 bound fractions, as well as RNA recovered from bound fractions at rounds 2, 3 and 4, was amplified by emulsion PCR with two sets of primers as described previously [**?**]. Amplified DNA was purified by 2% agarose gel electrophoresis and sequenced using Illumina's HiSeq 2500 device with 100-cycle paired-end sequencing protocol.

B.3 AptaMUT- Derivation and Conversion Analysis

The diversity of an initial pool in a SELEX experiment is a function of the sequence length. Here, we provide a mathematical estimation of the expected number of aptamers of size n with at least K % similarity. Let $\frac{1}{k},\ k \in \mathbb{N}$ be the threshold for the sequence dissimilarity according to the edit distance. Furthermore, we define n as the length of the aptamer and assume $\frac{n}{k} \in \mathbb{N}$. The expected fraction $F(n,k)$ of aptamers with at most k% dissimilarity can then be calculated the sum over all possible sequences with i variable nucleotides divided by the number of all permutations of sequences of size n:

$$F(n,k) = \sum_{i=1}^{\frac{n}{k}} f(i,n) \tag{B.1}$$

$$\text{where } f(i,n) = \frac{\binom{n}{i}3^i}{4^n} \tag{B.2}$$

Equation B.1 can be approximated as follows. For $i \leq \frac{n}{2}$ we have

$$f(i,n) > 3 * f(i-1,n) \tag{B.3}$$

because

$$f(i,n) = \frac{\binom{n}{i}3^i}{4^n} = \frac{\binom{n}{i-1}3^{i-1} * 3}{4^n} * \frac{n-i+1}{i} \tag{B.4}$$

$$= f(i-1,n) * \underbrace{\left(\frac{n+1}{i} - 1\right) * 3}_{\geq 1 \text{ for } i \leq \frac{1}{2}n} \geq 3 * f(i-1,n) \tag{B.5}$$

Thus, we can approximate an upper bound to (B.1) using the last term of the expansion:

$$F(n,k) \approx f(i,\frac{n}{k}) = \frac{\binom{n}{\frac{n}{k}}3^{\frac{n}{k}}}{4^n} = \left(\frac{3^{\frac{1}{k}}}{4}\right)^n * \frac{n!}{\frac{n}{k}!(n-\frac{n}{k})!} \tag{B.6}$$

Substituting $x!$ with the Stirling approximation $x! \approx \sqrt{2\pi x} * \frac{x}{e}^x$ we get

$$F(n,k) \approx \left(\frac{3^{\frac{1}{k}}}{4}\right)^n * \frac{\sqrt{2\pi n} * \left(\frac{n}{e}\right)^n}{\sqrt{2\pi\frac{n}{k}} * \left(\frac{\frac{n}{k}}{e}\right)^{\frac{n}{k}} * \sqrt{2\pi\left(n-\frac{n}{k}\right)} * \left(\frac{n-\frac{n}{k}}{e}\right)^{n-\frac{n}{k}}} \tag{B.7}$$

$$= \left(\frac{3^{\frac{1}{k}}}{4}\right)^n * \frac{\sqrt{2\pi n} * \left(\frac{n}{e}\right)^n}{\sqrt{2\pi\frac{n}{k}} * \left(\frac{k}{ne}\right)^{n\frac{1}{k}} * \sqrt{2\pi n\left(1-\frac{1}{k}\right)} * \left(\frac{n*(1-\frac{1}{k})}{e}\right)^{n*(1-\frac{1}{k})}} \tag{B.8}$$

$$= \underbrace{\left(\frac{3^{\frac{1}{k}}}{4}\right)^n}_{(A)} * \underbrace{\frac{\sqrt{2\pi n}}{\sqrt{2\pi n\frac{1}{k}} * \sqrt{2\pi n\left(1-\frac{1}{k}\right)}}}_{(B)} * \underbrace{\frac{\left(\frac{n}{e}\right)^n}{\left(\frac{k}{ne}\right)^{n\frac{1}{k}} * \left(\frac{n(1-\frac{1}{k})}{e}\right)^{n(1-\frac{1}{k})}}}_{(C)} \tag{B.9}$$

