Exploring the Functional and Structural Topology of Synthetic DNA

by

Symon Benjamin Levenberg

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2021 by the
Graduate Supervisory Committee:

Neal Woodbury, Co-Chair
Kevin Redding
Chad Borges
Giovanna Ghirlanda

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Exerting bias on a diverse pool of random short single stranded oligonucleotides (ODNs) by favoring binding to a specific target has led to the identification of countless high affinity aptamers with specificity to a single target. By exerting this same bias without prior knowledge of targets generates libraries to capture the complex network of molecular interactions presented in various biological states such as disease or cancer. Aptamers and enriched libraries have vast applications in bio-sensing, therapeutics, targeted drug delivery, biomarker discovery, and assay development. Here I describe a novel method of computational biophysical characterization of molecular interactions between a single aptamer and its cognate target as well as an alternative to next generation sequencing (NGS) as a readout for a SELEX-based assay. I demonstrate the capability of an artificial neural network (ANN) trained on the results of screening an aptamer against a random sampling of a combinatorial library of short synthetic 11mer peptides to accurately predict the binding intensities of that aptamer to the remainder of the combinatorial space originally sampled. This machine learned comprehensive non-linear relationship between amino acid sequence and aptamer binding to synthetic peptides can also make biologically relevant predictions for probable molecular interactions between the aptamer and its cognate target. Results of SELEX-based assays are determined by quantifying the presence and frequency of informative species after probing patient specimen. Here I show the potential of DNA microarrays to simultaneously monitor a pool of informative sequences within a diverse library with similar variability and reproducibility as NGS.

DEDICATION

I would first like to dedicate this work to my given family. My parents Susan and Sam have provided the most nurturing and supportive environment for me to navigate my personal and professional life. I love them both and would not have made it this far without them putting up with my antics for the last 30 years. My sister Claire deserves equal recognition for her support and love. Thank you for opening your home and providing a regular escape when I needed a break.

In addition to my given family, I dedicate this work to my chosen family. To David Sydiongko, Evan Balbona, Ruben Favaro, Mike Carungi, Mario Mendez, Jesse Davenport, Maria Balderas, Kelly Smith, Skipper Arnold, Natalie Goldfarb, Naomi Newman, and the rest of the Vine Reunion crew. I miss and love all of you. To my friends in and around Dirty Epic, thanks for teaching me how to party proper. To Miquah James, thank you for showing up early, staying late, getting weird, and keeping me present and accountable. To Jesse Frank, thanks for every record, disco ball, and sweaty dance floor. Thank you to the cycling crew, Team Cretins, Jesse Schlaefer, Chris Hildreth, Max Haggard, and every one of my teammates on the Cobra Arcade Bar/ Heavy Pedal team for being role models, friends, and going fast on bikes together. To my internet bestie and party girl party girl TJ Davis. You're great dude, you bring so much joy into my life.

And finally I would like to dedicate this work to the most important person in my life, Hayley Andersen. No need to explain further. Thanks bud!

ACKNOWLEDGEMENTS

My deepest gratitude to Caris Life Sciences and David Spetlzer for fostering a collaborative relationship with Arizona State University. This partnership between academia and industry has given me unparalleled resources to complete my research, and I am incredibly thankful. I would also like to acknowledge my co-chair Neal Woodbury. He has been an exceptional mentor, I strive to one day be as proficient a scientist and researcher as him. Thanks also to my colleagues at Caris Life Sciences; Jim Abraham, Anthony Helmstetter, Dan Martin, Varun Maher, Radhika Santhanam, Patrick Kennedy, Gerri Ortiz, Adam Stark, Xixi Wei, Matthew Rosenow, Teresa Tinder, Heather O'Neil, Mark Miglarese, Valeriy Domenyuk, Tassilo Hornung

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1. INTRODUCTION 1

1.1 Overview of systematic evolution of ligands exopnentially 1

1.1.1 Library Design 2

1.1.2 Sample selection and target immobilization 3

1.2 Library Sequencing 3

1.2.1 Sanger sequencing 3

1.2.2 Next generation sequencing 5

1.3 NGS applications in SELEX 9

1.4 Characterization of aptamers 9

1.4.1 Secondary structure 9

1.4.2 Target identification and binding site determination 10

1.5 Aptamer applications 11

1.5.1 Diagnostics 13

1.5.2 Therapeutics 13

1.6 Scope of thesis 14

1.6.1 Aptamer binding characterizations through machine learning and combinatorial chemistry 15

CHAPTER Page

1.6.2 Readout methods comparison of NGS and DNA microarrays for SELEX-based assays 15

2. PREDICTING MOLECULAR RECOGNITION BETWEEN PEPTIDES AND APTAMERS THROUGH COMBINATORIAL CHEMISTRY AND MACHINE LEARNING 16

2.1 Abstract 16

2.2 Introduction 16

2.3 Materials and Methods 18

2.3.1 Aptamer binding to peptide arrays 18

2.3.2 Neural network structure 20

2.4 Results and Discussion 22

2.4.1 Aptamer binding to peptide arrays is concentration dependence 22

2.4.2 Aptamer binding to peptide arrays is reproducible 23

2.4.3 Differential binding of aptamers to peptide arrays 24

2.4.4 Binding values predicted by neural networks have strong correlation to measured binding values 25

2.4.5 Impact of training steps and training set size on neural network predictive capabilities 27

2.4.6 Prediction varaiblity due to random selection of training set 28

2.4.7 Neural network ability to predict aptamer specificity 30

2.4.8 Machine learned characteristics of amino acids 31

CHAPTER Page

2.4.9 Probability of binding to amino acids in cognate targets based on neural network binding value predictions 32

2.5 Conclustion 37

3. METHODS COMPARISON OF NEXT GENERATION SEQUENCING AND DNA MICROARRAYS FOR MONITORING POOLS OF APTAMERS IN A SELEX-BASED BREASTCANCER DIAGNOSTIC LIBRARY 40

3.1 Abstract 40

3.2 Introduction 40

3.3 Methods 45

3.3.1 Libray Sequencing 45

3.3.2 DNA microarray content design 45

3.3.3 Array hybridiation 47

3.3.4 Initial washing 47

3.3.5 Secondary detection 47

3.3.6 Secondary washing 48

3.3.7 Scanning and feature extraction 48

3.4 Results 48

3.4.1 Pure library sequencing of L2000 by NGS 48

3.4.2 L2000 hybridization to complementary probes on a DNA microarray 52

3.4.3 Reproducability and variability of microarrays for detection of SELEX libraries 53

CHAPTER Page

3.5 Discussion 54

3.5.1 Variability and reproducability of NGS platform for monitoring SELEX libraries of varying diversity 54

3.5.2 L2000 hybridization to DNA microarray is concentration dependent 55

3.5.3 Variability and reproducability of DNA microarrays for monitoring SELEX libraries of varying diversity 55

3.5.4 Relationship between fluorescene signal from DNA microarray hybridization and copy number determined from NGS 55

3.6 Conclusion 57

3.7 Supplmental Information 58

4. A DNA-DIRECTED LIGHT HARVESTING AND REACTION CENTER SYSTEM 59

4.1 Abstract 59

4.2 Introduction 60

4.3 Results and discussion 63

4.3.1 Assembly of Light-Harvesting/Reaction Center Complex 63

4.3.2 Excitation and Energy Transfer Efficiency 68

4.3.4 Enhancement of Reaction Center Charge Separation 74

4.4 Conclusion 77

4.5 Supplemental Material 77

4.5.1 Reaction Center Protein Preparation 77

4.5.2 RC-DNA Conjugation and Purification 80

4.5.3 DNA-dye Conjugation and Purification 81

4.5.4 3arm-RC Preparation 82

4.5.5 Spectrocopic Analysis 83

5. CONCLUSIONS AND OUTLOOK 101

REFERENCES 102

LIST OF TABLES

Table Page

1.1 Starting library design for enrichment that lead to L2000 library 3

1.2 Comparison of critical features of aptamers and antibodies 13

1.3 Current approval and clinical trial status of aptamer-based therapeutics 15

2.1 Summary of aptamers included 21

2.2 Summary of aptamer's binding signature replicates 26

2.3 Variability of neural network generated test predictions for different training sets 32

3.1 Summary of experimental and measured sequencing conditions 55

4.1 3arm-to-RC ratio of different constructs 70

4.2 Fitting parameters for the Cy3 lifetime data in different constructs 77

LIST OF FIGURES

Figure Page

1.1 Scheme of the SELEX process 2

1.2 Illustration of Sanger sequencing workflow 5

1.3 Cluster generation on the Illumina sequencing platforms 8

1.4 Sample flowcell TIFF scans after one cycle of base incorporation 9

1.5 MS/MS spectra for the C1036 binding site on target protein hnRNP U 12

2.1 Summary of workflow and data processing 23

2.2 C1036 binding to peptide microarrays at varying concentrations 25

2.3 Aptamer binding signature similarity shown as heatmap 27

2.4 Neural network predicting binding versus measured binding scatter plots for all peptides in the test set of each aptamers model 29

2.5 Correlation between measured and predicted binding of the test set as a function of number of training steps and as a function of training set size 31

2.6 Scatter plots illustrating a neural network's capability to learn and predict aptamer binding specificity 33

2.7 Amino acid similarity matrix based on machine learned characteristics 35

2.8 Neural network-generated probability of C1036 binding to each amino acid in the cognate target (hnRNP U) sequence 37

2.9 Neural network-generated probability of Rt1 binding to each amino acid in the cognate target (HIV-1 reverse trancriptase) sequence 38

2.10 Neural network-generated probability of 2008s binding to each amino acid in the cognate target (*Pf*LDH) sequence 39

Figure Page

2.11 Neural network-generated probability of TBA binding to each amino acid in the cognate target (TBA) sequence 40

2.12 Neural network-generated probability of ARC1172 binding to each amino acid in the cognate target (VWF a1) sequence..... 41

3.1 Schematic of poly-ligand profiling workflow..... 46

3.2 Visualization of DNA microarray layout and feature packing..... 51

3.3 Gel confirmation of amplification with Illumina adapter sequences 54

3.4 Violin plots illustrating varaibility across PCR reaction and across sequencing runs... 56

3.5 Array hybridization summary..... 58

4.1 Modified structure of reaciton center construct with corresponding absorption spectra 65

4.2 FPLC purification trace for RC-DNA conjugation..... 69

4.3 Absorption spectra of 3arm-DNA-dye-RC constructs 71

4.4 Fluorescence emission spectra of 3arm-DNA-dye-RC constructs..... 73

4.5 Energy transfer efficiency of 3arm-DNA-dye-RC constructs..... 75

4.6 Light minus dark spectra of RC with and without DNA-dye constructs 80

4.7 Cytochrome c oxidation monitored at 550nm 82

CHAPTER 1

INTRODUCTION

1.1 Systematic Evolution of Ligands Exponentially Overview

Aptamers are short single stranded DNA or RNA ligands with highly specific binding to their target. Their versatility in analytics, diagnostics, and therapeutics has been demonstrated ad nauseum since the introduction of systematic evolution of ligands by exponential enrichment (SELEX) in the 1990s. While there have been improvements to the speed and efficiency of SELEX over the past three decades, the conceptual process remains unchanged. SELEX is a bottom up iterative artificial selection process that reduces the randomness of a highly diverse library of oligonucleotides, while simultaneously favoring selection of sequences with high binding affinity for the target of the selection [1]. The SELEX process consists of repeated rounds of incubating a library of single stranded oligonucleotides (ssONs) with the selection target (binding), isolation

of bound sequence and removal of any unbound sequences (partitioning), and amplification of bound sequences through PCR (Figure 1.1).

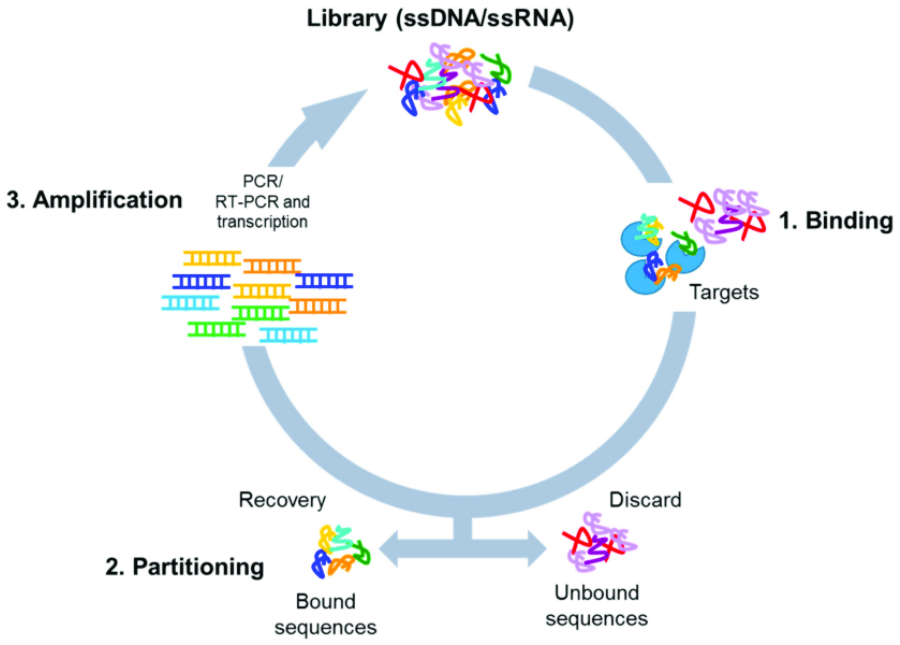


Figure 1.1: Scheme of the SELEX process. The procedure involves repeated cycles of: 1. Incubation of the high complexity library with the targets (binding); 2. Removal of unbound sequences and recovery of the bound oligonucleotides (partitioning); 3. Amplification of the bound sequences by PCR (for DNA library) or RT-PCR and transcription (for RNA library) [2].

