Universiteit Leiden

Opleiding Informatica

Datamining the

Peptide Sequenome

Name: E. Partodikromo

Date: 21/08/2015

1st supervisor: Dr. E.Schultes
2nd supervisor: Dr. S.Goultiaev

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

I would like to thank the LUMC BioSemantics Group, the LIACS Datamining Group and Ricardo Cachucho for their support and help with this research-project.

Contents

Chapter 1

Introduction

1.1 Background

Proteins are naturally occurring polymers of amino acids which are of fundamental importance to all biological processes. There are 20 chemically distinct amino acids found in life on earth. These can be arranged in arbitrary sequence order and have arbitrary sequence length. Peptides are short proteins, having fewer than 50 amino acids. In this study, we focus on peptides having 7 amino-acids.

In proteins and peptides, it is known that a particular sequence of amino acids determines a 3-dimensional folded molecular structure and associated chemical and physical properties. Hence, the amino acid sequence also determines the ultimate biological function.

Conceptually, protein structure can be divided in 4 levels:

1. Primary: The sequence of amino-acids.
2. Secondary: The formation of helical and β-sheet elements .
3. Tertiary: The formation of long range bonds between the secondary structure elements.
4. Quaternary: The formation of stable bonds between distinct protein molecules.

All 4 levels of protein structure can influence the function of a protein. With this in mind, it is sometimes desirable to search for peptide sequences which bind to a certain biological target. For example, peptides having structures that specifically bind bone tissue can act as drug targeting systems. The search (or selection) for these peptides is done through a laboratory technique called Phage Display Screening [1], which works as follows:

1. Using recombinant DNA methods [2], the genetic information for all possible 7-mer peptides is encoded into a bacterio-phage library. (Bacterio-phage are viruses that infect bacteria and can be easily grown in bacterial cultures). Each phage particle displays a unique 7-mer peptide on its surface.

2. The phage particles are then exposed to the target. Some phage will encode peptides having a structure which will bind the target.

3. The phage which did not bind are washed away. The bound phage are then extracted from the solution and amplified by replicating the phage (on bacterial colonies).

4. Using the phage isolated from the previous round, the whole process repeats again. Each iteration enriches the pool in phage particles that have peptides binding the target.

5. The genetic information from the residual phages is determined using DNA sequencing methods.

6. Once the peptide sequences have been determined, large quantities of particular peptides can be chemically synthesized and their binding properties independently validated and studied.

More recently, Next Generation Sequencing (NGS) technology has been applied in Phage Display, isolating millions of candidate peptide sequences [1]. However, from these large lists of candidate sequences, only a few can be taken into expensive validation experiments and developed as products. The main problem then is how to choose the very best peptides for the further research? This problem is made more difficult because even with the application of NGS, still less than 1 percent of the selected phage pool can be sequenced.

Presently, Phage Display yields no direct information about the folded structures of peptides. If this gap between sequence information on one hand, and structure information on the other could be closed, then it would be possible to prioritise peptide candidates in a rational and systematic way, bringing much added value to Phage Display technology.

As a step toward this ambitious goal, the LUMC-LIACS Biosemantics group created a very large database of all possible hepta-peptide sequence-structure combinations simulating the folding process of all possible 7-mer sequences. This database is called the Peptide Sequenome. Given the 20 distinct amino acids and the 7 positions in the peptides, there are 20 to the power 7 ″1.28 billion″ sequence possibilities in the database. The intention is to use the Peptide Sequenome to rationalize, at a structural level, the outcome of Phage Display-NGS experiments. Ultimately, using correlations on the entire sequence-structure landscape it may be possible to locate the global optimum peptide sequence from only a single round of Phage Display screening.

In this research project, we mine the Peptide Sequence database for global patterns in the sequence-to-structure mapping, and compare these patterns to empirical sequence data derived from Phage Display experiments conducted at the LUMC. This is a Big Data research project, where we have millions of sequences (from the lab) being compared to a billion structures (from the computer). The primary goal of this research is to see if the AGADIR structure predictions can explain features of the empirical data. The methodological beauty of this approach lies in the convergence of combinatorial optimization theory, large-scale computer simulation and data mining, and well-controlled high-throughput laboratory validation experiments.

1.2 Thesis Overview

In chapter 2, we introduce the Peptide Sequenome database. We begin with a discussion of the computer simulations of peptide folding and then describe the methods employed for data mining and characterizing the global sequence-structure landscape.

In chapter 3 we describe the large sequence data-sets obtained from Phage Display-NGS experiments carried out at LUMC. We also demonstrate how data from the Peptide Sequenome can be combined with data from Phage Display-NGS experiments.

In chapter 4 we analyze how the 7-mer structures predicted by computer simulation map onto the 7-mer sequences isolated by Phage Display laboratory experiments. We find a correspondence between the simulation and experimental data and interpret these findings considering potential limitations of the computational and experimental techniques.

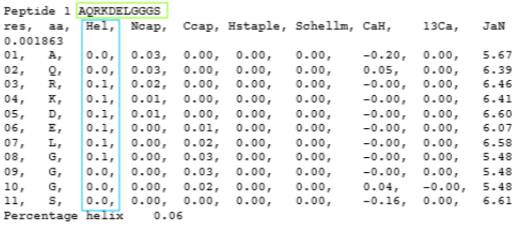
Last, we summarize the findings and conclusions and pose new questions for follow up investigations in chapter 5.

Chapter 2

Mining the Peptide Sequenome Database

2.1 Peptide Sequenome Database

In a collaboration between the LUMCs Human Genetics Department (E. Schultes) and the University of Amsterdams Computational Science Department (A. Belloum) a high-performance work-flow was created to generate the Peptide Sequenome database. First, the space of possible heptamer sequences was exhaustively enumerated. Each sequence was then submitted to a pseudo-molecular dynamics protein folding algorithm called AGADIR [3–9]. Among other things, AGADIR predicts the helical structure of the peptide. To be more exact, AGADIR outputs the expectation that each amino acid position in the peptide participates in a helical secondary structure (Figure 2.1). In the simulations run here, the 7-mer sequences were augmented with a C-terminal [10] 4-amino acid constant linker sequence (GGGS). This linker sequence was added because it is present in the Phage Display system (the linker was intentionally engineered to maximize exposure of the 7-mer peptide to its potential target without interference from the much larger phage particle to which it is attached). By including the linker sequence in the computer simulations, we can account for potential



```
Peptide 1 AQRKDELGGGS
res,  aa,  Hel,  Ncap,  Ccap, Hstaple, Schellm, CaH,    13Ca,   JaN
0.001863
01,   A,   0.0,  0.03,  0.00,  0.00,    0.00,    -0.20,  0.00,   5.67
02,   Q,   0.0,  0.03,  0.00,  0.00,    0.00,    0.05,   0.00,   6.39
03,   R,   0.1,  0.02,  0.00,  0.00,    0.00,    -0.00,  0.00,   6.46
04,   K,   0.1,  0.01,  0.00,  0.00,    0.00,    -0.00,  0.00,   6.41
05,   D,   0.1,  0.01,  0.00,  0.00,    0.00,    -0.00,  0.00,   6.60
06,   E,   0.1,  0.00,  0.01,  0.00,    0.00,    -0.00,  0.00,   6.07
07,   L,   0.1,  0.00,  0.02,  0.00,    0.00,    -0.00,  0.00,   6.58
08,   G,   0.1,  0.00,  0.03,  0.00,    0.00,    -0.00,  0.00,   5.48
09,   G,   0.0,  0.00,  0.03,  0.00,    0.00,    -0.00,  0.00,   5.48
10,   G,   0.0,  0.00,  0.02,  0.00,    0.00,    0.04,   -0.00,  5.48
11,   S,   0.0,  0.00,  0.00,  0.00,    0.00,    -0.16,  0.00,   6.61
Percentage helix    0.06
```