We can rewrite (C) as follows:

$$\frac{\left(\frac{n}{e}\right)^n}{\left(\frac{k}{ne}\right)^{n\frac{1}{k}} * \left(\frac{n(1-\frac{1}{k})}{e}\right)^{n(1-\frac{1}{k})}} = \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e} * \frac{1}{k}\right)^{n\frac{1}{k}} * \left(\frac{n}{e} * (1-\frac{1}{k})\right)^{n(1-\frac{1}{k})}} \tag{B.10}$$

$$= \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e}\right)^{n\frac{1}{k}} * \left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(\frac{n}{e}\right)^{n(1-\frac{1}{k})} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{B.11}$$

$$= \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e}\right)^{n\frac{1}{k}+n(1-\frac{1}{k})} * \left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{B.12}$$

$$= \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e}\right)^{n(\frac{1}{k}+1-\frac{1}{k})} * \left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{B.13}$$

$$= \frac{1}{\left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{B.14}$$

$$= \left(\frac{1}{\left(\frac{1}{k}\right)^{\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{(1-\frac{1}{k})}}\right)^n \tag{B.15}$$

$$= \left(\frac{1}{\left(\frac{1}{k}\right)^{\frac{1}{k}} * \left(\frac{k-1}{k}\right)^{(1-\frac{1}{k})}}\right)^n \tag{B.16}$$

$$= \left(\frac{1}{\frac{1^{\frac{1}{k}} * (k-1)^{1-\frac{1}{k}}}{k^{\frac{1}{k}} * k^{1-\frac{1}{k}}}}\right)^n \tag{B.17}$$

$$= \left(\frac{1}{\frac{1^{\frac{1}{k}} * (k-1)^{1-\frac{1}{k}}}{k^{\frac{1}{k}=1-\frac{1}{k}}}}\right)^n \tag{B.18}$$

$$= \left(\frac{k}{(k-1)^{1-\frac{1}{k}}}\right)^n \tag{B.19}$$

$$= \left(\frac{k(k-1)^{\frac{1}{k}}}{k-1} \right)^n := (D) \tag{B.20}$$

Combining (A) and (D) yields:

$$\left(\frac{3^{\frac{1}{k}}}{4} \right)^n * \left(\frac{k(k-1)^{\frac{1}{k}}}{k-1} \right)^n = \underbrace{\left(\frac{(3(k-1))^{\frac{1}{k}} * k}{4k-4} \right)^n}_{(E)} \tag{B.21}$$

Hence, an estimator for $F(n,k)$ can be written in the form of (A) * (E):

$$F(n,k) = \frac{\sqrt{2\pi n}}{\sqrt{2\pi n^{\frac{1}{k}}} * \sqrt{2\pi n \left(1-\frac{1}{k}\right)}} * \left(\frac{(3(k-1))^{\frac{1}{k}} * k}{4(k-1)} \right)^n \tag{B.22}$$

Note, that for $k \geq 2$ it follows that (E) decreases with k and $(E) < 1$. Hence $F(n,k)$ decreases at least exponentially with n where the base of the exponent decreases with k.