Repeated rounds of enrichment following this scheme influences library content to shift from high diversity low copy number to lower diversity with higher average copy number per species. Interrogating the information content stored in the enriched library is what leads towards nomination of a sequence as an aptamer [3].

1.1.1 Library Design

Designing the randomized library is the initial step prior to any SELEX. While initial randomized nucleic acid libraries vary in design and nucleotide sequence depending on application, there are three domains highly conserved across all library design; a variable region, to introduce the necessary randomness, and two primer regions for amplification between subsequent enrichment rounds as well as downstream sequence identification and analysis. Figure 1.2 shows an example of the library design for the starting library used in an enrichment to differentiate between plasma from biopsy positive and biopsy negative breast cancer patients [4].

Table 1.1: Design of the library used in the enrichment from Domenyuk et al. 2017.

Synthesis	Sequence	Length	%GC in variable region
IDT Ultramer	5' CTAGCATGACTGCAGTACGT-35N-CTGTCTCTTATACACATCTGACGCTGCCGACGA 3'	88	50%
	5' CTAGCATGACTGCAGTACGT-35N-ACTGTCTCTTATACACATCTGACGCTGCCGACGA 3'	89	50%
	5' CTAGCATGACTGCAGTACGT-35N-GACTGTCTCTTATACACATCTGACGCTGCCGACGA 3'	90	50%
	5' CTAGCATGACTGCAGTACGT-35N-TGACTGTCTCTTATACACATCTGACGCTGCCGACGA 3'	91	50%

The 35 base long variable region is depicted by the “35n” in the center of each sequence. Primer regions are indicates with blue and red text, with the blue primer region containing a sequence complimentary to the Illumina sequencing primers, a common NGS sequencing platform. Including the sequencing primer sequence allows for the removal of one of the multiple PCR steps normally included in library preparation for sequencing.

A fully random library, with even sequence motif dispersal, incorporates all 4 nucleotides in equimolar amounts. Starting libraries with equal probability for adenine, guanine, cytosine, and thymine to be incorporated into any given spot maintains a high diversity non-biased starting library. Biased libraries utilize constant nucleotides placed

at specific intervals and locations to insert common secondary structural motifs. These biased libraries are beneficial if trying to isolate an aptamer for a target with that particular secondary structural motif, but including these biases decrease the diversity of the starting library by orders of magnitude [5].

1.1.2 SELEX sample selection and target immobilization

1.2 Library sequencing

Assessing enriched library performance, and identifying aptamer targets involved a two pronged attack. The first of which is interrogating the entire pool to isolate species that favor on-target binding over off-target binding, or bind exclusively to the target to identify their sequence for downstream characterization. DNA sequencing is paramount to monitoring SELEX progress. Sequencing platforms have improved drastically in their cost, efficiency, and depth of coverage since the implementation of SELEX [6].

1.2.1 Sanger Sequencing

In traditional aptamer research, in order to know the primary sequence of informative monoclonal aptamers, the enriched pool of sequences needs to be PCR amplified, cloned into an appropriate vector, and subject to Sanger sequencing[7]. In traditional Sanger sequencing, target sequences are amplified with a small amount of fluorescently labeled chain-terminating deoxynucleotides spiked into the standard dNTPs. Chain-terminating dNTPs lack the 3’-OH group required for phosphodiester bond formation. This generates PCR product with the DNA sequence of interest being terminated at random lengths. In manual Sanger sequencing, four separate PCR reactions are set up, with only a single chain-terminating dNTP (ddATP, ddTTP, ddCTP, or

ddGTP) used in each reaction. In automated Sanger sequencing, all four chain-terminating dNTPs are labeled with a different fluorophore, and mixed in a single reaction. The PCR product with target DNA amplified to random lengths is then separated by size using capillary gel electrophoresis, and the terminal nucleotide of each fragment determined by its fluorescence (Figure 1.3). Determining aptamer candidates by sequencing in this fashion only represents the diversity of tens to hundreds of clones within the enriched pool of sequences. With number of sequences present in enriched libraries ranging in the millions of species, the majority of the information stored in the enriched library is lost. Despite a successful SELEX, historic sequencing measures provide low probabilities for identifying the top performing species in an enriched library[8].

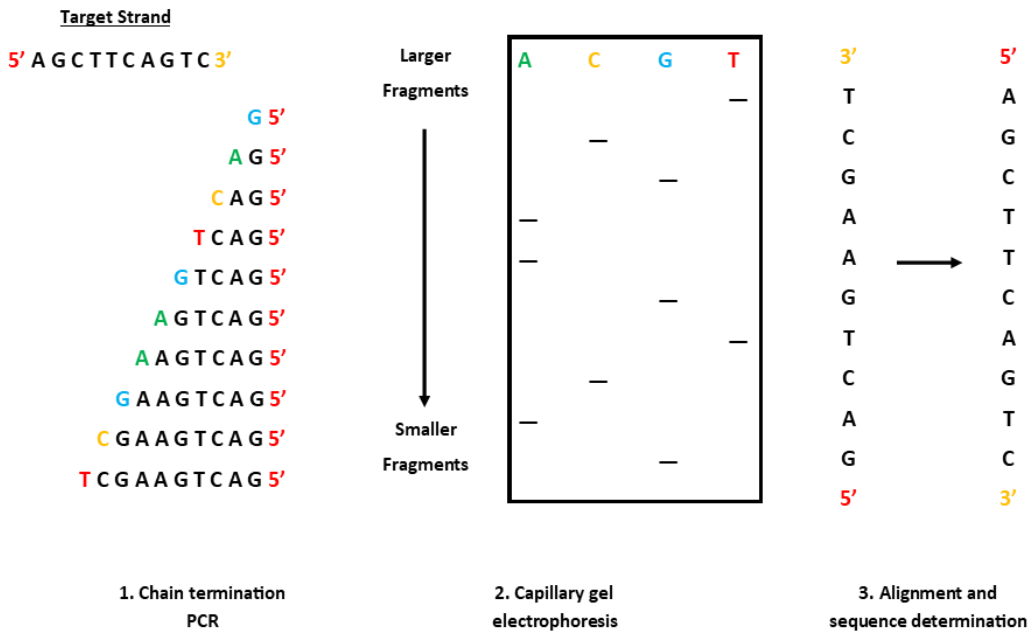


Figure 1.2: Generic scheme for Sanger sequencing. 1) Chain termination PCR amplification of target strand. Colored bases indicate the fluorescently labeled ddNTP. 2)

Size separation of PCR fragments using capillary gel electrophoresis. The largest fragment, fully polymerized target strand with chain terminating at the last base, will run the slowest through the gel. The smallest fragment, chain terminated at the first base, will run the fastest through the gel. 3) Resulting DNA sequence is determined from reading the gel from the top down, with the target sequence being the reverse compliment of the sequence read from the gel[8].

1.2.2 Next Generation Sequencing

The advent of high throughput sequencing (HTS) technologies, also commonly referred to as Next-generation sequencing (NGS) vastly improved monitoring SELEX progress. The depth of coverage for a sample through NGS can provide millions of reads for sequences between 50 and 300 bases in a single run. NGS is responsible for the massive acceleration of the Human Genome Project, shortening the time for sequencing a whole human genome from a decade to a day [9]. The efficiency of NGS platforms has transformed DNA sequencing into routine clinical practices and lab research [10]. One of the most prominent sequencing platforms to emerge in the clinical and research space is Illumina, who's sequencing technology coined "sequencing by synthesis" (SBS) will serve as the example in this work for understanding the NGS technology available today. The Illumina workflow for sequencing is broken down into four main processes; sample preparation, cluster generation, sequencing, and data analysis [11].

1.2.2.1 Sample Preparation

All current sequencing platforms require some form of pre-processing of DNA material to generate a form of the library compatible for sequencing. While the workflow

for NGS was developed for genomic DNA, the same principles can be applied to sequencing any nucleic acid material, i.e. libraries of aptamers. After the DNA is isolated (and fragmented if working with genomic DNA), adapters containing specific adapter sequences necessary for the Illumina workflow are added to both ends of each fragment in the library. This can be done either through ligation, or through PCR using forward and reverse primers containing the specific Illumina adapter sequences [8].

1.2.2.2 Cluster generation

Once the sequencing library has the necessary adapter sequences attached to both ends of the fragments, the sample is loaded onto a flowcell. Illumina flowcell surfaces are functionalized and coated in probes that are complimentary to the adapter sequences on both ends of the sequencing library. Single copies of each sequence in the library are not adequate for signal detection, clonal cluster generation through bridge amplification is used to amplify signal. The functionalized surface of the flowcell is covered in probes that are complementary to the forward and reverse primer regions of the sequencing adapters. The term “bridge” in bridge amplification is due to the bridge formed between forward and reverse probes by the sample on the flowcell surface during cluster generation [8]. Figure 1.4 shows a schematic and results of cluster generation.

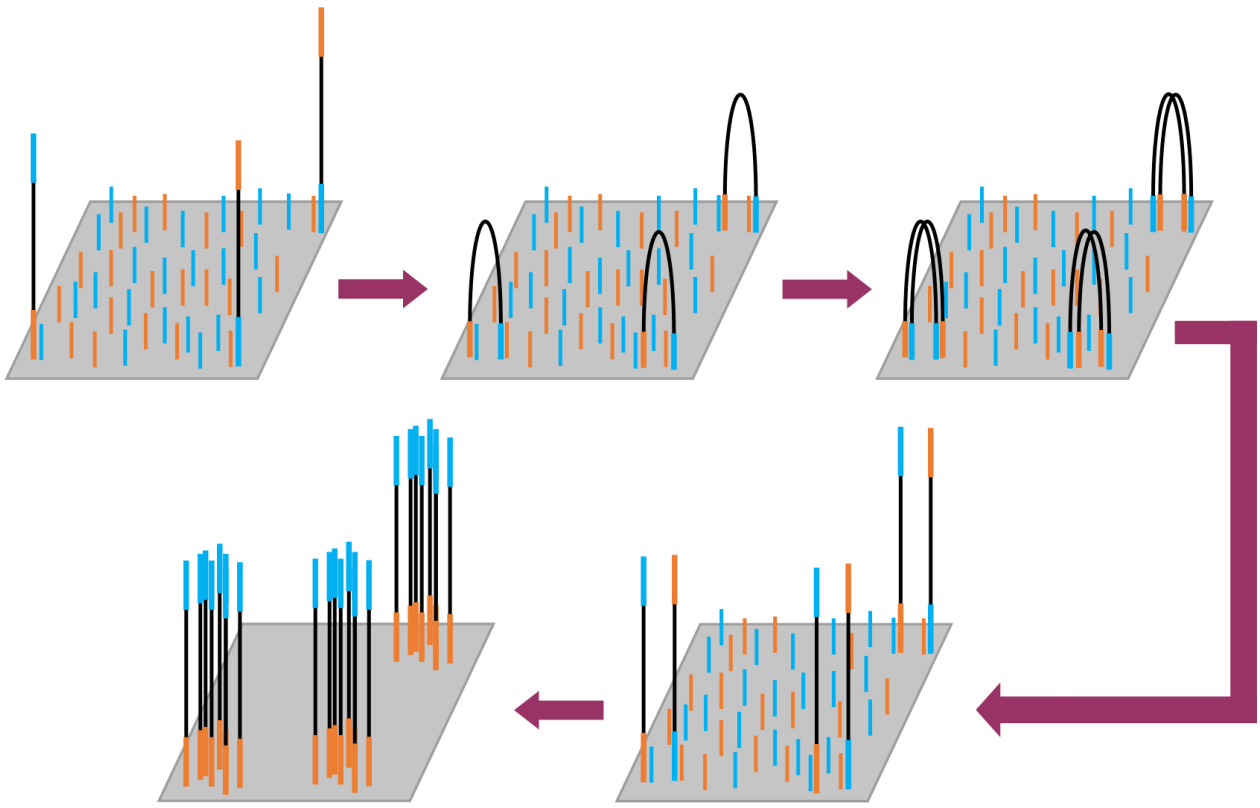


Figure 1.3: Schematic of cluster generation on Illumina flowcells. Sample containing Illumina sequencing adapters anneal to forward and reverse primers immobilized on the flowcell's surface. Several cycles of bridge amplification occur, generating dense clusters containing multiple copies of a given sequence.

1.2.2.3 Sequencing by synthesis

After cluster generation, sequencing proceeds by incorporating one base at a time, with each of the 4 bases being labeled with a unique fluorophore. Sequence of each cluster is determined by imaging the flowcell after each base incorporation. Figure 1.4 shows a sample of cluster imaging.