Figure 2.1: Example output of the AGADIR algorithm on a peptide from the Peptide Sequenome. The first column indexes the amino acid position (from N to C terminus). The second column shows the amino acid at that position using the single-letter amino acid code. The 7-mer sequence circled in green with its corresponding helical structure circled in blue. Note that the constant linker sequence is shown at positions 8-11, and is the same for all 1.28 billion 7-mer sequence possibilities. The seven remaining columns are other features of the peptides folded structure predicted by AGADIR, including NMR shifts (last 3 columns).

long-range interactions between amino acids in the 7-mer and the linker when forming folded structure. Furthermore, these extra positions comply with AGADIRs requirement for sequences to have minimally 10-amino acid positions.

The helical expectation at the amino-acid positions was found to range $[0.00, 52.80]$. The AGADIR output, that is the predicted structure of the peptide was thus represented as a real-valued vector $\{a_1, a_2, a_3 \cdots a_{11}\}$. The particular case where the helical expectation for all positions is 0.00, we named the NULL vector. Peptides (or proteins) associated with the NULL vector are predicted by AGADIR to contain no helical structure. However, this does not mean that such peptides have no structure whatsoever. It may well be that these peptides have other structures like β-sheet. This limitation of the AGADIR algorithm must be carefully considered when interpreting the results. Nonetheless, the AGADIR algorithm has 3 distinct advantages for this study:

1. AGADIR is computationally fast and is practical to use when computing over all possible 1.28 billion sequence possibilities.

2. AGADIR has three explicit parameters: temperature, pH and ionic strength. This permits the computation of structure predictions fitted to the specific conditions of real-world experiments.

3. Given (1) and (2), it is feasible to use AGADIR to explore the temperature, pH and ionic strength parameter space of the Peptide Sequenome. Establishing trends over this parameter space will yield thermodynamic insights for each peptide sequence that extend beyond static structure predictions.

In this study, ionic strength and pH were held constant at 60mM and 7.4 respectively, while temperature was adjusted to 3 distinct values: 277, 295 and 310 Kelvin (resp. 4, 22, 37 Celsius). In general for proteins and peptides, there is a relationship between folded structure and temperature. Typically, there will be a melting point where the folded structure at one temperature regime becomes denatured at a higher temperature regime, giving rise to an alternative structure or obliteration of unique structure altogether. The 3 temperature points investigated here were chosen with practical applications in mind: 4 degrees Celsius is the coolest temperature at which peptide experiments are typically be carried out; 22 degrees Celsius is a typical room temperature which is convenient and often used; 37 degrees Celsius equals the human body temperature, which for human drug development is the most relevant temperature. This resulted in a large sequence-structure MySQL database we called the Peptide Sequenome.

For the purposes of data mining, we partition the Peptide Sequenome into 6 main tables:

1. 3 tables for each of the 3 different temperatures. We refer to these as the original tables σ.

2. 3 tables containing the unique structures μ derived from each σ. These 3 tables are created because tables σ contain duplicate structures (due to a many-to-one sequence-structure relationship). Furthermore, these unique structures appear in multiple temperatures.

To get a better understanding of how folded structure is distributed in the total sequence space, we did the following. For each σ the amount of NULL vectors was filtered out to calculate the percentage that has helix structure. Next, we search for the unique structures of μ, for the purpose of further analysis (Table 2.1). It occurs that μ contains overlapping structures of all σ. Structures found in σ-310 can also be present in σ-277 or σ-295 and vica versa. Consequently, μ can be filtered into 7 temperature categories (Figure 2.2).

Temperature Category σ	% having helical structure	Number of sequences having helical structure
310	50,57	647.251.502
295	58,60	750.072.633
277	67,23	860.598.890

Table 2.1: Fraction of sequences in the Peptide Sequenome that have NULL vectors at different temperatures. These results correspond with the logical denaturation of proteins: high temperatures causes the breakdown of structures and thus higher percentage having NULL structure.

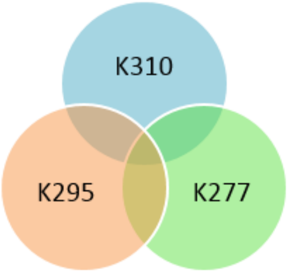


Figure 2.2: This Venn diagram shows the possible overlap between AGADIR structure vectors at the 3 temperature categories of the μ tables. Each circle represents the space of all possible 1.28 billion sequences. The 4 overlapping sectors represent those sequences having identical structure vectors at multiple temperature regimes. The 4 overlapping and 3 non-overlapping categories create 7 unique sequence-structure temperature relationships (see Tables 2.2 and 2.3).

2.2 Filtering algorithms

The following algorithms were used to partition μ. The α, β and γ variables represent the different temperature categories 277, 295 and 310.

1. Obtaining unique structures of μ in 1 temperature category:

   ⇒ insert into temporary_table A select structures from μ-α left join μ-β

   ⇒ insert into temporary_table B select structures from A left join μ-γ

   ⇒ insert into final_table select structures from B

2. Obtaining unique structures of μ in 2 temperature categories:

   ⇒ insert into temporary_table A select structures from μ-α inner join μ-β

   ⇒ insert into temporary_table B select structures from A left join μ-γ

   ⇒ insert into final_table select structures from B

3. Obtaining unique structures of μ in all 3 temperature categories:

   ⇒ insert into temporary_table A select structures from μ-α inner join μ-β

   ⇒ insert into temporary_table B select structures from A inner join μ-γ

   ⇒ insert into final_table select structures from B

The first algorithm utilises *left joins* only. A left join extracts data exclusively present in the left table and not the right. Therefore, the first algorithm can obtain the structures exclusively present in 1 temperature category. The second uses a combination of left and *inner joins*. It first isolates the structures *both* present in 2 temperature categories with an inner join. Afterwards it filters out any remaining structures of the third temperature (overlap) with a left join. Finally, the last algorithm obtains the structures exclusively present to **all** temperatures by using 2 inner joins. The following table displays the result of the filtering: The first 3 entries are the same as the 3 original μ tables. The amount of original structures in the next 3 categories is equal to the number of all structures belonging to both of the mentioned temperatures. The overlap of the third temperature is filtered out afterwards. The last category compromises the original structures of the first 3 categories, but with the removal of duplicates. Finally, the sum of all filtered structures totals the amount of original structures of μ-α_β_γ. This proves a check on the filtering process.

Temperature Category of μ	Number of original structures	Number of structures after filtering
Unique structures only in K310	1.472.710	574.218
Unique structures only in K295	2.231.423	997.595
Unique structures only in K277	3.922.242	2.663.905
Unique structures only in K277K295	1.102.707	490.966
Unique structures only in K277K310	767.371	155.630
Unique structures only in K295k310	742.862	131.121
Unique structures only in K277K295K310	5.625.176	611.741

Table 2.2: Sequence-structure-temperature partition of the Peptide Sequenome.