B.4 AptaMUT - List of Analyzed Mutants in Cluster 1

Legend: Seed Sequence | Enriched Mutants | Depleated Mutants

Pool Size 4: 1923823
Pool Size 5: 4621438

Cluster ID	Aptamer	Count Round 5	Fraction R5	Enrichment	Count Round 4	Fraction R4	Log Score
SEED 1	**TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC**	**351921**	**7.61E-02**	**30.31212404**	**4833**	**2.51E-03**	
p1	TAACACTCGATTCTCCTAGCCCTCTAGAAATTCCCCTCCC	3097	0.000670138	429.7421301	3	1.56E-06	-19.16434217
p2	TAACACTTGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	1093	0.000236506	227.4982959	2	1.04E-06	-6.862747998
p3	TAACACTCGATTCTCCTAGCCCGCAAGAAATTCCCCTCCC	336	7.27E-05	139.8708644	1	5.20E-07	-2.642604337
p4	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCCCCC	1131	0.000244729	117.7038301	4	2.08E-06	-4.484570278
p5	TAACACTCTATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	472	0.000102133	98.24263097	2	1.04E-06	-2.042432047
p6	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTTCC	432	9.35E-05	89.91698428	2	1.04E-06	-1.754571899
p7	TAACACTCGATTCTCCTAGCCCGCTAGAGATTCCCCTCCC	622	0.00013459	86.30920405	3	1.56E-06	-2.21397266
p8	TAACACTCGATTCTCCTATCTCGCTAGAAATTCCCCTCCC	204	4.41E-05	84.92159627	1	5.20E-07	-1.387821997
p9	TAACACTCGATTCTCCTAGCCCGCTAGATATTCCCCTCCC	404	8.74E-05	84.0890316	2	1.04E-06	-1.53965282
p10	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCTTCCC	388	8.40E-05	80.75877292	2	1.04E-06	-1.433547752
p11	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTACC	188	4.07E-05	78.26107891	1	5.20E-07	-1.228048933
p12	TAACACTCGATTCTCCTAGCCCGCTATAAATTCCCCTCCC	1445	0.000312673	75.19099669	8	4.16E-06	-2.98686202
p13	TAACACTCGATTCTCCTAGCCCGCTACAAATTCCCCTCCC	174	3.77E-05	72.43312623	1	5.20E-07	-1.130789174
p14	TGACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	167	3.61E-05	69.51914988	1	5.20E-07	-1.108986469
p15	TAACACTCAATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	645	0.000139567	67.12552646	4	2.08E-06	-1.582980065
p16	TACCACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	315	6.82E-05	65.5644677	2	1.04E-06	-0.901286456
p17	TAACACTCGATTCTTCTAGCCCGCTAGAAATTCCCCTCCC	305	6.60E-05	63.48305603	2	1.04E-06	-0.837647438
p18	TAACACTCGCTTCTCCTAGCCCGCTAGAAATTCCCCTCCC	152	3.29E-05	63.27491486	1	5.20E-07	-1.067580847
p19	TAACACTCGATTCTCCTAGTCCGCTAGAAATTCCCCTCCC	443	9.59E-05	61.47102475	3	1.56E-06	-1.091191611
p20	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCTCCTCCC	424	9.17E-05	58.83456996	3	1.56E-06	-0.99261414
p21	TAACACTCGATTCTCCTAGCCCACTAGAAATTCCCCTCCC	540	0.000116847	56.19811518	4	2.08E-06	-1.022545259
p22	TAACATTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	266	5.76E-05	55.36555051	2	1.04E-06	-0.595583672
p23	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCT	660	0.000142813	54.94926817	5	2.60E-06	-1.07750217
p24	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCACTCCC	372	8.05E-05	51.61900949	3	1.56E-06	-0.69557297
p25	TAACACTCGATTCTCCTAGCCCGCTAGAAATTACCCTCCC	123	2.66E-05	51.20272716	1	5.20E-07	-0.891868083
p26	TAACACTCGATTCTCCTATCCCGCTAGAAATTCCCCTCCC	1428	0.000308995	45.72701337	13	6.76E-06	-0.898079417
p27	TAACACTCGATTCTCCTAGCCCGCTAGAAATCCCCCTCCC	867	0.000187604	45.11459802	8	4.16E-06	-0.705078901
p28	TAACACTCGATTCTCCTAGCCCGCTGGAAATTCCCCTCCC	431	9.33E-05	44.85442156	4	2.08E-06	-0.502529456
p29	TAACACTCGATTCTCCTAGCCGGCTAGAAATTCCCCTCCC	106	2.29E-05	44.12592747	1	5.20E-07	-0.524085542
p30	TAACACTCGATTCTACTAGCCCGCTAGAAATTCCCCTCCC	105	2.27E-05	43.70964514	1	5.20E-07	-0.524085542
n1	TAACCCTCTATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	30	6.49E-06	12.48847004	1	5.20E-07	-2.136210865
n2	TAACACTCGATACTCCTAGCCCGCTAGAAATTCCCCTCCC	24	5.19E-06	9.990776031	1	5.20E-07	-2.628904498
n3	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCG	125	2.70E-05	8.672548638	6	3.12E-06	-0.545491351
n4	TAACACTCGATTGTCCTAGCCCGCTAGAAATTCCCCTCCC	40	8.66E-06	8.325646693	2	1.04E-06	-0.535680901
n5	TAACACTCGATTCTCCTAGCCCGCTAGAAAATCCCCTCCC	36	7.79E-06	7.493082023	2	1.04E-06	-0.891422702
n6	TAACACTCGATTCTCCTAGCCCGCTAGAAAGTCCCCTCCC	36	7.79E-06	7.493082023	2	1.04E-06	-0.891422702
n7	TAACTCTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	330	7.14E-05	6.868658521	20	1.04E-05	-2.56278133
n8	TAACCCTCAATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	10	2.16E-06	4.162823346	1	5.20E-07	-4.397795847
n9	TAACACGCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	30	6.49E-06	3.12211751	4	2.08E-06	-2.075661839
n10	TAACGCCCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	5	1.08E-06	2.081411673	1	5.20E-07	-5.049623067