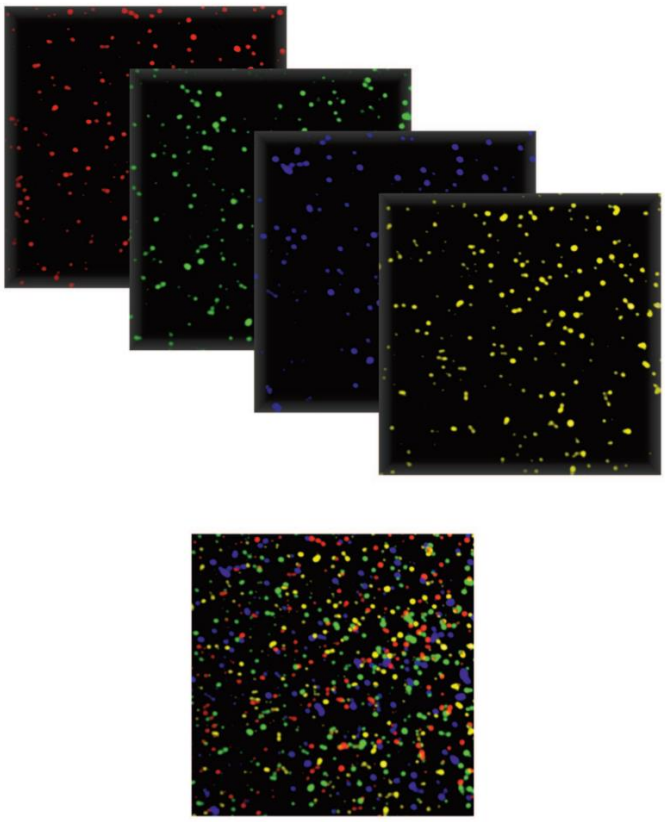


Figure 1.4: Pseudo color image generated from an Illumina flowcell. Each fluorescence signal originates from a clonally amplified cluster. The top panel shows the 4 different emission wavelengths for the incorporation of the 4 different nucleotides. The lower panel shows a composite image of the 4 fluorescence channels [9].

1.3 NGS applications in SELEX

Sanger sequencing could not adequately cover the sequence space possible in an enriched library, let alone monitoring the SELEX process concurrently. The efficiency and depth of data generated from NGS makes it a prime tool for characterizing the SELEX progress. Library diversity, being one indicator of SELEX progress, can be easily monitored using NGS. NGS also increases the probability of identifying high performance aptamer candidates just based on sheer numbers, being able to identify millions of sequences rather than only the hundreds accessible by Sanger sequencing [11].

1.4 Characterization of Aptamers

Once identified via sequencing, characterization is the next step in the aptamer life cycle. Assessing aptamer binding affinity and specificity to its enriched target, as well as biophysical characterization is important for downstream development of assays or therapeutics. Sequence alignment and grouping into families/categories narrows the scope of candidates for further characterization. Open source and web-based tools like ClustalW (http://www.genome.jp/tools/clustalw) provide accessible tools for sequence alignment [3].

1.4.1 Secondary Structure

Aptamer tend to form well defined structures based on complementary base pairing within their sequence [12]. This folding onto itself gives rise to secondary structure motifs like stems, loops, pseudoknots, G-quadraplexes, and hairpins [13, 14]. Circular dichroism (CD) is the most common biophysical technique for determining

structure of aptamers. CD probes structure by measuring absorption changes of circularly polarized light. The various secondary structure motifs effect known unique quantifiable changes in absorption. The sensitivity and robustness of the technique also allows for analyzing structural changes in response to changes in buffer conditions such as ion concentrations, acids, bases, alcohols [3].

1.4.2 Target Identification and binding site determination

SELEX primarily occurs against a known target. The libraries used in SELEX-based diagnostic assays are enriched without prior knowledge of the target or targets. Their utility lies in the binding signature of the library as a whole to differentiate between biological states. While the assay read-out is independent of target list, the binding partners of species within the library offer potential novel biomarkers, and actionable targets for therapeutics. Modified immunoprecipitation (using aptamers as the capture molecule instead of antibodies) followed by mass spectrometry has been used to identify the unknown targets of aptamers [4]. Aptamer-target molecular binding interfaces are characterized through aptamer-based affinity labeling (ABAL), photocleavable crosslinking, and protected trypsin digestion [15]. Figure 1.6 shows the MS/MS spectra of the trypsin-digested peptide containing the molecular weight tag transferred to the

aptamer binding site from a crosslinking experiment between heteronuclear ribonuclear protein U (hnRNP U) and it's aptamer C1036 [16].

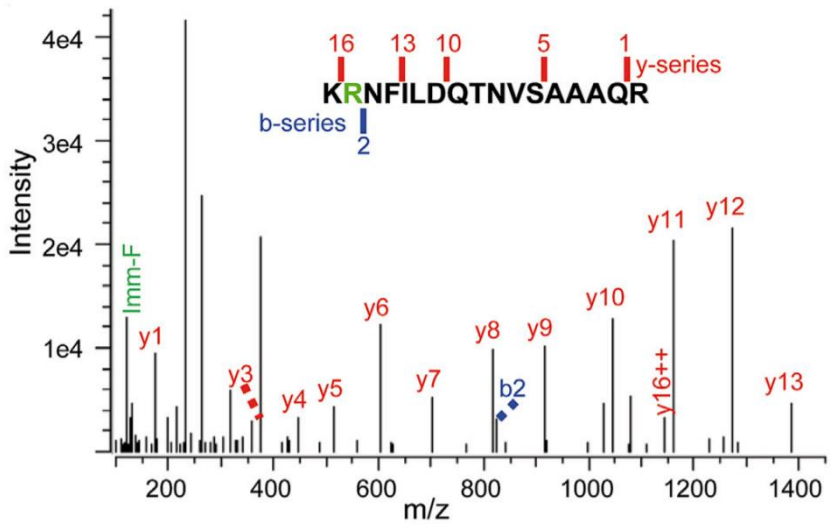


Figure 1.5: MS/MS spectra of a peptide of hnRNP U with the TMT-SDAD (molecular weight tag) label identifying the crosslinking site between the protein and aptamer C1036 at arginine 575 (highlighted in green) [16].

1.5 Aptamer Applications

Aptamers are analogous to monoclonal antibodies in the sense that they recognize and bind a specific target. With robust molecular recognition of their targets, aptamers have wide range of diagnostic and therapeutic applications as well as use in diverse research areas within biochemistry and medicine. Aptamers present an ideal alternative to antibodies. They are much smaller in size (6 - 30kDa, 2nm) compared to antibodies (150 – 180kDa, 15nm), increasing the potential for molecular recognition of smaller molecules and binding domains inaccessible to antibodies [12]. Their size however does contribute

to short in vivo circulation time and kidney filtration. A major benefit of aptamers over antibodies is the amount of time needed for generation and development is significantly shorter. Generation of antibodies is time intensive and expensive, and targets must trigger a strong enough immune response. The cost and generation time for aptamers is significantly less. Targets for aptamer selection are not limited to molecules initiating a strong immune response including toxic molecules [9] [12]. Table 1.1 shows a comparison summary of pros and cons between aptamers and antibodies.

Table 1.2: Comparison of critical features of aptamers and antibodies [17]

	Aptamers	**Antibodies**
Stability	Withstand repeated rounds of denaturation/renaturation. Temperature resistant: stable at room temperature. Long shelf life (several years). Can be lyophilized. Degradable by nucleases. Resistant to proteases.	Easily denatured. Temperature sensitive and require refrigeration to avoid denaturation. Limited shelf life. Must be refrigerated for storage and transport. Degradable by proteases. Resistant to nucleases.
Synthesis	In vitro SELEX takes only 2–8 weeks. No batch-to-batch variation. Cheap to synthesize.	Produced in vivo. More than 6 months. Batch-to-batch variations. Laborious and expensive.
Target potential	From ions and small molecules to whole cells and live animals.	Targets must cause a strong immune response for antibodies to be produced.
Size	Small molecules.	Relatively large by comparison.
Modifiability	Aptamers can readily and easily be modified without affinity loss.	Modifications often lead to reduced activity.
Affinity	High and increased in multivalent aptamers.	Dependent on the number of epitopes on the antigen.
Specificity	Single point mutations identifiable.	Different antibodies might bind the same antigen.
Tissue uptake/kidney filtration	Fast.	Slow.

1.5.1 Diagnostics

Diagnostic capabilities of aptamers are extensive. Aptamer-based biosensors have been used to detect bacterial, viral, and parasitic pathogens; as well as detecting environmental contaminants and food safety [3]. Aptamers have also been developed to detect cancer specific biomarkers and tumor associated cell surface proteins in living cancer cells [18] [19]. Aptamers have infiltrated the cancer imaging and diagnostic world as well. High sequence-diversity pools of aptamers have shown to differentiate between biopsy-positive and biopsy-negative breast cancer through liquid biopsy of 500 patients plasma. Enriched libraries of aptamers have demonstrated differential staining of cancer and normal tissue in patient FFPE tissue slides, and have even been developed into clinical companion diagnostics for determining patient response or non-response to chemotherapy, e.g. trastuzamab [4] [20].

1.5.2 Therapeutics

Aptamers have demonstrated ability to compete with small molecules and receptor ligands, inhibit targets, activate receptor function, and initiate cell internalization mechanisms making them ideal to act as or deliver therapeutic agents [21]. Table 1.2 shows a list of DNA and RNA aptamers both modified and unmodified and their current status as therapeutics [22].

Table 1.3

Name (company)	Composition	Target	Indication	Current phase
Pegaptanib sodium/Macugen (Pfizer/Eyetech)	2′-*O*-methyl purine/2′-fluoro pyrimidine with two 2′-ribo purines conjugated to 40 kDa PEG, 3′ inverted dT	Vascular endothelial growth factor	Age-related macular degeneration	Approved in the US and the EU
AS1411/ AGRO001 (Antisoma)	G-rich DNA	Nucleolin	Acute myeloid leukaemia	Phase II
REG1/RB006 plus RB007 (Regado Biosciences)	2′-ribo purine/2′-fluoro pyrimidine (RB006)/40 kDa PEG plus 2′-*O*-methyl antidote (RB007)	Coagulation factor IXa	Percutaneous coronary intervention	Phase II
ARC1779 (Archemix)	DNA and 2′-*O*-methyl with a single phosphorothioate linkage conjugated to 20 kDa PEG, 3′ inverted dT	A1 domain of von Willebrand factor	Thrombotic microangiopathies and carotid artery disease	Phase II
NU172 (ARCA biopharma)	Unmodified DNA aptamer	Thrombin	Cardiopulmonary bypass to maintain steady state of anticoagulation	Phase II
ARC1905 (Ophthotech)	2′-ribo purine/2′-fluoro pyrimidine conjugated to 40 kDa PEG, 3′ inverted dT	Complement component 5	Age-related macular degeneration*	Phase I
E10030 (Ophthotech)	DNA and 2′-*O*-methyl 5′-conjugated to 40 kDa PEG, 3′ inverted dT	Platelet-derived growth factor	Age-related macular degeneration*	Phase I
NOX-A12 (NOXXON Pharma)	L-RNA with 3′-PEG	CXCL12	Multiple myeloma and non-Hodgkin's lymphoma‡	Phase I
NOX-E36 (NOXXON Pharma)	L-RNA with 3′-PEG	CCL2	Type 2 diabetes, diabetic nephropathy	Phase I

1.6 Scope of Thesis

I have had the unique opportunity to conduct the research in a collaborative partnership between academia and industry. This has given me unparalleled access to clinical samples, data sets, and equipment not traditionally available in an academic research lab setting. The resources available to me as allowed me to explore the complex rules of self-folding and molecular recognition capabilities of the SELEX platform, as well as exploring Scientists using SELEX have carved out an invaluable niche in biomarker discovery, bio-sensing, diagnostics, and therapeutics. My aim with this thesis

is to demonstrate the ability to build structural systems using the simple rules of DNA, leverage the complex rules of secondary structures and molecular recognition through SELEX, and use machine learning to build new rules between structure and function.

1.6.1 Aptamer binding characterization through machine learning and combinatorial chemistry

The scope of the second chapter of this thesis is to describe a novel tool for characterizing the molecular interactions between aptamers and their targets. Most techniques rely on interrogating the intact aptamer-target complex for identification of binding interface. The first aim of this paper is to provide evidence that machine learning informed with combinatorial library screening data forms a comprehensive understanding of molecular interactions between synthetic peptides and aptamers, and can also make meaningful predictions about probable interaction sites on a target protein.

1.6.2 Assessment of readout method variability/reproducibility and orthogonal detection method

Our lab's previous work as shown that poly-ligand profiling of plasma with a pool of 2000 informative sequences selected from libraries enriched on plasma from breast cancer patients can differentiate between biopsy positive and biopsy negative patient specimens. Sequences from the pool exhibit differential binding and subsequent amplification dependent on the two clinical phenotypes. This differential binding is captured and quantified via NGS to determine a sample's classification. The aim of chapter 3 is to assess the variability of copy number quantified by NGS across PCR

reactions and sequencing runs as well as describe an orthogonal method for simultaneous sequence detection and quantification.