At this point, each unique structure is provided with a unique, though arbitrary numeric structure id for the sake of clarity (Table 2.3). The NULL-vector is assigned 5013436.

Temperature Category of μ	Index of structure_id
Unique structures only in K310	1 - 574.218
Unique structures only in K295	574.219 - 1.571.813
Unique structures only in K310	1.571.814 - 4.235.718
Unique structures only in K277K295	4.235.719 - 4.726.684
Unique structures only in K277K310	4.726.685 - 4.882.314
Unique structures only in K295k310	4.882.315 - 5.013.435
Unique structures only in K277K295K310	5.013.436 - 5.625.176

Table 2.3: Assignment of structure IDs for all 5.625.176 unique structures found over all three temperature regimes.

2.3 Temperature Dependence

Now we take a closer look at the temperature dependence of our peptide structures. Therefore, we plot the percentage of peptides having helical structure (Table 2.1, the 3 σ tables referring to the original calculated structure data) against the number of unique structures (Table 2.2, first 3 μ tables) for 277, 295 & 310 degrees Kelvin (resp. 4, 22 & 37 degrees Celsius).

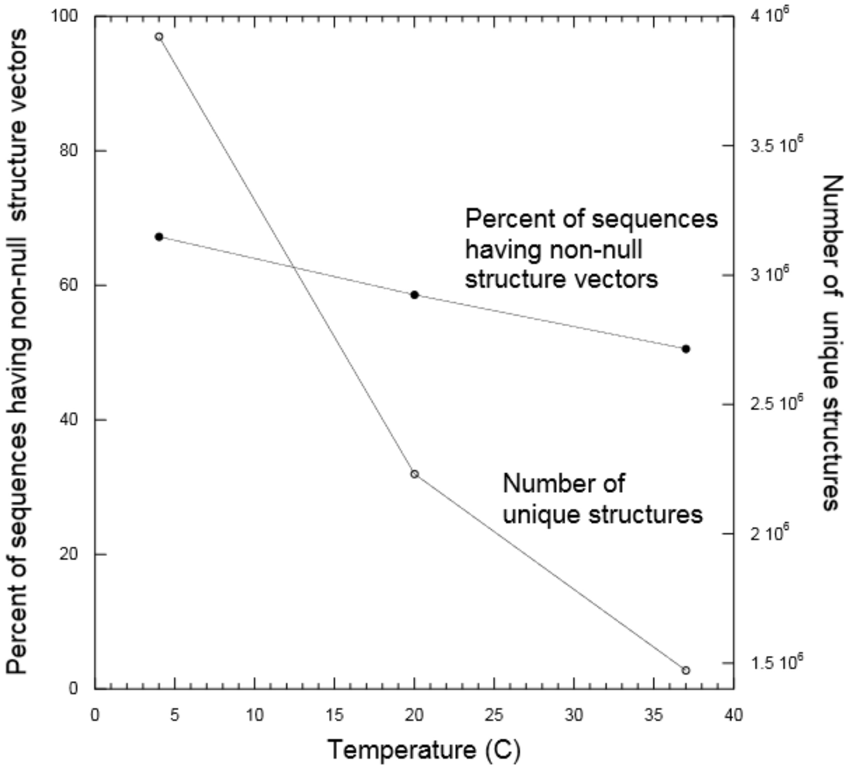


Figure 2.3: This figure shows the linear decrease of structure per temperature category, in percentage of the peptide sequences having structure. Secondly, it shows the decline of the number of unique structures per temperature category. The first y-axis represents the percentage of peptide sequences having non-null, i.e. helical structure. The second y-axis represents the number of unique structures. Finally, the x-axis depicts the temperature in Celsius.

2.4 Base Curve

Having now partitioned the data-sets structure and temperature, we focus exclusively on the peptide sequence space of 310 Kelvin.

From Table 2.2 we note that the amount of unique structures is relatively small (on the order of one million) compared to the number of unique sequences (on the order of one billion). Hence, there must be in general, a many-to-one sequence-to-structure relation in the peptide space. That is, in general, many different sequences are mapping to a given structure. This many-to-one relation guarantees a correlated structure landscape. We can express this relation by assembling frequency tables holding the the number of sequences per structure (**frequency counts**). But even before we can do that, we need to construct **sequence-structure** tables.

The sequence-structure tables are built by comparing the structure vectors of the σ tables to the 7 different structure tables. Specifically, each helical position ρ of $\{a_1, a_2, a_3 \cdots a_{11}\}^{\sigma}$ is compared to each helical position π of $\{a_1, a_2, a_3 \cdots a_{11}\}^{\mu}$. If and only if the 2 vectors are an exact match, the sequence name of the vector in question is retrieved from σ and mapped to its corresponding structure_id from μ. Consequently, 7 sequence-structures tables are obtained.

This mapping took about a week. This period of time could have been shortened by assigning indexes to the records within the σ tables. Another performance improvement could have been achieved by assigning the structure_id's to the 3 original μ tables. However, we chose to assign the id's after the partitioning because we preferred an ascending index-mapping for all temperature categories.

Now we can finally attach structures to the sequences obtained by Phage Display Selection through comparison by sequence-name. Even more, this enables us to construct the frequency tables for all temperature categories. Frequency counts is defined as the number of sequences mapping to a structure. Each structure_id is paired with a frequency count. The higher the frequency count, the more *common* the structures are. The data of the associated frequency tables F-310, F-277_310,F-295_310 & F-277_295_310 is combined and ordered by highest frequency counts. This totals to 1.472.410 million structure_id's, which covers the total 310 Kelvin sequence space. The graph produced from this data, is referred to as the Base Curve (Figure 2.2). We define this curve as a reference landscape, to which the Phage Display Data will be compared.

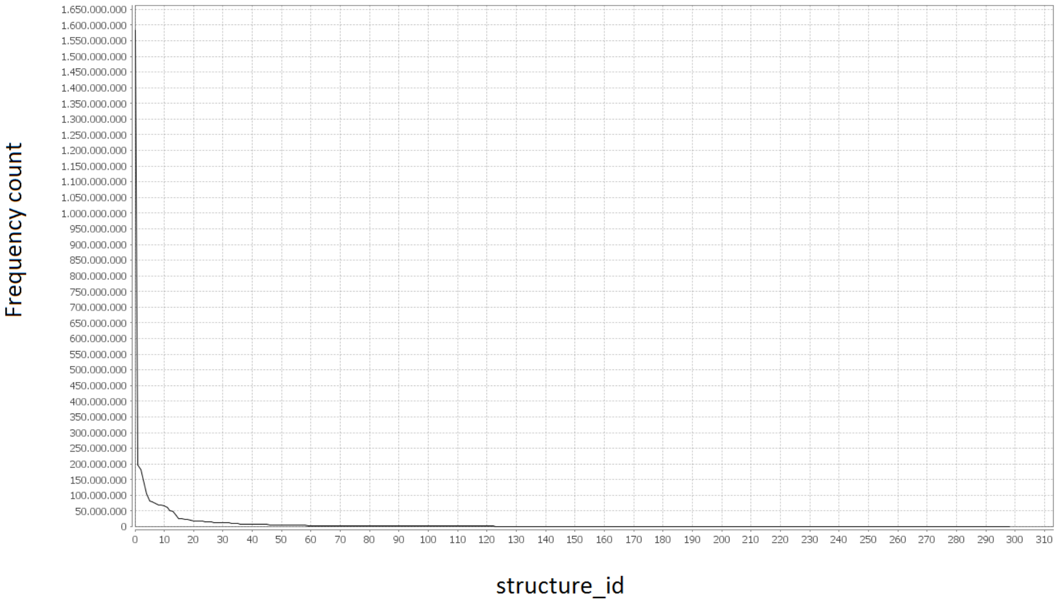


Figure 2.4: The base curve at 37C. The x-axis represents a rank ordering of the structure_ids according to the number of sequences that map to them (shown on the y-axis). Shown here are the first 300 of 1.472.710 unique structures found at 37C.