B.5 AptaMUT - List of Analyzed Mutants in Cluster 2

Legend: Seed Sequence | Enriched Mutants | Depleated Mutants

Pool Size 4: 1923823
Pool Size 5: 4621438

Cluster ID	Aptamer	Count Round 5	Fraction R5	Enrichment	Count Round 4	Fraction R4	Log Score
Seed 2	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	636712	0.152938639	3.971389919	56879	0.038510104	
p1	TCACAGTTCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	12284	0.002950625	26.4123662	165	0.000111714	-117.057533
p2	TCACATTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2112	0.000507304	28.81854249	26	1.76034E-05	-27.25102639
p3	TCCCAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	3590	0.00086232	16.75839037	76	5.1456E-05	-21.98465566
p4	TCACAGTCCCGGTGCCGCACTAAAACCCATTTTTGTGCGA	690	0.000165738	18.83029765	13	8.80169E-06	-8.671586115
p5	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTTTGCGA	749	0.00017991	13.98555364	19	1.2864E-05	-6.523900233
p6	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTTCGA	855	0.000205372	13.18832756	23	1.55722E-05	-6.443160461
p7	TCACCGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	640	0.000153728	14.19094895	16	1.08328E-05	-6.064025487
p8	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCTA	578	0.000138836	13.67061416	15	1.01558E-05	-5.460460433
p9	TCACAGTCCCGGTGCCGCACTAATACCCATTGTTGTGCGA	277	6.65356E-05	19.6544643	5	3.38527E-06	-5.023946965
p10	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTCGA	176	4.22753E-05	31.22008769	2	1.35411E-06	-4.91133152
p11	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGC	262	6.29326E-05	18.59014313	5	3.38527E-06	-4.65225577
p12	TCACAGTCCCGGTGCCGCCCTAAAACCCATTGTTGTGCGA	581	0.000139557	12.12491374	17	1.15099E-05	-4.643814886
p13	TTACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	905	0.000217382	9.729400607	33	2.23428E-05	-3.93995426
p14	TCACAGTCCCGGTGCCTCACTAAAACCCATTGTTGTGCGA	1283	0.000308177	8.753359377	52	3.52068E-05	-3.552525129
p15	TCACAGTCCCGGTGCCGCACTAAAACCTATTGTTGTGCGA	542	0.000130189	10.12038728	19	1.2864E-05	-3.348674248
p16	TCACAGTCCCGGTGCCGCACTAAAACCCATTGATGTGCGA	183	4.39567E-05	16.23089786	4	2.70821E-06	-3.281538437
p17	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGT	455	0.000109291	10.08887777	16	1.08328E-05	-3.037677076
p18	TCACAGTCCCGGTCCCGCACTAAAACCCATTGTTGTGCGA	401	9.63205E-05	10.16173309	14	9.47874E-06	-2.88871647
p19	TCCCCGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	115	2.76231E-05	20.39948912	2	1.35411E-06	-2.882937703
p20	TCACAGTCCCGGTGTCGCACTAAAACCCATTGTTGTGCGA	969	0.000232754	7.994784613	43	2.91133E-05	-2.553277719
p21	TCACAGTCCCGGTGCCGCACTAAAACCCCTTGTTGTGCGA	101	2.42603E-05	17.91607305	2	1.35411E-06	-2.400091518
p22	TCACTGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	196	4.70793E-05	11.58927498	6	4.06232E-06	-2.39723436
p23	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGTGA	911	0.000218823	7.695211009	42	2.84362E-05	-2.240324041
p24	TCACAGTCCCGGTGCCGCACTAAAACTCATTGTTGTGCGA	458	0.000110012	8.551913973	19	1.2864E-05	-2.213139922
p25	TCACAGTCCCGGTGCCGCACTATAACCCATTGTTGTGCGA	118	2.83437E-05	13.95443314	3	2.03116E-06	-2.146358736
p26	TCACAGTCCCGGTGCCGCACTAAAATCCATTGTTGTGCGA	561	0.000134753	7.961122362	25	1.69263E-05	-2.057182526
p27	TCACAATCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	432	0.000103767	7.298202318	21	1.42181E-05	-1.470565613