CHAPTER 2

PREDICTING MOLECULAR RECOGNITION BETWEEN PEPTIDES AND APTAMERS THROUGH COMBINATORIAL CHEMISTRY AND MACHINE LEARNING

Symon Levenberg, Dan Martin, Anthony Helmstetter, Neal Woodbury, David Spetzler

2.1 Abstract

2.2 Introduction

Aptamers have become an invaluable tool in diagnostics, differentiating diseases states, determining potential of a patient to respond to chemotherapy, and a powerful fishing tool for discovering actionable targets in these systems[3] [4]. Historically, enriching for single aptamer- single target systems has led to the identification of powerful biomarkers, and helped study protein-DNA interactions, but this dogma doesn't capture the complexity and heterogeneity of complex biological systems. Disease states such as cancer are not defined by a single biomarker but rather a complex network of protein, DNA, RNA, and metabolite interactions. The traditional single aptamer – single protein enrichments lack sufficient complexity of molecular recognition to capture complex biological states. A recent trend has shown utility of pools or libraries of aptamers ranging from 10^3 to 10^6 species that have been enriched via the systematic evolution of ligands exponentially (SELEX) to differentiate between complex heterogeneous states such as disease vs healthy, cancer vs non-cancer, and drug responder vs. non-responder [4, 20]. The differential capabilities of complex aptamer libraries like these does not depend on prior knowledge of the individual binding partners

for every species. However, determining the binding partners of individual aptamers within enriched libraries provides a powerful platform for downstream target analysis and potential actionable drug targets [23].

The popular method of identifying the targets of these libraries involves extensive probing and pull-down experiments on patient samples and characterization of binding site via mass spectrometry [15]. In this paper I explore the potential for unsupervised learning algorithms informed in a combinatorial chemistry space to elucidate these nuanced aptamer-protein binding interfaces. Previous studies have employed machine learning algorithms on small subsets of peptide combinatorial spaces to develop a comprehensive understanding of protein-peptide interactions without having to physically probe the entire combinatorial space [24]. Machine learning has also been used to map RNA binding sites on RNA binding proteins, and predict possible epitopes of monoclonal antibodies [25, 26]. Here we explore the potential for machine learning can to derive comprehensive, quantitative, nonlinear relationships between the sequences of amino acids and their affinity for synthetic DNA aptamers based on screening a small subset of peptides from a combinatorial library. This relationship is invaluable for characterizing aptamers, requiring minimal chemical synthesis while still retaining the unbiased nature of the peptide combinatorial space.

Combinatorial chemistry was developed as an alternative to rational molecular design [27]. While molecular design relies on prior knowledge of chemical interactions, combinatorial chemistry searches for optimum performance within high dimensional landscapes independent of said knowledge [27]. The utility of combinatorial chemistry hinges on how well you can sample the space you're interested in probing. When looking

at combinatorial libraries it is important to consider the number of molecules, with varying arrangements of building blocks that need to be measured in order to develop a comprehensive, predictive, and representative model for the whole of that molecular space. In this approach we attempt to answer this question using a well-defined system of peptides made from 19 of the natural amino acids (cysteine was withheld to avoid inter peptide crosslinks) with lengths ranging from 7 to 17 residues, and average length of 11 residues. With average length of 11 residues, the possible combinatorial space is equal to ~10^{14} molecules. Approximately 125,000 of peptides within that combinatorial space are synthesized in an array format with their sequences selected as randomly as possible under the synthesis constraints. These peptides are immobilized as high density arrays on a silica substrate, for simultaneous screening and detection [28]. Here, binding values for each peptide in the array were measured to each of 9 fluorescently labeled DNA oligonucleotides, aptamers, incubated to the array. This binding data is used to inform an artificial neural network, with the trained network able to predict aptamer binding intensity based on primary amino acid sequence.

2.3 Materials and Methods

2.3.1 Aptamer Binding to Peptide Arrays

Peptide arrays were purchased from the Peptide Array Core (PAC), an Arizona State University company. Slides contain 24 arrays per slide laid out in 8 rows and 3 columns. Grid geometry and feature size are virtually identical for both manufacturers, but the peptide sequence content, amino acids, and synthesis method differ. Peptides on the PAC arrays contain 19 out of the 20 naturally occurring amino acids (cysteine is withheld), and average peptide length on the array is 12 amino acids long. Peptides are

synthesized on an epoxy functionalized silica surface semi-automatically in an oxygen-rich environment. Table 2.1 shows the aptamers used for this study.

Table 2.1: Summary of aptamers used for this study

ID	Target	Length	References
trCLN3	Hepatocyte growth factor receptor (cMET)	42	[23, 29]
V7t1	Vascular endothelial growth factor (VEGF)	25	[30]
IGA3	Insulin	30	[31]
AS1411	Nulceolin	26	[32]
Rt1	HIV-1 reverse transcriptase	37	[33, 34]
C1036	Heteronuclear ribonuclear protein U (hnRNP U)	36	[21] [16]
ARC1172	Von Willebrand factor A1	41	[35]
TBA	Thrombin	15	[36, 37]
2008s	Lactose dehydrogenase from *Plasmodium falciparum*	35	[38]

Aptamers were ordered with a 5'-AF546 and 18 carbon spacer modification from IDT. All aptamers were incubated at 95 C for 5 min and allowed to cool to room temperature in 1x phosphate buffered saline (PBS) with 3mM $MgCl_2$ to ensure proper folding before all assays. Arrays on each slide were isolated from each other by sandwiching the slides with a rubber gasket between two metal cassettes. After cassette assembly, arrays were pre-treated at room temperature for 1 hour with gentle mixing with 100uL of 1xPBS 3mM $MgCl_2$. 50uL of each aptamer diluted in 1xPBS $MgCl_2$ was added to each array. Aptamers incubated with arrays for 1.5 hours at room temperature with

gentle mixing. After incubation, the cassettes were disassembled while submerged in 1xPBS $MgCl_2$, and slides were washed by dipping 30 times in fresh 1xPBS 3mM $MgCl_2$. Slides were then dipped in isopropanol 10 times and dried with ultra-high purity N_2 gas. Arrays were scanned on an Innopsys innoscan 910AL, and feature extraction was performed using GenePix software.

2.3.2 Neural Network structure

Neural networks were constructed in python using the PyTorch machine learning libraries. All networks contained an input layer, 1 hidden layer with width of 100 neurons, and 1 output layer. All peptide sequences used for network training were first converted into a binary sequence matrix of with dimensions of length of longest sequence (17) on the array by 20 amino acids. Figure 1 shows a graphical representation of the work flow, including the captured TIFF image of the peptide array, the extracted peptide sequence and corresponding fluorescence value, and the generation of the binary matrices for input into the neural network architecture. All fluorescence intensities for each peptide were log transformed prior to any model training, model validation, or binding intensity prediction. Once each peptide sequence is represented by a binary matrix, the full dataset, represented by a list of sequence-binding pairs, is separated into a training set and a testing set. While the size of the training set, and the number of training steps vary across experiments, each training step consists of the same parts. In each training step, a batch of 200 sequence-binding pairs are selected at random from the training set. The linear transformed vector of each sequence's binary matrix is fed forward through the network, and the output (predicted binding) is calculated. The loss between the predicted binding value and measured binding value for all 200 sequence-binding pairs is

calculated, and the weight function of each of the nodes in each hidden layer are optimized to minimize the loss via backpropagation. SmoothL1Loss and Adagrad are used as the loss and optimizer functions respectively in all models[39] [40]. After the last training step, the sequence-binding pairs from the test set are fed forward through the trained model to generate predicted values for each sequence. Quality of the trained model is assessed via the Pearson correlation coefficient (R) between the predicted binding and measured binding of the test set. The number of sequence-binding pairs used in the training set, and the number of training steps were varied based on the experiment. Figure 2.1 depicts the workflow for the work describe here.

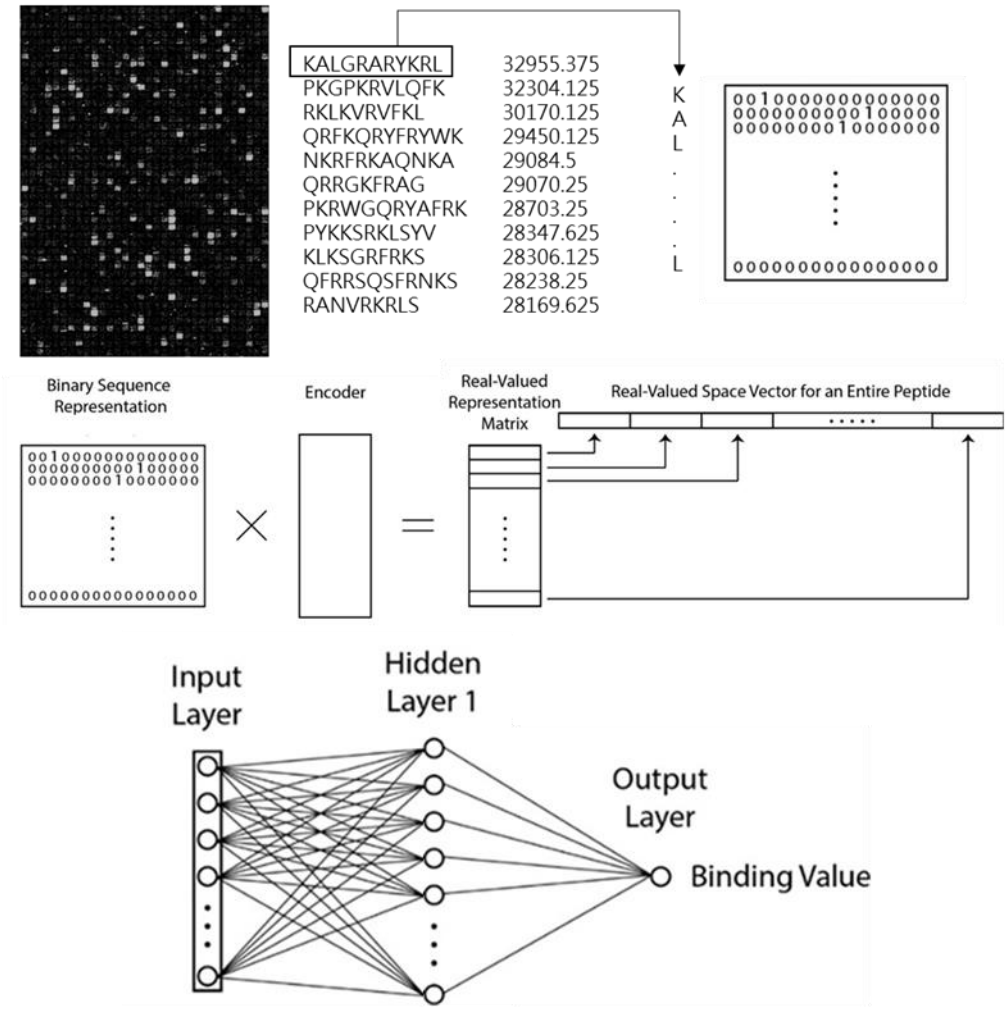


Figure 2.1: Workflow for data processing and neural network training. Starting with the raw TIFF image of a peptide array with an aptamer bound. Overlaying a map with the global positions of each peptide on the array on top of the TIFF image generates a list of peptides and corresponding measured fluorescence values. Each peptide sequence is converted into a binary matrix by 1 hot encoding each amino acid into a zero filled matrix with dimensions 19 X 17. Each matrix representation of a peptide sequence gets converted into a vector. The peptide sequences represented by vectors and their corresponding fluorescence values are then divided into the training set and test set for subsequent neural network training and validation.

2.4 Results and Discussion

2.4.1 Aptamer Binding to Peptide Arrays is Concentration Dependent

Fluorescently labeled C1036 was incubated with PAC arrays at 300, 60, 6, 0.6 nM concentrations, with 4 technical replicates for each concentrations. Arrays were scanned, feature extraction performed, and average fluorescence per peptide calculated for each replicate. Average fluorescence per peptide was averaged across technical replicates shown in figure 2.2A, with error bars representing standard error of the mean. A logarithmic regression was applied to concentration plotted against average fluorescence per feature, with an R^2 of 0.9401 shown in figure 2.2B. The optimal binding concentration determined from C1036 will be used for all further binding experiments.