Chapter 3

Phage Display Data analysis

3.1 Phage Display Data & Naïve Libraries

The data generation process for the phage display data is divided into 2 phases. The first phase compromises **Phage Display screening**, the second **Deep Sequencing**. Here, we provide a recap of the process.

Phage Display screening is a laboratory technique which utilizes the surface display of peptides on phage particles to aid the selection of peptides, through a process of filtering and amplifying. It was first described by George P. Smith in 1985 [11]. Since then, Phage Display has become a widely used technique throughout the life sciences [12, 13] and has become commercially available from New England BioLabs [1].

In the Phage Display, the genetic information for all possible 7-mer peptides (1.28 billion sequence combinations) is encoded in bacterio-phage genomes. These phage are then exposed to a target of interest. Phage particles that have a peptide capable of binding the target will non-covalently bind with some level of affinity and specificity. Phage that bind with little or no affinity will be subsequently washed away. The bound phage, having been separated (selected) from the larger pool can now be eluted from the target and amplified. The amplification of the phage is accomplished by infecting bacterial hosts with the selected phage particles, whereby the selected phage viruses undergo many instances of replication. This process of selection and amplification constitutes a single round of selection.

The selection-amplification process can be iterated multiple times, each round enriching the selected population in phage binding to the target. Over the course of multiple rounds, the diversity of peptide sequences in the phage library tends to decrease from 1.28 billion to a lower value depending on the details of the experiment (for example, number of rounds, the nature of the target, the stringency of the selection protocol).

In this way, Phage Display is a physical instantiation of a combinatorial optimization algorithm not unlike a genetic algorithm or simulated annealing. In principle, after sufficient rounds of selection-amplification, Phage Display Screening will converge onto peptide sequences that are optimal in binding to the given target under the experimental conditions.

The outcome of the Phage Display Screen is a list of peptide sequences that bind the target. The sequence information is obtained by sequencing that part of the phage genome that encodes the peptide. In general, over the course of selection, strongly-binding peptide sequences will become enriched in the population and will have multiple sequence reads. It is assumed that the more sequence reads a peptide has (the higher its count in the sample) the better it is at binding the target. Until recently, sequencing was done is a manual way, producing at most a few hundred sequence reads. In 2012, the research group at LUMC was the first to apply high-throughput Next Generation Sequencing (NGS) technology to Phage Display [13]. DNA from the bound phage was processed in large amounts, which result in millions of sequence reads representing the selected peptides. The target in this case was human bone cell in tissue culture.

The application of NGS technology to Phage Display gives a high-resolution picture into the dynamics of the phage population with each round of selection-amplification. Using NGS, even rare sequences including sequences that occur only a single time can be identified in each round of selection-amplification. The ultimate goal in applying NGS to Phage Display was to determine if NGS technology could, by looking deeper into the selected pools, reduce the number of rounds of selection-amplification that were necessary to obtain suitable peptides. In order to assess the potential of NGS technology in Phage Display, samples were taken from the naïve (pre-selected) libraries as well as from each of 4 rounds of selection. Then, the sequence counts of each unique sequence was noted, creating a distribution of sequence-counts per round. Hence, the starting point for the Phage Display data analysis in this thesis is as follows (summarized in Table 3.1):

1. **4 rounds** of Phage Display data previously described in [13].

2. **3 Naïve Libraries** named A,B & C.These data reflect the distribution of peptides in the original Phage Display library before they are exposed to the target. In principle, each of the 1.28 billion possible 7-mer sequences should be uniformly represented in the original pools. However, physical, chemical, and biological effects are known to induce biases in the distribution of peptide sequences even before the selection step is initiated. Library B is a sequenced, non-amplified library. Library A is the amplified library from B after one round of Phage Display screening. Finally, C is the first round amplified naïve library used for the Phage Display screening and sequencing of the 4 Phage Display rounds.

Data-set	Original reads	Unique sequences
Round 1	14.725.587	1.792.654
Round 2	16.619.643	1.228.666
Round 3	12.647.578	280.909
Round 4	13.540.211	236.516
Naïve A	25.962.770	8.236.941
Naïve B	7.871.525	5.415.512
Naïve C	6.189.787	3.489.292

Table 3.1: Experimental data: NGS sample size, and the number of unique sequences found, for the Naïve Libraries and the 4 rounds of Phage Display data.

Henceforth, we refer to these data-sets as *experimental data*. This is in contrast to the AGADIR-based structure prediction data-sets discussed in the previous chapter, which we call *simulation data*. The 1.28 billion peptide sequences are common elements of both data-sets and hence, the sequences act as a bridge linking experimentally determined NGS sequence counts on one hand and the AGADIR-predicted frequency counts on the other. The most fundamental question we can ask is: Are there correlations between these Big data-sets? In other words, are there features of the AGADIR-predicted structure landscape that can be used to explain features of the NGS Phage Display sequence counts? If so, can the AGADIR-predicted structure landscape be used to optimize peptide binding for this, and perhaps for other targets? Answering these questions imply an exploratory mining of Big Data. To provide focus in our search for correlations, we formulate and test a simple hypothesis:

Given that some AGADIR-predicted structures are more common in sequence space (Chapter 2), we assume that Phage Display screening tends to result in sequences having common structures. That is, the sequences (and structures) recovered in Phage Display experiments are those that are most easily found throughout the peptide sequenome. Although sequences having rare structures could in principle also bind to the target, they will be more susceptible to noise in the system and have lower probability of successful amplification. If this hypothesis is true, then we expect to see a correlation between experimentally determined NGS sequence counts on one hand and the Base Curve on the other.

If this hypothesis was not supported by the data (i.e., we find there is no correlation between the experimental and simulated data-sets), then either Phage Display screening selects for other features of peptide structure or the AGADIR prediction algorithm fails to capture structural features relevant for peptide binding. In the next section we describe the assembly of a new data-set whereby the peptide sequence is used to combine the experimental and simulated data-sets so that we can search for correlations between NGS sequence counts and AGADIR-predicted frequency counts. We then interpret the results in Chapter 4.

3.2 Combining experimental and simulated data-sets

First, we build new tables where the sequences of the experimental data are provided a structure_id and helical structure vector (at 37C). The structure_id is associated with a sequence by comparing the sequence name of the experimental data with sequence name of the sequence-structure tables of the 310 Kelvin space. The helical structure is assigned to its corresponding sequence by matching the sequence name from the experimental data to sequence names of the σ-310.