p28	TCACAGTCCCGGTGCCGTACTAAAACCCATTGTTGTGCGA	430	0.000103286	7.264414344	21	1.42181E-05	-1.470565613
p29	TCACAGTCCCGGTGCCGCACTAAAACCCTTTGTTGTGCGA	114	2.73829E-05	10.11105113	4	2.70821E-06	-1.393542682
p30	TCACAGTCCCGGTGCCGCATTAAAACCCATTGTTGTGCGA	621	0.000149165	6.676196438	33	2.23428E-05	-1.270741846
p31	TCACAGTTCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	67	1.60934E-05	11.88491975	2	1.35411E-06	-1.265229786
p32	TCACAGTCCCGGTGCCGAACTAAAACCCATTGTTGTGCGA	164	3.93929E-05	8.311841529	7	4.73937E-06	-1.250908609
p33	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTCCGA	233	5.59668E-05	7.514752513	11	7.44758E-06	-1.224306127
p34	ACACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	116	2.78633E-05	8.230750392	5	3.38527E-06	-1.017681384
p35	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTACGA	501	0.000120341	6.347915558	28	1.89575E-05	-1.007569904
p36	TCACAGTCCCGGTGCCGCACTAAAACCCGTTGTTGTGCGA	347	8.33496E-05	6.479288534	19	1.2864E-05	-0.949563116
p37	TCACAGTCCCAGTGCCGCACTAAAACCCATTGTTGTGCGA	2226	0.000534687	5.806811097	136	9.20792E-05	-0.883078415
p38	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTCTGCGA	191	4.58783E-05	6.776178124	10	6.77053E-06	-0.854657601
p39	TCACAGTCCTGGTGCCGCACTAAAACCCATTGTTGTGCGA	803	0.000192881	5.813944902	49	3.31756E-05	-0.78047461
p40	TCACAGTCCCCGTGCCGCACTAAAACCCATTGTTGTGCGA	405	9.72813E-05	5.986806589	24	1.62493E-05	-0.759172506
p41	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCAA	675	0.000162135	5.701720561	42	2.84362E-05	-0.679284661
p42	TCACAGTCCCGGTGCCGCACTAAAACCCATAGTTGTGCGA	130	3.12261E-05	6.58865487	7	4.73937E-06	-0.648015293
p43	TCACAGTCCCGGTGCCGCACTTAAACCCATTGTTGTGCGA	47	1.12894E-05	8.337182509	2	1.35411E-06	-0.640776194
p44	TCACAGTCCCGGTGCCGCTCTAAAACCCATTGTTGTGCGA	173	4.15547E-05	6.137585421	10	6.77053E-06	-0.584369492
p45	TCACAGTCCCGGTGCCGCACTCAAACCCATTGTTGTGCGA	63	1.51326E-05	7.450248199	3	2.03116E-06	-0.575824787
p46	TCACAGTCCCGGTACCGCACTAAAACCCATTGTTGTGCGA	641	0.000153969	5.414522784	42	2.84362E-05	-0.513959741
n1	GCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	610	0.000146522	0.535673197	404	0.000273529	-8.878266926
n2	TCAAAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	414	9.94431E-05	0.597058218	246	0.000166555	-4.00194793
n3	TCACAGGCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	593	0.000142439	0.947661343	222	0.000150306	-3.932926632
n4	TCACAGTCCCGTGGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.129008627	11	7.44758E-06	-3.267316863
n5	TCAAAGTCCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.129008627	11	7.44758E-06	-3.267316863
n6	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGGGA	102	2.45005E-05	0.314668868	115	7.78611E-05	-3.252188056
n7	TCACAGTCGCGGTGCCGCACTAAAACCCATTGTTGTGCGA	66	1.58532E-05	0.650418494	36	2.43739E-05	-2.645311843
n8	TCACAGTCCCGGTGCCGCACTAAAACCCAGTGTTGTGCGA	69	1.65738E-05	0.281372264	87	5.89036E-05	-2.62296625
n9	TCACAGTCACGGTGCCGCACTAAAACCCATTGTTGTGCGA	164	3.93929E-05	0.684504596	85	5.75495E-05	-2.511193192
n10	TCACAGTCCCGGTGCCGCACTACAACCCATTGTTGTGCGA	75	1.8015E-05	0.239711976	111	7.51529E-05	-2.398060114
n11	TCACAGTCCCGGTGCCGCACGAAAACCCATTGTTGTGCGA	37	8.88742E-06	0.625077513	21	1.42181E-05	-2.38028024
n12	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTGGTGCGA	91	2.18583E-05	0.768676402	42	2.84362E-05	-2.365210473
n13	TCACAGTCCCGGTGCCGCACTAAAACCCATTGGTGTGCGA	106	2.54613E-05	0.854682153	44	2.97903E-05	-2.258368502