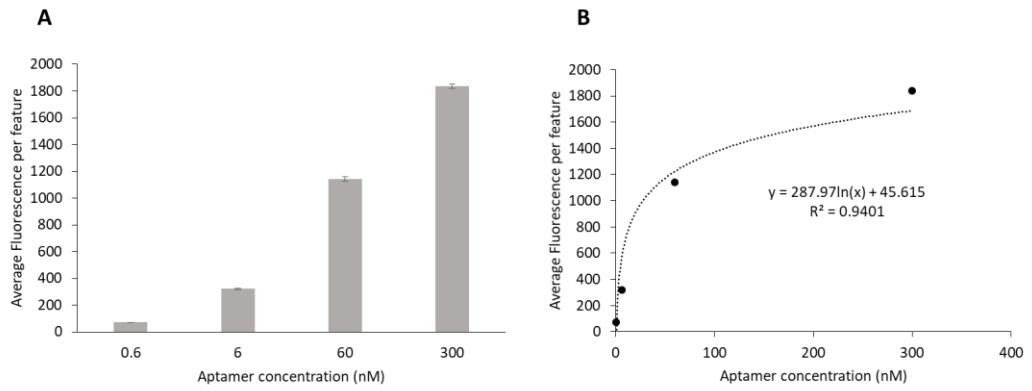


Figure 2.2: Fluorescently labeled aptamer, C1036, bound in varying concentrations to the 125K peptide microarray. A) Average fluorescence per feature was taken for the ~125,000 peptides on the array for each input, error bars are standard error of the mean. B) Logarithmic relationship between aptamer concentration and average fluorescence per feature with R^2 of 0.9401.

2.4.2 Aptamer Binding to Peptide Arrays is Reproducible

To determine reproducibility of binding signature for a given aptamer, 60nM of each aptamer was incubated on a PAC array with 4 technical replicates for each aptamer, accept for ARC1172 which used 3 technical replicates due to a large scratch across the surface of the 4th replicate's array. Arrays were scanned, and features extracted for each technical replicate. For each aptamer, a Pearson correlation coefficient (R) was calculated between every inclusive pair of replicates, and the average R was calculated with error for each aptamer. Table 2.2 shows average correlation between replicates with error for all aptamers, and includes a no aptamer control. The no aptamer control assay was performed in identical buffer conditions, incubation conditions, and washing conditions as the assays containing aptamer, sans aptamer.

Table 2.2: Summary of technical replicates for each sample, and the average R with error between linear binding values of every technical replicate pairwise of a particular sample. Assays were run under the same conditions.

Sample	Replicates	Average correlation between replicates w/ error
trCLN3	4	0.9807 ± 0.0111
V7t1	4	0.9342 ± 0.0390
IGA3	4	0.9847 ± 0.0105
AS1411	4	0.9769 ± 0.0088
RT1	4	0.9534 ± 0.0264
C1036	4	0.8673 ± 0.0528
ARC1172	4	0.8326 ± 0.0553
TBA	4	0.7846 ± 0.0712
2008s	4	0.8227 ± 0.0571
No Aptamer	4	0.0132 ± 0.0111

2.4.3 Aptamer Binding Signature on Peptide Arrays is Unique to the Individual Aptamer

To determine similarity of binding signatures across samples, peptides' fluorescence values were averaged across all technical replicates for each aptamer, generating a single list of ~125,000 sequence-binding pairs for each aptamer. Figure 2.3 shows a similarity matrix between samples as a heat map with color of the gradient indicating Pearson correlation coefficient between each samples' binding signature.

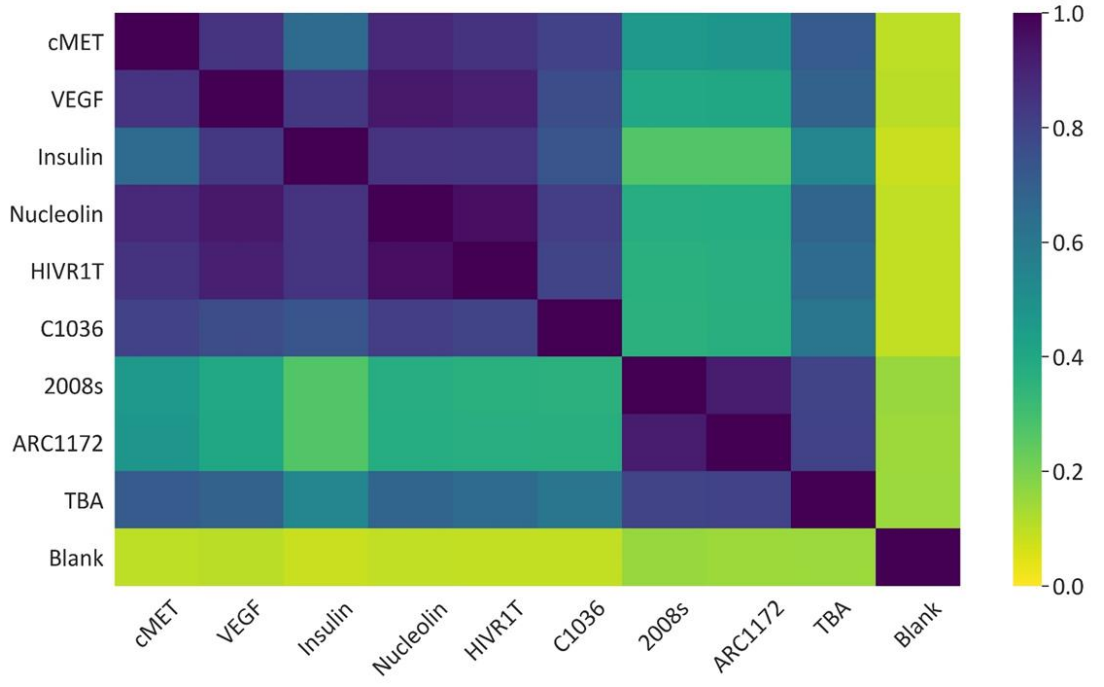


Figure 2.3: Comparison of measured binding across sample types. Heat map shows Pearson correlation coefficient for all peptide fluorescence values between samples. All Assays were run under the same conditions. Correlation is represented by a gradient on the right for each sample to every other sample, including to itself. The darker the blue, the stronger the similar the binding signatures between the pairs is.

2.4.4 Binding Values predicted by Neural Network Have Strong Correlation to Measured Binding Values

Each neural network (NN) was trained on one aptamer's binding signature. All networks used for this study are comprised of the same architecture: 1 hidden layer with width of 100 nodes. To train each neural network, 200 sequence-binding pairs were selected at random from the training set, and the NN was optimized based on the loss between the measured and predicted binding value for each pair. This act of calculating the loss and optimization is called a training iteration. This is repeated 20,000 times, this number of training iterations was chosen to ensure high probability that the network would encounter ever sequence-binding pair in the training set at least once during training. After 20,000 training steps, the test set of sequence-binding pairs were fed through the trained network, and their predicted binding values calculated. The NN predicted binding values were plotted against the measured binding values and the R calculated between predicted and measured for the full test set for each aptamer's NN (Figure 2.4)

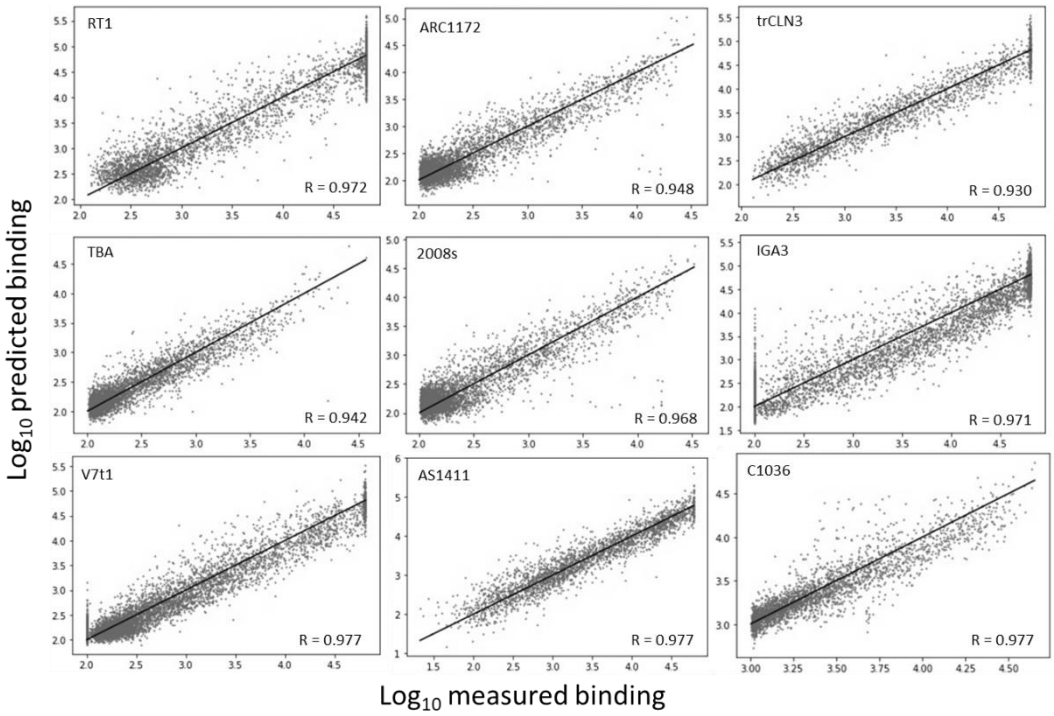


Figure 2.4: Log of predicted binding intensity plotted against log of measured binding intensity for the validation set of sequence binding pairs for individual neural networks trained on each aptamer's binding signature. Each network was trained on 90% of the sequence-binding pairs, the remaining 10% of sequences were fed through the trained model, and their predicted values plotted against the experimentally measured binding values. Each point in the scatter plots represents one peptide. Top left corner of each scatter plot shows the aptamer binding signature used to train the model, and the bottom right shows the R between measured and predicted binding. Each point on a scatter plot represents one peptide from the test set.

2.4.5 Impact of Number of Sequence-Binding Pairs Used in Training and the Number of Training Steps on Correlation Between Measured and Predicted Binding Values

To determine the impact the number of training steps has on the ability of the NN to make accurate predictions, the average R (n=10) between measured and predicted binding values of the test set was plotted as a function of the size of the training set for each aptamer. Training set for this experiment was 90% of the sequence-binding pairs in the full data set for each aptamer's binding signature with the remaining 10% used as the test set. To ensure the R value was indicative of the NNs performance and not just an artifact of a particular set of sequence-binding pairs used for the training, the NN was reinitialized with a randomly selected training set for each of the 10 replicates. R of the test was calculated for a range of training steps, with all other NN parameters remaining constant. The impact that the number of sequence-binding pairs used in training has on the ability of the NN to make accurate predictions can be modeled in a similar fashion. The R between measured and predicted test set by holding the number of training steps constant while varying the size of the training set. Figure 2.5 shows graphical representation of R as a function of training steps, and as a function of size of the training set for each of the aptamers.

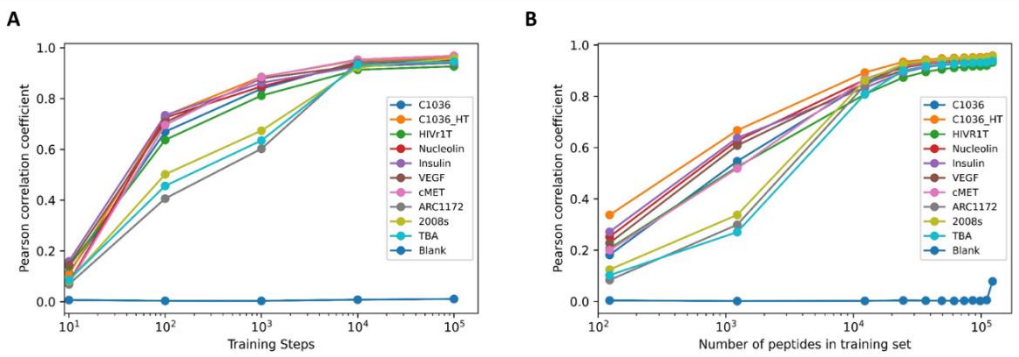


Figure 2.5: R as functions of training steps or training set size. A) R between measured and predicted binding values of the test set as a function of the number of training steps. B) R between measured and predicted binding values of the test set as a function of the number of sequence-binding pairs included in the training set. All points represent the average R across 10 NNs.

2.4.6 Variability due to Random Selection of Training Set

To analyze the variability across models built from the same data sets, the dataset from each aptamer was used to train 10 models, using 80% of the data as the training set, selected randomly for each model. Each model was trained using 20,000 training steps. Because the training set was chosen at randomly for each model, 7500 randomly generated peptides were used as the test set to ensure consistent sequences for each model.. The 7500 randomly generated 12mers used 19 out of the 20 naturally occurring amino acids, to account for amino acids withheld in the synthesis from both array manufacturers. The predicted binding value for each sequence in the randomly generated "test" set was calculated by each of the 10 models built. The %CV was calculated between all 10 predicted values, and the average %CV for each aptamer in the test set

was taken. Table 3 shows the Average %CV between 10 models trained on different subsets of sequence-binding pairs from each aptamer.

Table 2.3: Variability of predicted binding values. The dataset from each aptamer was used to train 10 unique models, using a randomly selected 80% of the dataset for training each time. Because there is no overlap between sequences on the PAC and HT arrays, a list of 7,500 randomly generated peptide sequences, with length 10 amino acids were used as the validation set for each model built. Average %CV represents the variability of those 7500 sequences across 10 unique models for each aptamer's data set, with training on either log transformed data, or non-transformed data.