Sequence name	Sequence counts	Helical Vector	Structure_id
AAAAAFWGGGS	1	{0, .2, .2, .2, .2, .1 ,0,0,0,0,0}	5158499
AAAAAPGGGGS	1	{0,0,0,0,0,0,0,0,0,0,0,}	5013436
AAAAAPPGGGS	4	{0,0,0,0,0,0,0,0,0,0,0,}	5013436
AAAAASWGGGS	3	{0, .1, .1, .2, .2, .1,0,0,0,0,0}	5085886
AAAADTCGGGS	4	{0,0,0,0,0,0,0,0,0,0,0,}	5013436

Table 3.2: A representative example of the combined experimental-simulated data-sets. Listed here are the 5 sequences, their corresponding structure_id and their Helical Structure Vector, as well as the Sequence Counts from Round 1 of the Phage Display screening.

Now, we build tables, containing for each structure_id, the **frequency counts** F (number of sequences mapping to the AGADIR structure) from these data-sets to construct the 7 core tables τ. To these tables, we also add the sum S of the sequences counts per structure to each data-set from τ. Thus, τ consists of structure_id's with their associated frequency count and S of sequence counts. Finally, we normalize by dividing F and S by the number of **original** sequences per round.

Structure_id	Frequency count	Sum of sequence counts
5013436	802.9479572	6726.7973766
5085886	0.6580383	3.9237825
5158499	1.8213196	15.2421768

Table 3.3: This is the τ-Round1 data-set associated with the original Round 1. The exact same structure_id's are also shown. It occurs that the NULL-vector has an extremely large F and S, in comparison to the other structures.

The frequency counts and sum of sequence counts both serve as measures which will determine whether our hypothesis holds. By mapping them onto the base-curve, we hope to confirm and the correlation between our experimental data and the AGADIR data.

Chapter 4

Results

By having the τ, i.e. unique structure tables, from our 7 data-sets on hand, we now can now analyze the experimental and simulation data. First, we investigate possible trends in sequence and structure diversity over rounds of phage display selection. Then we look for correlations between NGS sequence counts and AGADIR-predicted frequency counts.

1 Data-set	2 Original Reads
NaïveA	25.962.770
NaïveB	7.871.525
NaïveC	6.189.787
Round1	14.725.587
Round2	16.619.643
Round3	12.647.578
Round4	13.540.211

Table 4.1: The first column lists the experimental data-set, either the naïve library or selection round. The second column (Sample size: original number of reads) gives the number of NGS reads acquired.

As described previously (Chapter 3 & [13]), we would expect the diversity of sequences to diminish from the naïve libraries through Round 4 (Table 4.2 columns 2 and 3). Presumably, this is because relatively few (approximately 12.1% to 1.7%, Table 4.2 column 2) peptide sequences from the 1.28 billion possibilities are binding the target and are being enriched in the population by the process of phage display. However, unexpectedly, the number of unique sequences having helical structure (Table 4.2 columns 4 and 5) decrease from 34% to 29% over the course of the selection. In other words, phage display appears to be selecting for sequences that are associated with the Null-vector.

1 Data-set	2 Unique sequences	3 % Unique Sequences	4 US having helical structure	5 % Having helical structure	6 Unique Helical structures	7 % of Unique Helical structure
NaïveA	8.236.941	31.73	3.022.421	37	47.145	1.5
NaïveB	5.415.512	68.80	1.960.641	36	35.705	1.8
NaïveC	3.489.292	56.37	1.271.764	36	27.880	2.2
Round1	1.792.654	12.1	610.266	34	15.644	2.6
Round2	1.228.666	7.39	409.218	33	11.644	2.8
Round3	280.909	2.32	84.210	30	4.062	4.8
Round4	236.516	1.74	67.705	29	3.458	5.1

Table 4.2: Trends in the diversity of peptide sequences and structures over rounds of phage display selection. The first column lists the experimental data-set. Because some sequences are found multiple times among the NGS reads, the second column gives the number of Unique Sequences. Some of the Unique Sequences, have non-Null-vector helical structure, and these listed in the 4th column. As with sequence, many of these helical structures appear multiple times. In the 6th column, we list the number of unique structures found in the original sample (these are the number of unique helical structures from the τ tables which we extracted in Chapter 3). Columns 5 and 7 are normalized as follows: The percentage having helical structure is calculated by dividing the unique sequences having helical structure (column 4) by associated number of unique sequences, per data-set (column 2). The percentage of unique helical structures is obtained by dividing the unique helical structures (column 6) by amount of unique sequences having helical structure, per data-set (column 4). Last, the percentage of unique sequences is calculated by dividing the number of unique sequences (column 2) by the original reads (column 2 of Table 4.1).

Also surprising, regardless of the trend of increasing Null-vector, it is evident that the diversity of unique helical structures (Table 4.2 columns 6 and 7) increases per round of Phage Display screening (nearly doubling from 2.6% to 5.1%).

In figure 4.1, the percentage of unique structures per the unique sequence sample size (Table 4.2 column 6 divided by column 2) is plotted against the percentage of sequences having helical structure (Table 4.2 column 4). Note (1), there is a clear trend from lower right to upper left between the naïve libraries and extending through rounds 1-4, (2) there is a gap between rounds 1-2 and rounds 3-4, (3) there is a gap between experiment and random samples. It shows that Phage Display selection filters out all redundant sequences (noise) first, and selects for more specific ones in rounds 3 and 4. The fact that the random samples are situated far away also suggests that Phage Display does not select random sequences with respect to AGADIR assigned structure_id's.

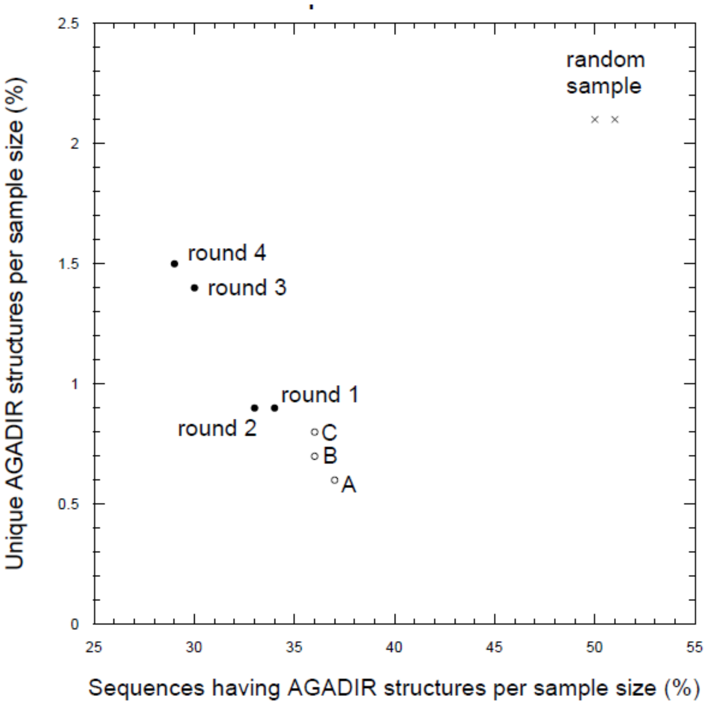


Figure 4.1: On the y-axis, the percentage of unique structures per the unique sequence sample size is plotted while the x-axis depicts the percentage of sequences having helical structure. There is a trend from the naïve libraries through rounds 1-4 is visible, as well as the gaps found between rounds 1-2 and rounds 3-4 and between the experimental and random samples.