n14	TCACAGTCCCGGTGCCGCAGTAAAACCCATTGTTGTGCGA	72	1.72944E-05	0.774051761	33	2.23428E-05	-2.238622405
n15	TCACAGTCCCGGTGGCGCACTAAAACCCATTGTTGTGCGA	158	3.79517E-05	1.019168152	55	3.72379E-05	-2.181143995
n16	TCACAGTCTCGTTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.236515816	6	4.06232E-06	-1.96587893
n17	TAAAAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.236515816	6	4.06232E-06	-1.96587893
n18	TCACAGTCTCTGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.118257908	6	4.06232E-06	-1.96587893
n19	GCACAGTCCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.118257908	6	4.06232E-06	-1.96587893
n20	TCACAGTCTCGGTGCCGCACTAAAACCCATTGTTGTGCGA	743	0.000178469	1.432591722	184	0.000124578	-1.912815245
n21	GCACAGCCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.283818979	5	3.38527E-06	-1.693592934
n22	TCACAGTCCCGGGGCCGCACTAAAACCCATTGTTGTGCGA	867	0.000208254	1.297843116	237	0.000160462	-1.645064777
n23	TCACAGTCCCGTTTCCGCACTAAAACCCATTGTTGTGCGA	7	1.6814E-06	0.354773724	7	4.73937E-06	-1.578331946
n24	TCACAGTCCCGGTGCCGCACTAAAGCCCATTGTTGTGCGA	762	0.000183033	1.590221044	170	0.000115099	-1.542056568
n25	TCACAGTCCCTGTGCCGCACTAAAACCCATTGTTGTGCGA	882	0.000211857	1.596481757	196	0.000132702	-1.533162901
n26	TCCAAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.354773724	4	2.70821E-06	-1.444223753
n27	TCACAGGCCCGGTGGCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.177386862	4	2.70821E-06	-1.444223753
n28	TCAAAGTCCCTGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.177386862	4	2.70821E-06	-1.444223753
n29	TCACAGTCCCGGGGCCGCACTAAAGCCCATTGTTGTGCGA	2	4.80401E-07	0.177386862	4	2.70821E-06	-1.444223753
n30	TCACAGTCCAGGTGCCGCACTAAAACCCATTGTTGTGCGA	184	4.41969E-05	1.388901387	47	3.18215E-05	-1.439128166
n31	TCACAGTCCCGGTGACGCACTAAAACCCATTGTTGTGCGA	323	7.75848E-05	1.637027325	70	4.73937E-05	-1.244068404
n32	TCACAGTCTCGGTTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.236515816	3	2.03116E-06	-1.243820777
n33	GCACAGTGCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.236515816	3	2.03116E-06	-1.243820777
n34	GCACAGTCCCGGTGCCGCACTAAAACCCATTGCTGTGCGA	2	4.80401E-07	0.236515816	3	2.03116E-06	-1.243820777
n35	TCACAGTCCCGGTGCCGCACTAAACCCCATTGTTGTGCGA	194	4.65989E-05	1.5642296	44	2.97903E-05	-1.161986439
n36	TCACAGCCCCGGTGCCGCACTAAAACCCACTGTTGTGCGA	4	9.60803E-07	0.709547448	2	1.35411E-06	-1.138469215
n37	TCACAGTCCGGGCGCCGCACTAAAACCCATTATTGTGCGA	4	9.60803E-07	0.709547448	2	1.35411E-06	-1.138469215
n38	TCACAGTCCCTTTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.709547448	2	1.35411E-06	-1.138469215
n39	TCACAGTCCCGGCGCCGCACTAAACCCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n40	GCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGG	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n41	TCACAGTGTCGGTGCCGCACTAAAACCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n42	TCACCGTCCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n43	TCACAGTCCCGGGGCCGCACTAAAACCCATTGTTGTGCGG	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n44	TCACAGGCCCGGTGCCGCCCTAAAACCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n45	TCACAGTCCCGGTGCCGCAATAAAACCAATTGTTGTGCGA	2	4.80401E-07	0.354773724	2	1.35411E-06	-1.138469215
n46	GCACAGACCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.354773724	2	1.35411E-06	-1.138469215
n47	TCAAAGTACCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.354773724	2	1.35411E-06	-1.138469215