Aptamer	Average %CV (log space)	Average %CV (Real space)
No Aptamer	0.07%	0.47%
C1036	1.13%	8.41%
2008s	0.47%	3.22%
TBA	0.46%	3.54%
ARC1172	0.48%	3.24%
C1036	0.64%	4.55%
HIVr1t	1.09%	8.10%
AS1411	1.06%	7.96%
IBA	1.37%	10.33%
V7t1	1.02%	7.46%
cMET	0.89%	6.39%

2.4.7 Ability for Neural Networks to Predict Aptamer Specificity

The aptamers V7t1, RT1, AS1411, IBA, and trCLN3 were measured under identical conditions at the same time. This provides an opportunity to explore how well a NN can predict binding specificities, ratios between two aptamers' binding signatures. For each pair of aptamers, binding specificity for each peptide was represented as the ratio of the measured binding between each aptamer pair. 90% of the sequencing-binding ratio pairs were used to train a NN. The log10 of the ratio between predicted binding values for the remaining 10% of the peptides in the validation set were plotted against the log10 of their measured binding ratio. The Pearson correlations for each comparison shown in the subplots of Figure 2.6 were all greater than 0.6.

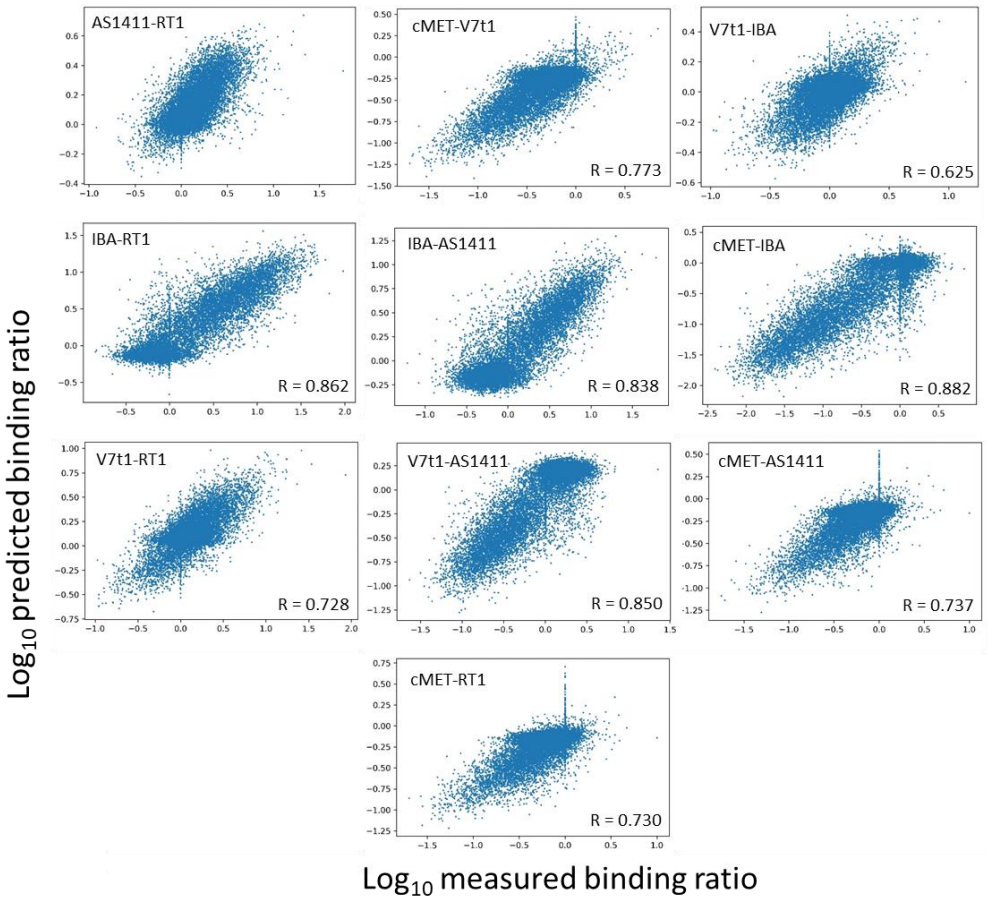


Figure 2.6: Predicted versus measured binding specificity. Binding specificity is defined as the log10 of the ratio of measured binding values between two aptamers listed above. Log10 of the measured binding specificities are along the x-axis, and values predicted by a neural net trained on the given aptamer pairs is along the y-axis. These calculations require that the two aptamers be measured under identical assay conditions, this is true for all five aptamers listed above. Calculated R between measured and predicted shown for each scatter plot. Each point represents one of ~20,000 peptides from the validation set.

2.4.8 Machine Learned Characteristics of Amino Acids

In all of the analyses above, each amino acid is assigned a vector representation by an encoder matrix, and that encoder matrix is one of the layers optimized during training of the neural network. These resulting descriptors should presumably describe each amino acids chemical properties such as charge, size, pKa, polarity, etc… For all analyses, the neural network was allowed to optimize 10 parameters (chosen arbitrarily) to describe each amino acid in the encoder matrix. Previous studies have shown model performance as a function of the number of descriptors allowed to be optimized on has little no to effect on the correlations between predicted and measured values. A target specific amino acid similarity matrix can be calculated using the final learned vector representations in the encoder matrix. These amino acid similarities are represented as normalized dot products and is given a heat map (Figure 2.7). A value of 1 means that the vector representing the two amino acids being compared are related parallel by a positive proportionality constant (e.g. two very similar amino acids like lysine and arginine). A value of 0.5 on the heat map indicates the two amino acids have no discernable difference

with relation to binding on the array, and a value of 0 indicates that the two amino acids being compared are related parallel by a negative proportionality constant. The heat map generated from the encoder matrix learned via the C1036 data set shown in figure x generally shows chemically intuitive similarities (D&E, R&K, G&S) and dissimilarities (like D&E vs R&K). The aptamers tested in this body of work all have comparable results, the optimized encoder matrix reflects an average of the molecular interactions at the aptamer-peptide interface.

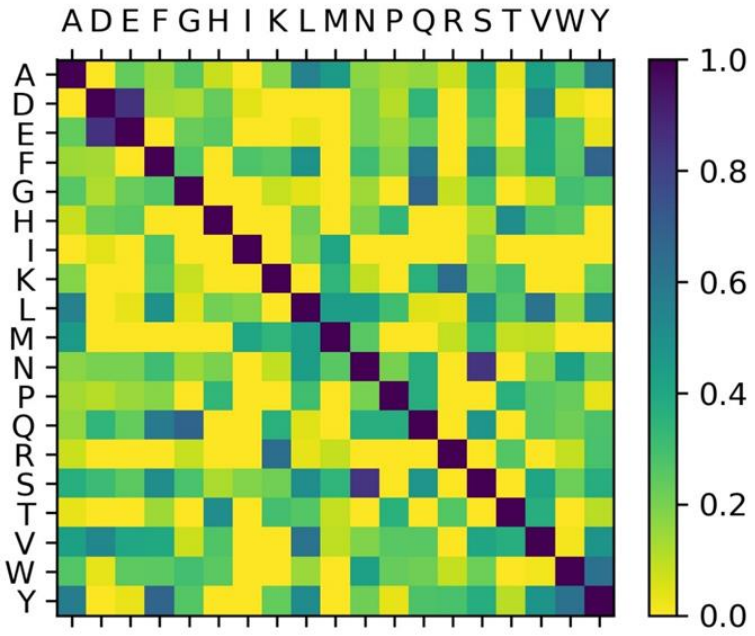


Figure 2.7: Amino acid similarity matrix. Magnitude normalized dot-products for each pair of amino acid vector representations extracted from a model trained on C1036 binding signature. The model was allowed 10 parameters to describe each amino acid.

2.4.9 Predicted probability of binding to amino acids in cognate targets

Thus far, NNs have demonstrated capacity to learn a comprehensive relationship between amino acid sequence and aptamer binding intensity despite only being informed on a minute subset of the combinatorial space of all possible 11mer peptides. This is illustrated by the high correlations between measured and predicted binding for all aptamers used in this study. The test set used to calculate those correlations were essentially a random sampling from the entire combinatorial space despite being present on the array surface. The next logical step is to predict the binding intensities of 11mers in the combinatorial space that are also present in the aptamer's cognate protein target. The amino acid sequence for C1036's target, hnRNP U, was truncated into 11mers, and the predicted binding for each 11mer recorded. A rolling average with period 11 was taken, to generate a binding value for each amino acid in the protein's sequence. Probability of binding was determined by dividing every tiled subsequence of the protein's predicted binding value by the binding value of the subsequence with the highest predicted value. Figure 2.8 shows a plot of the probability of binding between each amino acid in hnRNP U and C1036 with the tryptic peptide identified by Sonal et. Al as the binding interface overlayed. Figure 2.9 through 2.12 shows the other aptamers probability of binding plots

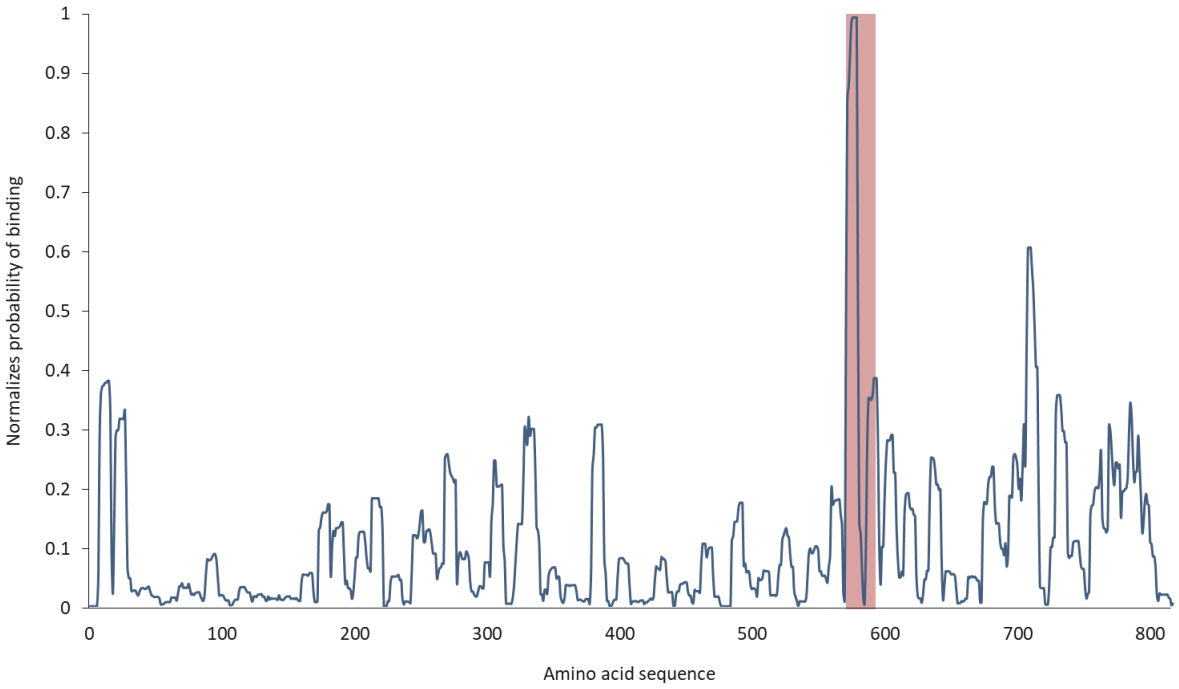


Figure 2.8: Plot of how probable an amino acid in the primary sequence of hnRNP U is to interact with aptamer C1036. Normalized probability is shown as a blue line. Red bars indicate the previously published, experimentally determined binding interface between C1036 and hnRNP U [16].

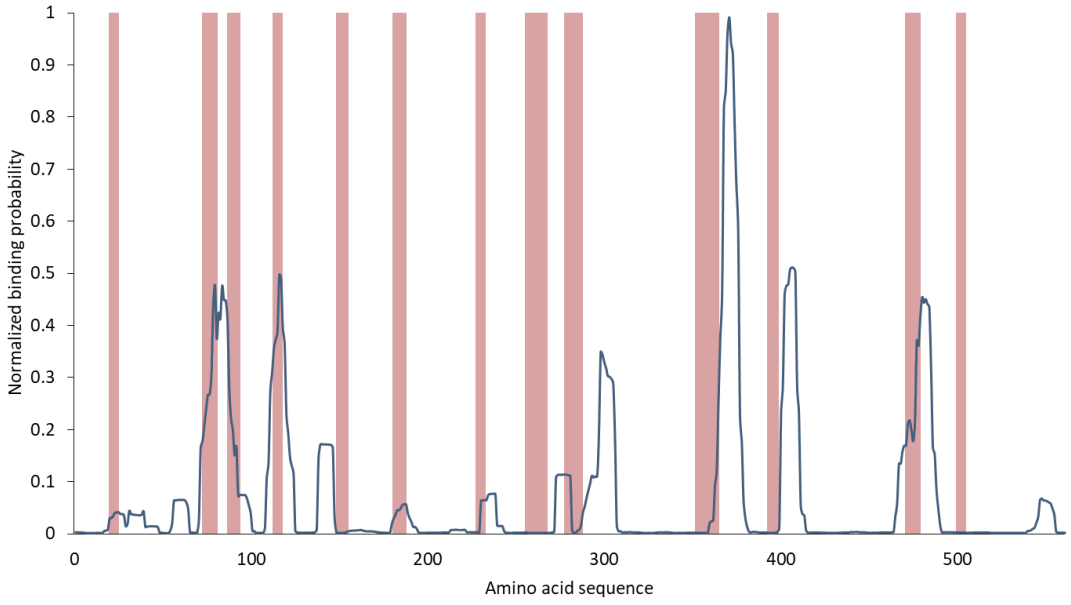


Figure 2.9: Plot of how probable an amino acid in the primary sequence of the HIV-1 reverse transcriptase is to interact with aptamer Rt1. Normalized probability is shown as a blue line. Red bars indicate amino acids within hydrogen bonding distance from the solved 3D structure of the Rt1-RT complex [33].