Next, we present plots of the unique structure tables of experimental and random data and look for correlations between NGS sequence counts and AGADIR-predicted frequency counts. These data-sets (Figures 4.2 and 4.3) are all ordered with respect to the frequency count of the Base Curve. They show the first 300 (of more than 1.4 million) structure_id's. Since the NULL-vector has an extreme amount of frequency counts and sums of sequence counts, we chose to disregard it from the sets.

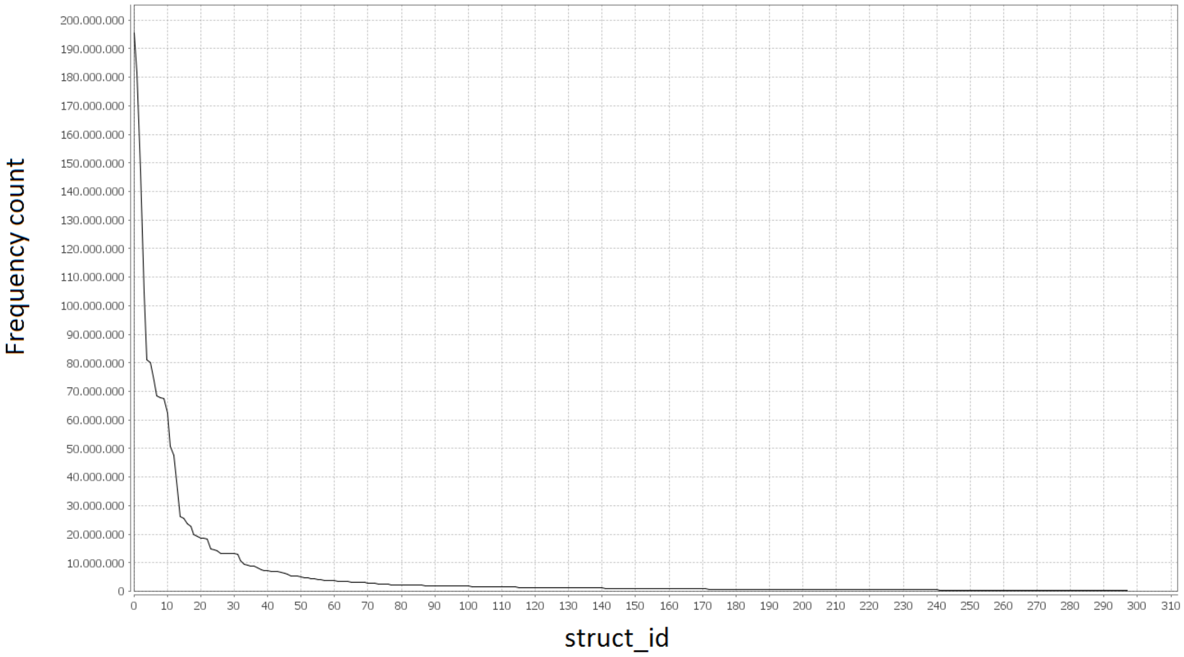


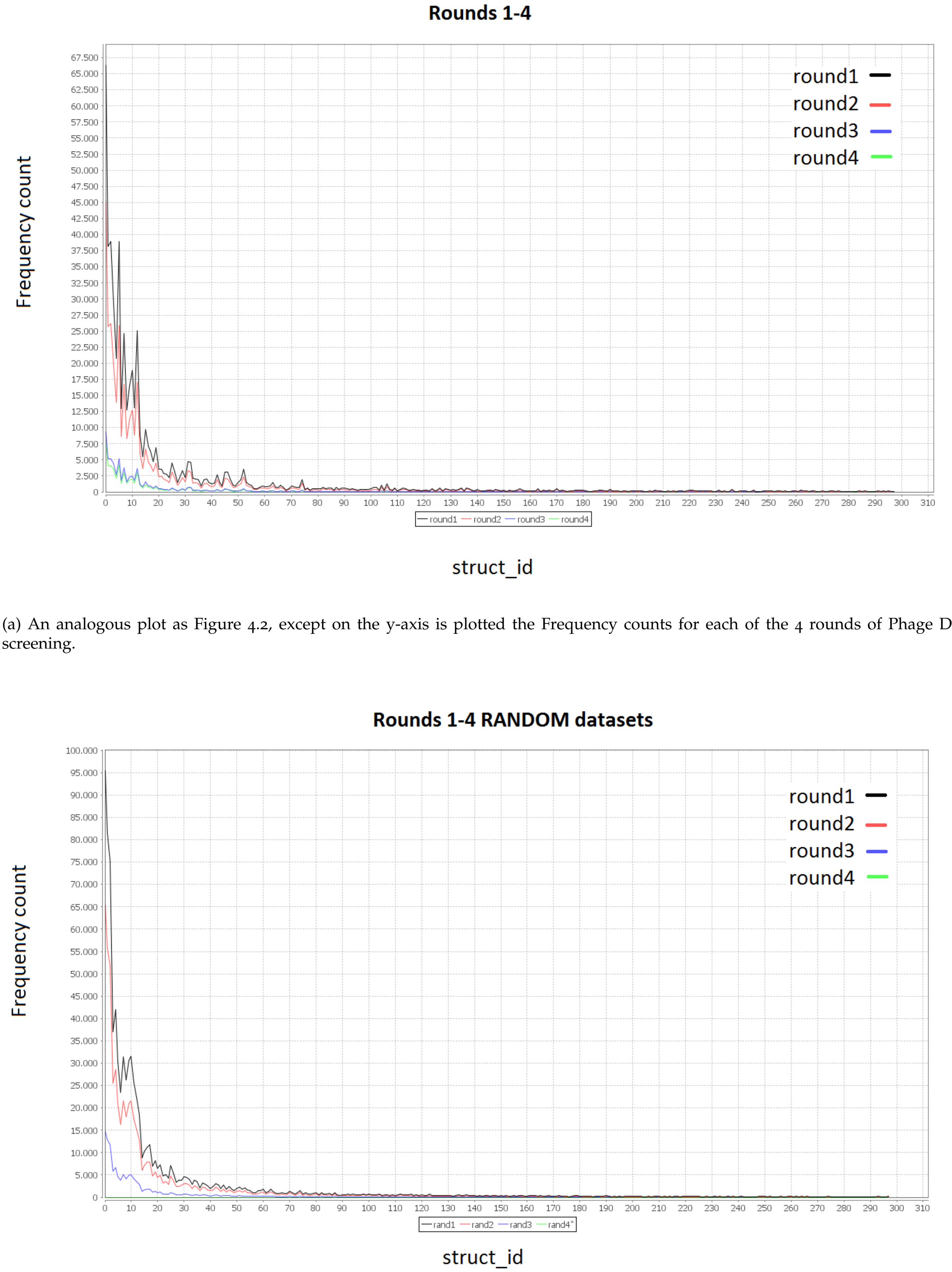
Figure 4.2: The Base Curve computed in Chapter 2 depicted here without the NULL-vector.

In figure 4.2, the 300 top-ranking structure_ids are plotted on the x-axis and their frequency counts (number of sequences mapping to each structure_id) on the y-axis. The Base Curve demonstrates a power-law-like distribution of sequences mapping to helical structures: Relatively few structures (those on the left) are commonly found throughout sequence space, while a long tail of structures (those on the right, and continuing to 1.4 million other structure_id's) are rarely found in sequence space.