B.6 AptaTRACE - Experimental Details

CELL-SELEX Experimental Details

Cell-based SELEX was performed using an RNA library containing a randomized 30-nt region flanked by fixed primer sequences. Cell-based selection was performed as previously described [25], by employing open PCR for DNA amplification during each selection round. Positive selection was performed on Hela cells transduced with a bicistronic lentiviral vector expressing the target surface receptor and GFP, while unmodified Hela cells, which lack expression of the target receptor, were used for negative selection. High throughput sequencing (HTS) was performed on the positive selection at rounds 0, 1, 3, 5, 6, 7, 8, and 9. Significant molecular enrichment of receptor-specific aptamers was observed after five rounds of selection.

Flow Cytometry Analysis of Cell Surface Binding

PCR was used to amplify DNA oligomer templates which encoded our wildtype and mutant aptamer sequences. RNA containing 2'F-modified pyrimidines was transcribed from these templates using DuraScribe T7 Transcription Kit (Epicentre, Madison, WI). After purification via size-exclusion columns (Micro Bio-Spin Columns with Bio-Gel P-30, Bio-Rad), DNase I treatment, and ethanol precipitation, RNA was labeled with Cy3 using Label IT Nucleic Acid Labeling Kit (Mirus Bio, Madison, WI). HeLa cells expressing CCR7 (FUG-CCR7-W HeLa) or lentiviral vector backbone only (FUG-W HeLa) were dissociated with Accutase cell detachment solution (Innovative Cell Technologies, San Diego, CA). 2x105 cells per binding reaction were washed twice with pre-warmed DPBS containing calcium and magnesium (binding buffer) and then incubated with 100 1.55 µmg/mL of yeast tRNA for 15 min. Cells were then incubated with Cy3-labeled RNA (200 nM working concentration) and competitor tRNA in binding buffer at room temperature for 30 min. After incubation with RNA, cells were washed twice with binding buffer and resuspended in DPBS containing DAPI. Flow cytometry was used to analyze the percentage of cells bound by fluorescent RNA.