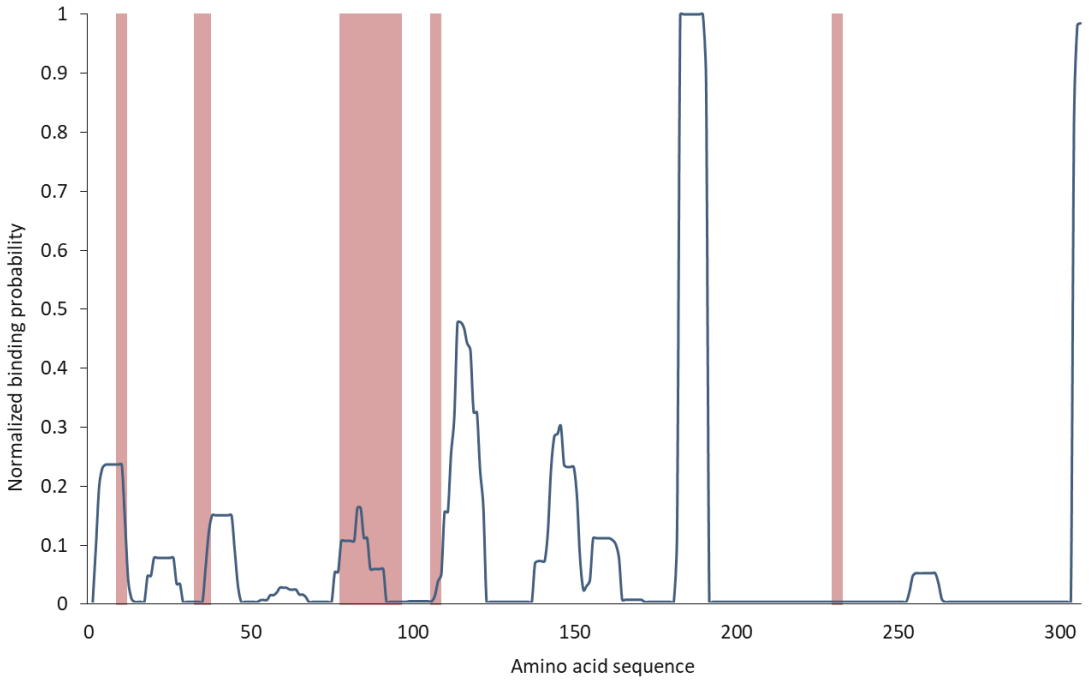


Figure 2.10: Plot of how probable an amino acid in the primary sequence of PfLDH is to interact with aptamer 2008s. Normalized probability is shown as a blue line. Red bars indicate amino acids within hydrogen bonding distance from the solved 3D structure of the 2008s-PfLDH complex [38]

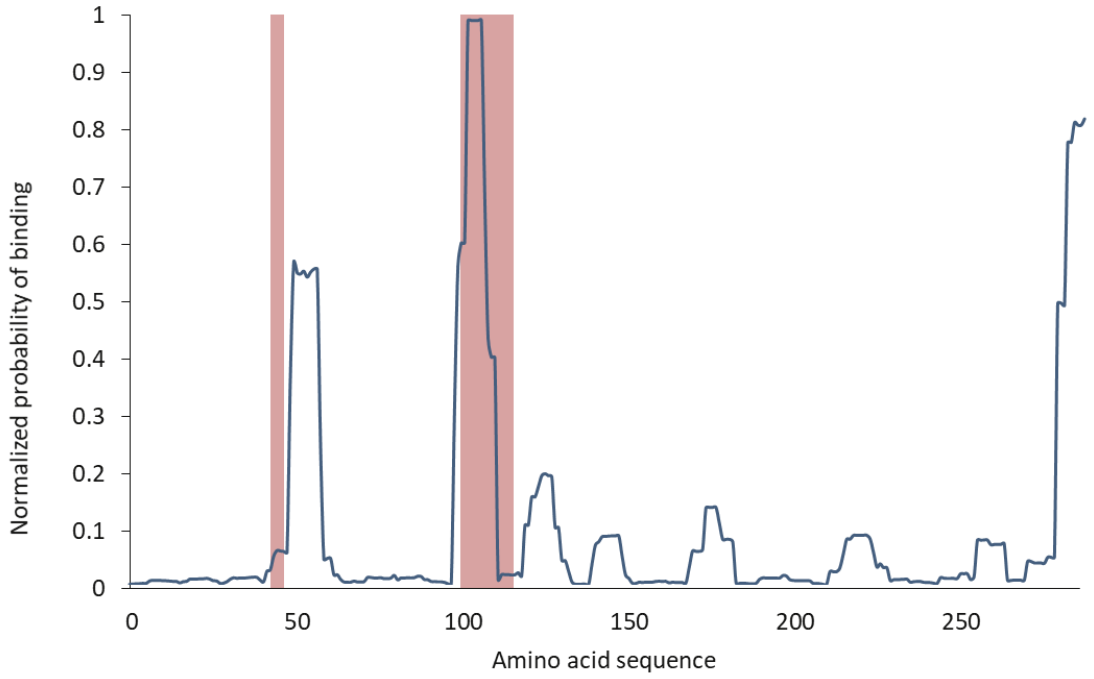


Figure 2.11: Plot of how probable an amino acid in the primary sequence of Thrombin is to interact with aptamer TBA. Normalized probability is shown as a blue line. Red bars indicate amino acids within hydrogen bonding distance from the solved 3D structure of the TBA-Thrombin complex [36].

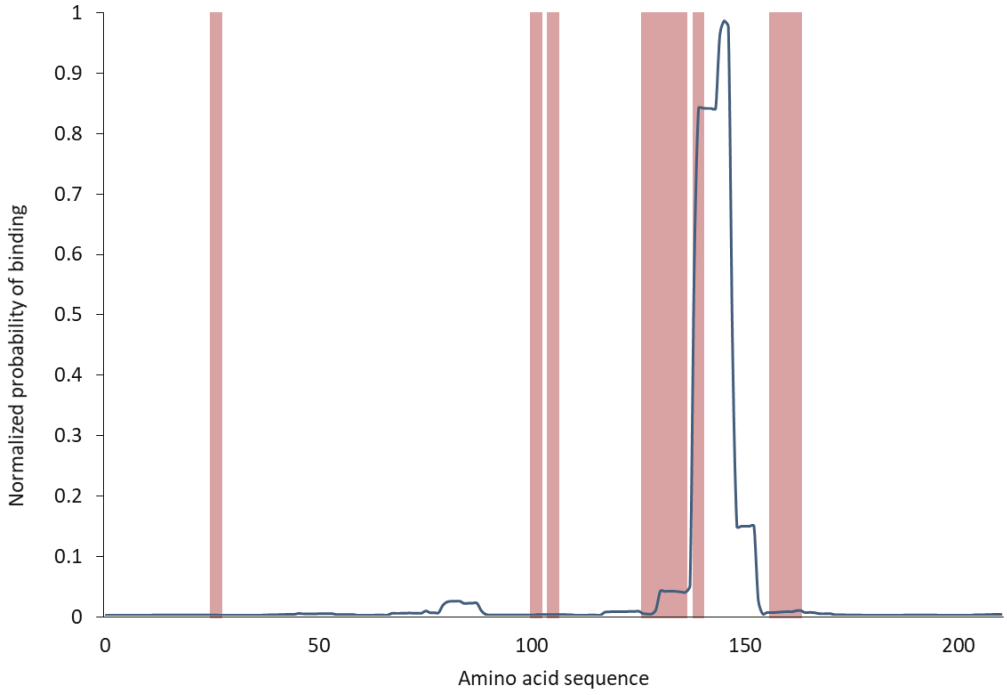


Figure 2.12: Plot of how probable an amino acid in the primary sequence of VWF-a1 is to interact with aptamer ARC1172. Normalized probability is shown as a blue line. Red bars indicate amino acids within hydrogen bonding distance from the solved 3D structure of the ARC1172-VWF-a1 complex [32].

2.5 Conclusion

Traditionally, a combinatorial chemical space requires brute force and computationally heavy simulations to explore fully. These methods have utility for certain applications, but none provide a comprehensive landscape across the entire combinatorial space being explored. Previous work has shown the utility of machine

learning to provide a comprehensive relationship between structure and function of peptide-protein interactions by probing a small subset of a simple combinatorial space of ~10 residue linear sequences of peptides using only 16 out of the 20 naturally occurring amino acids. In this work, machine learning using only a small subset of a 11mer combinatorial peptide library was sufficient enough to build a robust, comprehensive relationship between amino acid sequence and aptamer binding that can be extrapolated to the rest of the combinatorial space in question.

Aptamer binding to the arrays is concentration dependent, and therefore not random associations as indicated by increasing fluorescence signal intensity as concentration of aptamer increases. The relationship between signal intensity and concentration seems to be logarithmic as indicated by the high R^2 of the logarithmic forecast. This is understandable, the microarray scanner's photodetector has a detection limit, and there is a limited number of copies of each peptide synthesized for each feature. Aptamer binding to the peptides on the arrays is reproducible, with high R and minimal variation between technical replicates for a single aptamer (table 2.3). Similarity between binding signatures for each aptamer was to be expected, as there are some general characteristics of peptide-DNA interactions that apply, but there is enough variability between binding signatures to conclude that each aptamer has measurable differential binding.

Neural networks learn comprehensive predictive binding relationships between peptides and aptamers by sampling only a small subset of species in the possible sequence space. The neural networks can begin to make positive associations between amino acid sequence and aptamer binding by training on as few as 100 peptides in the 11mer combinatorial space (Figure 2.5B). This machine learning process is independent

of structural binding information between the aptamer and its protein target and of any chemical characteristics of the combinatorial space other than the sequence of its component building blocks. Despite the propensity for negatively charged DNA to interact non-specifically to positively charged peptides, the binding signature of each of the 9 aptamers used in this work exhibit a unique binding signature to the same ~125,000 peptides. Additionally the algorithm can pick up on those nuanced variations to accurately reproduce binding differences between two aptamers (Figure 2.6).

The true test of a neural networks capability is to determine if the binding predictions it generates match the experimentally confirmed molecular interactions within the aptamer-target interface. The binding likelihoods generated by a NN trained on C1036's binding to random synthetic peptides, correlate to the previously published binding site (Figure 2.8). While there is no structure solved for the C1036-hnRNP U complex, additional peaks in the probability curve indicate potential amino acids within close geometric proximity to the highest predicted peak. ARC1172, TBA, 2008s, and Rt1 do however have structures solved for aptamer-target complexes. The overlap between the predicted probability of binding and amino acids within hydrogen bonding distance of the aptamer in the solved complex structure is striking. While some peaks in the various probability curves line up almost exactly to the known interaction partners, other peaks seem slightly shifted. This could be due in part to the nature of solving the 3D structures of biomolecular complexes in a rigid conformational state. The shifts in the predicted probability peaks could be explained in part by the dynamic nature of proteins in vivo.

CHAPTER 3

METHODS COMPARISON OF NEXT GENERATION SEQUENCING AND DNA MICROARRAYS FOR MONITORING POOLS OF APTAMERS IN A SELEX-BASED BREASTCANCER DIAGNOSTIC LIBRARY

3.1 Abstract

3.2 Introduction

Aptamers have become an invaluable tool in diagnostics, differentiating diseases states, differentiating responder-non-responder to treatments for enrollment in clinical trials, and a powerful fishing tool for discovering actionable targets in complex biological systems. Historically, large diverse libraries of synthetic oligodeoxynucleotides (ODNs) are enriched to develop a pool of sequences with high affinity to a specific target. Individual sequences within that pool of high affinity binders are further characterized. Aptamers present a profound detection systems for single biomarkers, but complex biological states such as cancer are comprised of dynamic networks of molecular interactions. The traditional single aptamer – single protein enrichments are inherently uninformed enough to capture, reproducibly, complex biological states.