In figure 4.3a, these data track the trend of the Base Curve seen in Figure 4.2. Note, the divergence from rounds 1-2 and rounds 3-4. However, it is too early to draw a plausible conclusion from these findings. Therefore, we included a similar plot of data-sets from randomly generated sequences to be able to understand our previous discovery.

The random data-sets show the same divergence as the Phage Display rounds (Figure 4.3a), but round 4 resembles a flat line. Presumably, this line is a result of (a) the sample size and (b) frequency count of 1 which was assigned to each random unique sequence. Hence sample size clearly effects the distribution of selection results on this ranking of the structure_id's. From these plots, we cannot tell the difference between random and phage display selected samples.

Figure 4.3: Frequency count plots of the Phage Display selection and random data rounds.



(a) An analogous plot as Figure 4.2, except on the y-axis is plotted the Frequency counts for each of the 4 rounds of Phage Display screening.

(b) This plot shows the distribution of randomly generated sequences of 4 rounds with the same sample size as the rounds 1-4 of Phage Display screening. With the exception of round 4 (nearly flat line), these data depict the same trend we observed earlier in Figures 4.3.

In the previous plots, we are seeing the effects of sample size. In an effort to control for that, we normalize our data using the sum of sequence counts (Figure 4.4). Again, the sum of sequence counts represents the total amount of sequence counts per unique structure. Moreover, the previous plots are also naïve, but it would be hard to understand the normalised plot without them.

In Figure 4.4, we observe numerous anomalies where certain structure classes (indicated with arrows) are over-represented in the phage display selection. The sum of sequence counts for these anomalous structure classes also increases sequentially over each round. Apparently, Phage Display is selecting for many different sequences having the same structure, presumably because these structures have especially high binding affinity and specify for the target. Note, the sum of sequence counts for random samples gives no peaks. Therefore this is a real signal from the selection.

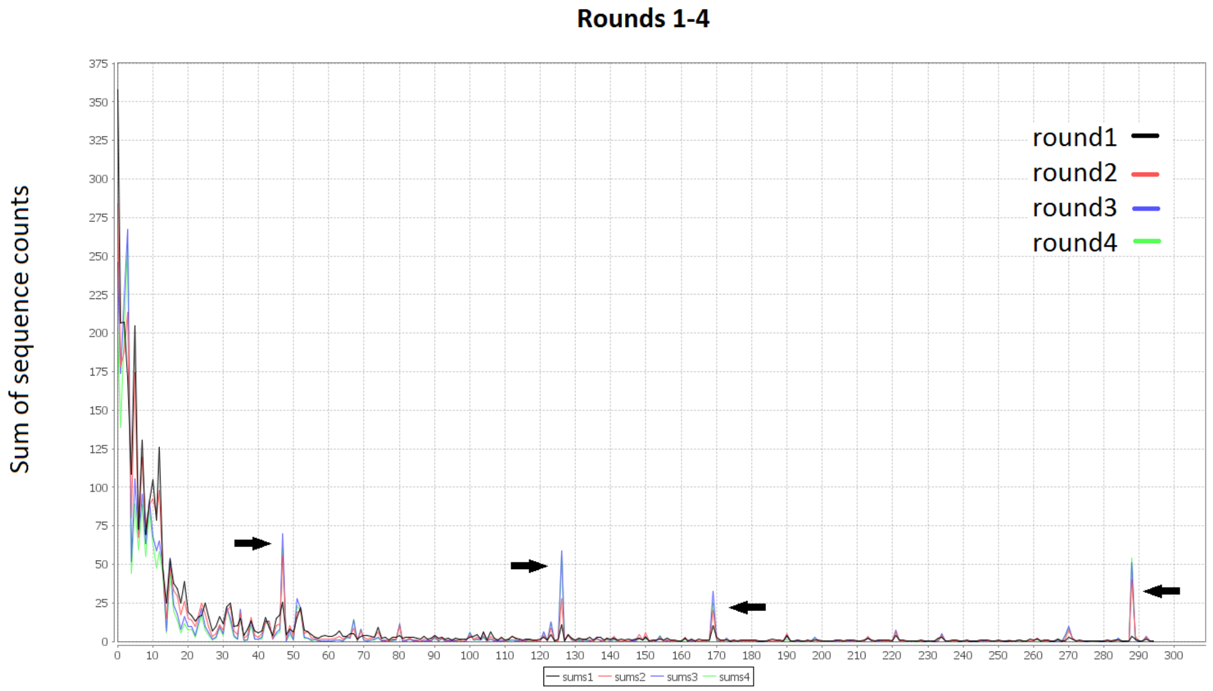


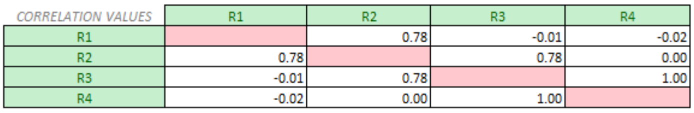
Figure 4.4: Sum of sequence counts of the 4 rounds of Phage Display screening (as described in Chapter 3). In contrast to the absolute frequency counts, these data were normalized by their original reads. Although we see the same trend in the distribution as in the previous plots, we also see some anomalies indicated with arrows (high peaks).

Next, we demonstrate an interesting correlation between the sum of sequence counts columns from the 4 rounds of Phage Display screening (Figure 4.5). This is done by calculating the *correlation coefficient* [14] for all possible pairs x, y, representing these columns, of our Phage Display rounds:

$$C = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the sample means of x and y. The coefficient C is defined as a value between $[1, -1]$, with positive 1 indicating a perfect positive correlation and minus 1 a perfect negative correlation.

As can be seen, the successive rounds of the selection have high correlation, while the other possible correlations are minimal. This demonstrates a trend in the sum of sequence counts over the course of selection.



CORRELATION VALUES	R1	R2	R3	R4
R1		0.78	-0.01	-0.02
R2	0.78		0.78	0.00
R3	-0.01	0.78		1.00
R4	-0.02	0.00	1.00	

Figure 4.5: This matrix depicts correlation values of the sum of sequence counts of the 4 Phage Display selection rounds.

Finally, we conclude this chapter with a distribution analysis of the *common* and *rare* structures of the Phage Display selection rounds, i.e. we want to evaluate the distribution of the structures coming from the experimental data of the Phage Display rounds 1-4 within the sequence space. These data will test directly our hypothesis that Phage Display selects for common structures.

The first step in defining common and rare structures is to calculate the 80%-20% threshold of the Base Curve. The 80%-20% threshold of the Base Curve was found at rankorder_id 1282 with its associated structure_id 5086460. We arbitrarily refer to structure_id's left this threshold as common structures and right from this threshold as rare structures. Thus, we obtained a list of these common structures, which we compared to the unique structure_id's (extracted from their associated τ tables) of the 4 Phage Display selection rounds. The amount of the matched common structure_id's was afterwards subtracted from the total amount of unique structure_id's per round, giving us the number of rare structure_id's. Therefore, providing the partitioning of the common and rare structures. We normalized this data with respect to the total amount of unique structure_id's. In Figure 4.6, the blue bar represents the fraction of common structure_id's per round while the orange bar shows the fraction of unique structure_id's which belong to the rare structures. The plot depicts a clear trend: the amount of *common* structures rises over the course of the selection while the amount of *rare* structures decreases.