B.7 AptaTRACE - Parameters used in this Study

AptaTRACE: We set $\alpha = 2$ and $k = 6$ for our simulation data and $\alpha = 3$ and $k = 5$ for CELL-SELEX data. Furthermore a p-value of 0.01 was chosen as threshold for filtering out non-significant k-context traces from the selection. We set $\gamma = \frac{2}{3}$ for both simulation and CELL-SELEX data and we also allow the users to modify this parameter as required.

In order to choose which k-mers are to be merged with the cluster seed, we use the following decision scheme: K-mers fully overlapping with the cluster seed must differ from this k-mer by at least one mismatch. For partially overlapping k-mers and $k \leq 6$, the longest common substring (LCS) with the seed must be at least 4 nucleotides in length, while the LCS

Abstract: Aptamers are short, 15-150 nt long RNA/DNA molecules capable of binding, with high affinity and specificity, a specific target molecule via sequence and structure features that are complementary to the biochemical characteristics of the target's surface. The spectrum of aptamer targets spans from small organic molecules, over transcription factors and other proteins or protein complexes, to the surfaces of viruses and entire cells. This broad range of targets makes aptamers suitable candidates for a variety of applications including molecular biosensors, drug delivery systems, and antibody replacement.

Aptamers are typically identified via the High-Throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX) protocol. HT-SELEX leverages the well established paradigm of *in vitro* selection by repetitively enriching a pool of initially random RNA/ssDNA sequences (species) with those that strongly bind a target of interest. Specifically, based on the assumption that a large enough initial pool of randomized (oligo)nucleotides contains some species with favorable sequence and structure allowing for binding to the target, these binders are then selected for through a series of selection cycles. Each such selection cycle involves (a) incubating the pool with the target, (b) partitioning target-bound species from non-binders and (c) removing the latter from the pool, followed by (d) elution of the bound fraction from the target, and (e) amplifying the remaining sequences, one portion of which forms the input for the subsequent round and the remainder is sequenced.

However, optimal utilization of the HT-SELEX process has lagged behind the wide range biomedical applications due to the lack of dedicated computational approaches. The key challenges in HT-SELEX data analysis include the identification of target-affine aptamers and aptamer families which are selected for, as well as the elucidation of common sequence-structure binding motifs. In addition, next-generation sequencing of the aptamer pools enables studies of important properties of the SELEX protocol itself, such as the mutational landscape of aptamer sequences due to error prone amplification (PCR) which also require to be informed by computational tools. Finally, user friendly, aptamer oriented software for demultiplexing and quality control of the raw sequencing data is also of great relevance.

To close this gap we, have developed several novel computational methods designed to tackle these challenges and to elucidate previously unappreciated properties of the SELEX protocol. For standardizing the preprocessing of raw sequencing data, we introduce AptaPLEX, a standalone and platform independent demultiplexer and quality control tool specifically designed for HT-SELEX data. Next, in order to study the effect of the selection pressures exerted during SELEX as well as to test our algorithms, we developed AptaSIM, aimed at realistically recreating the general purpose SELEX protocol *in silico*. AptaSIM simulates error-prone PCR and many aspects of *in vitro* experiments such as sampling effects, the presence or absence of pool contaminants, and aptamer affinity. To aid the identification of aptamer candidates, we introduce AptaCLUSTER, a novel technique that scales well with next generation sequencing data to efficiently cluster aptamer families in each selection round and to trace their behavior throughout consecutive cycles. Building on the results of AptaCLUSTER we then provide the first in-depth analysis of the mutational landscape of HT-SELEX experiments and propose a theoretic model capable of discriminating favorable mutants from those which decrease the binding affinity to the target (AptaMUT). Finally, with our new AptaTRACE algorithm we tackle the challenging task of sequence-structure motif identification in HT-SELEX data. AptaTRACE is built on the idea of tracing the dynamics of the SELEX process itself to uncover motif-induced selection trends. It is robust enough to be applicable to a broad spectrum of RNA/ssDNA HT-SELEX experiments independent of the target's properties, and capable of elucidating an arbitrary number of binding sites along with their corresponding structural preferences.

The results of our approaches can be visualized via AptaGUI which provides, to the best of our knowledge, the first graphical, multi-user, and platform independent user interface for navigating high throughput sequencing data from HT-SELEX experiments.