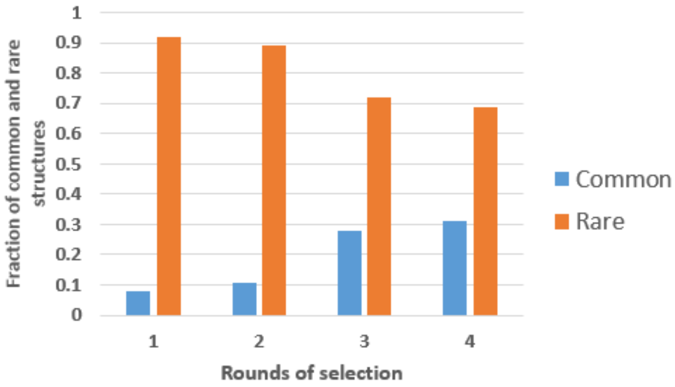


Figure 4.6: The distribution analysis of the 4 Phage Display selection rounds. The fraction of unique common and rare structure_id's is plotted on the y-axis while on the x-axis, is plotted the rounds of selection. The blue bar represents the fraction of the common unique structure_id's per round and the orange bar shows the fraction of unique structure_id's which are rare.

Chapter 5

Conclusion

Having produced interesting results through combining our experimental and simulated data, we can now form some conclusions.

In chapter 1 we provided a general introduction regarding our research-project. In chapter 2, we then discussed the datamining of the computer simulated data, our AGADIR predicted peptide structures. In chapter 3, we subsequently explained the parsing and mapping of frequency counts onto structures. Finally in chapter 4, we reviewed our results acquired from the datamining. Here we review particular results and derive conclusions regarding the role of predicted secondary structure in the outcome of Phage Display experiments.

The first result (Figure 4.1) showed a trend that the percentage of unique structure rises from the naïve libraries through the 4 Phage Display rounds. Since the amount of unique sequences having helical structure decreases, you would expect that the percentage of unique structures also decreases. This would be logical if there existed a one-to-one sequence-structure relationship. Because many unique sequences map to a structure, and the amount of these sequences decreases, the percentage of unique structures would be expected to increase for reasons having nothing to do with target binding. In other words, this trend is not relevant regarding our hypothesis. With additional simulations and analysis on the experimental data, this explanation could be validated. If this explanation was not validated then this trend would likely indicate a correspondance between AGADIR predicted structures in silico and real molecular structures in vitro.
Another result from Figure 4.1 is a large gap visible between rounds 1-2 and rounds 3-4. When selecting for peptides through rounds 1-4, a big decrease of unique sequences between rounds 1-2 and between rounds 3-4 was noted (Table 4.2, column 2). This observation is consistent with how Phage Display works; it is common knowledge among experimentalists that at least 3 rounds of Phage Display screening are needed to obtain adequate binders. However, based on these data it is not possible to interpret the structural meaning of this

gap. From the test of correlation between sum of sequence counts of the Phage Display rounds (Figure 4.5), we see a trend between successive rounds. This implies an evolution of the sequence-structure distribution between the early rounds of selection and later rounds of selection.

A third observation noted from Figure 4.1 is the large gap noted between the experimental and randomly generated data as described in Chapter 4. We can deduce from this that Phage Display is doing something other than randomly selecting sequences. Although in principle the naïve libraries should represent random samples of sequence space, similar to the randomly generated data, the observed gap shows that in practice the naïve libraries are biased with respect to the AGADIR predicted structures. Furthermore, the bias in the naïve strongly impacts the course in the selection experiment.

Figure 4.4, which demonstrated the sum of sequence counts distribution of the 4 Phage Display rounds, showed a similar trend to that observed in the Base Curve, but also contained anomalies where certain structures where overrepresented. This observation can be interpreted that Phage Display selects for many different sequences having the same structure, apparently because these structures have a high binding affinity and focus on the target. However, we have not determined if these anomalies have resulted from a single sequence or many sequences. This would be interesting material for a follow-up study.

Our initial hypothesis was that Phage Display selects for common structures. We expected the distribution of common structures to go up through rounds of selection, and the distribution of rare structures to decrease. This was confirmed in the distribution analysis (Figure 4.6).

Given these findings we conclude that, (1) there exists a correlation between the AGADIR predicted structures and the experimental data, and (2) Phage Display selects for sequences having common structures.

Appendix

An overview of the definitions used in this thesis.

1. Sequence - A peptide's amino-acid sequence, i.e. 'peptide'.
2. Phage - A virus.
3. Helical structure - A secondary protein (peptide) structure.
4. μ - The tables containing unique structures, per temperature category.
5. σ - AGADIR's prediction of helical structure for 3 different temperatures for the 1.28 billion 7-mer sequences.
6. F - Frequency count, the number of sequences mapping to a structure.
7. τ - The core tables of the Experimental data, which consist of unique structures plus their frequency counts and sum of sequence counts.
8. Sequence count - The number of times a sequence is encountered in a round of Phage Display screening.
9. S - Sum of the sequences counts per structure.
10. Base Curve - A reference landscape of the 1.4 million structure_id's and their frequency counts of the 310 Kelvin space.
11. Naïve Library - A library which has unselected sequences.

Bibliography

[1] "Phage Display Screening." https://www.neb.com/applications/protein-analysis-and-tools/~/media/1231BB0939D54AAA96E04FE0C8EA7437.ashx.

[2] "Recombinant DNA technology." http://www.britannica.com/science/recombinant-DNA-technology.

[3] "AGADIR general information." http://agadir.crg.es/.

[4] V. Munoz and L. Serrano, "Elucidating the folding problem of helical peptides using empirical parameters," *Nature Struct. Biology*, 1994.

[5] D. Mattes and L. de Groot, "Secondary structure propensities in peptide folding simulations: A systematic comparison of molecular mechanics interaction schemes," *Biophysical Journal*, vol. 97, pp. 599–608, 2009.

[6] V. Munoz and L. Serrano, "Elucidating the folding problem of helical peptides using empirical parameters ii: Helix macrodipole effects and rational modification of the helical content of natural peptides," *J. Mol. Biology*, 1994.

[7] V. Munoz and L. Serrano, "Elucidating the folding problem of helical peptides using empirical parameters iii: Temperature and ph dependence.," *J. Mol. Biology*, 1994.

[8] V. Munoz and L. Serrano, "Development of the multiple sequence approximation within the agadir model of a-helix formation. comparison with zimm-bragg and lifson-roig formalisms.," *Biopolymers*, vol. 41, pp. 495–509, 1997.

[9] E. Lacroix, A. Viguera, and L. Serrano, "Elucidating the folding problem of a-helices: Local motifs, long-range electrostatics, ionic strength dependence and prediction of nmr parameters.," *J. Mol. Biology*, vol. 284, pp. 173–191, 1998.

[10] "C-term and N-term definitions." http://academic.brooklyn.cuny.edu/biology/bio4fv/page/c_and_n_.htm.

[11] G. P. Smith, "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface," *Science*, vol. 228, pp. 1315–1317, 1985.

[12] Z. F. Zang, "Screening and selection of peptides specific for esophageal cancer cells from a phage display peptide library," *Journal of Cardiothoracic Surgery*, vol. 9, 2014.

[13] P. A. C. t. Hoen, "Phage display screening without repetitious selection rounds," *Analytical Biochemistry*, vol. 421, pp. 622–631, 2012.

[14] "Correlation value information." `http://www.stat.wmich.edu/s216/book/node122.html`.