Enriched libraries of ODNs are generated through the systematic evolution of ligands exponentially (SELEX). For a starting library with a variable region of 35 bases, the combinatorial possibilities are in the magnitude of 10^{21} possible sequences. Enrichments towards a specific target or biological state reduce that complexity down to the range of 10^3 to 10^6 species. A small set of individual sequences are traditionally selected for further characterization of affinity and performance. The informative and

differentiating capabilities of an enriched library are lost when only looking at single species after enrichment. Recent studies have demonstrated the utility of ignoring the specific targets of individual species in a library and looking at the presence, absence, and frequency of sequences within the library as an indicator. This platform of using the entire enriched library to capture the phenotypic diversity of specific biological state is termed poly-ligand profiling (PLP). PLP allows access to molecular interactions otherwise unavailable to single ligand profiling of a biological state. PLP has been applied to determining responder-non responder to trastuzumab-treated breast cancer patients, as a novel liquid biopsy for breast cancer, and there are other ongoing studies on applying PLP to other cancers and treatment responses. While single species within these complex libraries can identify unique actionable binding partners, the informative potential of the library as a whole deserves exploration. PLP's utility in the liquid biopsy space was shown using a library ODNs of interest is termed L_{2000}, published in Nature Scientific Reports by Domenyuk et al. in 2017. This enriched library contains 2000 unique species, which gives an area under the curve (AUC) of 0.73 when comparing plasma from healthy donors with biopsy-positive breast cancer patients with n = 500. L_{2000} differentiation capability is entirely independent of knowledge of the binding partners for any of the 2000 sequences represented in the library. Figure 3.1 lays out the workflow of PLP on patient plasma.

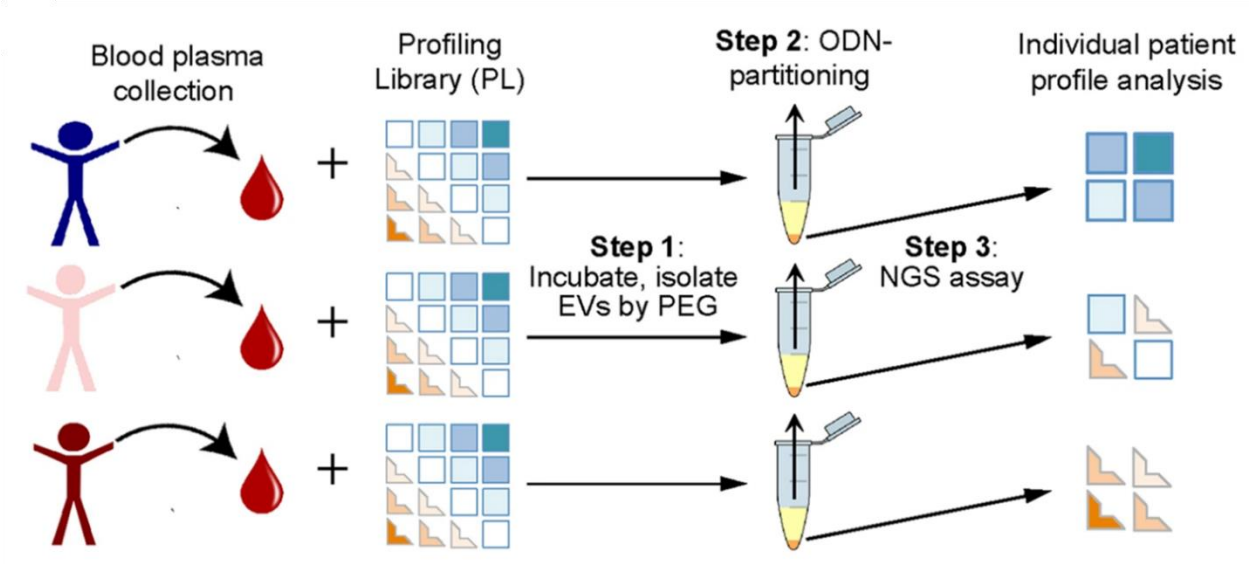


Figure 3.1: Schematic for Poly-ligand profiling (PLP) workflow. Plasma is collected from patients. Each specimen is incubated with an aliquot of the profiling library. Bound and unbound species from the profiling library are partitioned by isolating exosomes and other supramolecular structures above a certain size by precipitation with polyethylene glycol (PEG). The species bound to the precipitated complexes are pelleted and saved for downstream analysis, while the unbound species in the supernatant are discarded. The information content (nucleotide sequences, and ODN frequency in the bound fraction of the profiling library are ascertained by Next Generation Sequencing (NGS), with abundance of specific sequences serving as the indicator for the assay.

The tried and true method of dissecting sequence content, and species frequency in an enriched library is next generation sequencing (NGS). NGS provides a powerful platform with massive depth of coverage to look at sequence and copy number for species in a library of oligonucleotides. NGS is inherently an indirect method of detection after probing however. Libraries must undergo amplification to add specific tag

sequences for hybridization to substrate used for sequencing. Additional “barcode” sequences need to be added for data deconvolution, and if the sequencing method is sequencing by synthesis (SBS) that then sequencing itself is essentially a PCR reaction. While the errors are minimal within a single amplifications, they start to accumulate over multiple amplifications needed in the NGS process. This provides opportunity for adding mutations to the library and moving further away from directly detecting specific sequences in the library. The readout of sequencing libraries in this method lends to favoring sequences that amplify via PCR easier, and those sequences are not necessarily the most informative species in the library. One of the major issues with the Illumina platform comes when looking at a subset of sequences across multiple experiments. Despite including equimolar mixture of barcoded and purified sequences, identified barcodes are not evenly distributed in NGS readout data when the pool is loaded on multiple flowcells. High abundance species in libraries are usually an indicator of PCR performance, and not binding performance, and using NGS to assess performance favors easily amplifiable sequences. NGS is powerful for analyzing libraries as a whole after enrichment to identify potential informative sequences to explore further, but the simultaneous monitoring of large groups of sequences across multiple experiments needed for an enriched library to break into the clinical diagnostic assay space is a barrier that needs to be overcome.

High density DNA microarrays are a potential alternative to sequencing for direct detection of large groups of specific sequences in an enriched oligonucleotide library. DNA microarray technology provides for simultaneous monitoring of large numbers of target sequences. DNA microarrays have diverse proven functionality for profiling gene

expression, detecting single nucleotide polymorphisms, and detecting gene fusions. The vast majority of commercially available DNA microarray kits follow similar processing methods. Probes for specific targets are identified and immobilized on a solid substrate either by spotting the full sequence, or printing via a modified photolithographic method directly on the surface. DNA, RNA, or most commonly mRNA is isolated from samples, cDNA is amplified from isolated genetic material, fluorescently labeled, and allowed to hybridize to the array. High resolution image of the hybridized and washed array is taken, and a list of targets and their measured fluorescence values is generated through image analysis. Since the fluorescence signal intensity is directly proportional to the number of labeled molecules present, the measured fluorescence value can quantify amplifications and increased copies of a target present, and also deletions. The resolution of microarrays is determined by the copies of each probe in a single feature, with a higher density allowing for more precise differentiation between copy number differences in the sample. Their ability to detect highly specific sequences and frequency of particular targets serves as evidence for utility in detecting large groups of synthetic oligonucleotides in an enriched library.

While DNA microarrays rely on simple base pairing for detection of target molecules, "aptamer microarrays" can detect their target molecules directly. Optimizing aptamer microarrays proves difficult, because aptamer-target binding is largely due to 3D conformations that aptamers take, and immobilizing sequences can lead to loss of functionality. The work around for this is to retain the simplicity of DNA microarray hybridization and make probe sequences complementary to target sequences of interest in diverse enriched aptamer libraries. DNA microarray technology lends itself to precise

simultaneous measurement of vast number of target sequences in a sample. In this study we present a proof-of-concept platform for precisely measuring frequency of a sub-pool of species in a large complex enriched library of aptamers through direct detection on DNA microarrays.

3.3 Methods

3.3.1 Library sequencing

The L_{2000} and NE libraries were ordered from IDT, resuspended in water, and diluted to 4ng/uL. All 8 titer points in the serial dilutions of the L_{2000} library were made with volumes of same concentration of the NE library. An adapter sequence needs to be added to every ODN for sequencing on the Illumina NextSeq platform. 0.4ng of each titer point was amplified in triplicate with primers containing the Illumina adapter sequence, and a unique barcode sequence for downstream analysis via PCR. 24 unique barcoded adapter primers were used for 8 titers of the L_{2000} in triplicate. Samples were amplified for 10 cycles, and run on a 4% agarose gel with a 50bp ladder for confirmation of amplification. All 24 PCR samples were purified using the Beckman Coulter amPURE XP PCR cleanup protocol. The elution concentrations measured using the Sigma Aldrich QuantIT DNA quantification assay. A portion of each of the 24 samples was normalized to 0.415ng/uL and then pooled for sequencing on an Illumina NextSeq. The pool was sequenced three separate times, on three separate days, on the same instrument to assess reproducibility. Due to degradation when stored at low concentrations, the pool of 24 samples was renormalized and repooled for each separate sequencing run. Samples were demultiplexed according to their unique barcode sequence. Total sequences present in

each sample was measured, and then copy numbers determined for each of the 2000 sequences in the L_{2000} library across all titers and replicates.

3.3.2 DNA microarray content design

All DNA microarrays were purchased as custom arrays from Agilent. Probes on the DNA microarray are reverse complements to each of the sequences in the L_{2000} library. The "negative controls" for the array are 1000 randomly generated sequences with a Levenshtein difference of 10 from any of the 2000 sequences in the L_{2000} library, meaning that at least 10 bases need to be changed in order for any of the negative controls to have sequence alignment with any of the sequences in the L_{2000} library. Each of the 2000 positive controls and 1000 negative controls appear in quadruplicate on every array. Each feature on the array is circular with a diameter of ~60microns. Each array contains a total of 15,744 features (192 rows of 82 features). Features are orange packed, meaning the columns and rows are offset from each other. There are 8 arrays per slide.

Figure 3.2 shows a raw TIFF image at different scales to visualize feature geometry and packing as well as feature dimensions.

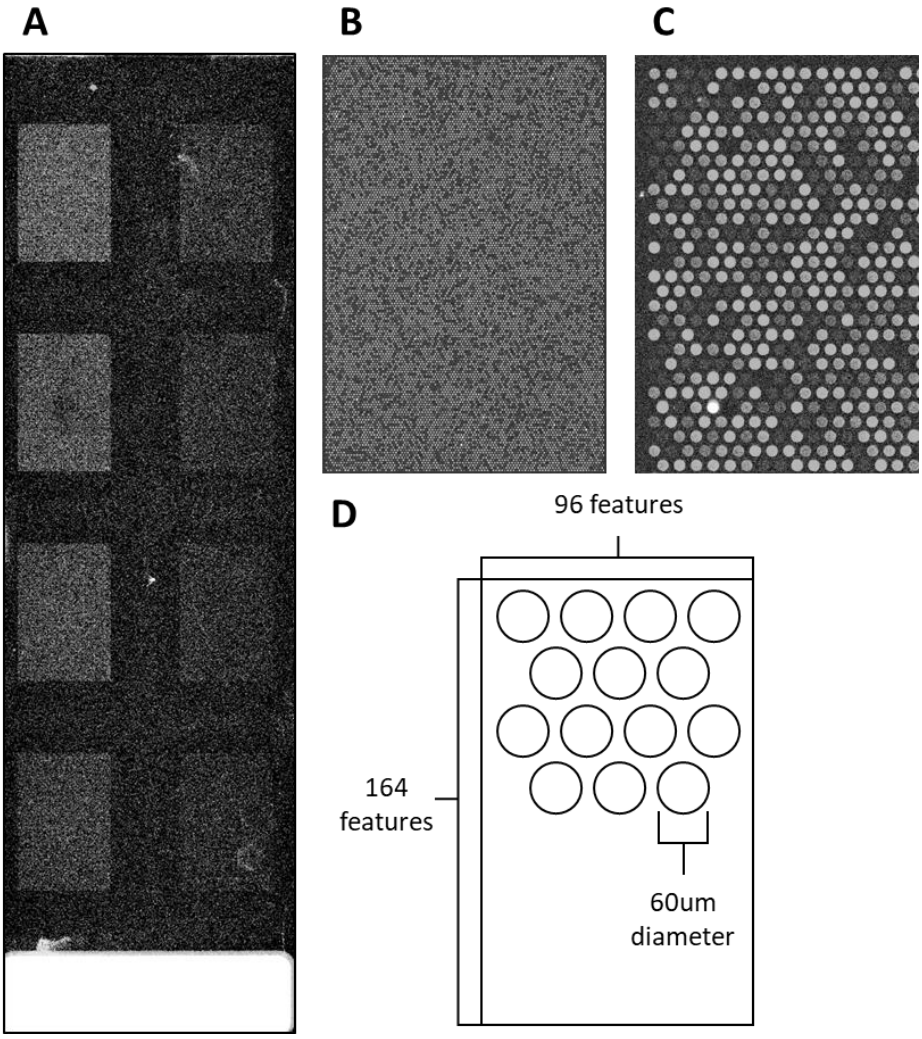


Figure 3.2: Visual representation of a slide (A) containing 8 microarrays(B). Enhanced image of a single array to visualize the orange packed grid of features (C) and Graphic representation of grid geometry and feature size (D).

3.3.3 Hybridization

A modified 2-color gene expression microarray protocol from Agilent was used for hybridization. Initial sample preparation differs but the hybridization and washing protocol remains the same. Each hybridization reaction contains 40uL with 1x of the proprietary hybridization buffer from Agilent and desired input of the enriched library based on the experimental needs. 40uL of hybridization reaction is loaded onto a gasket slide containing 8 chambers corresponding to the 8 arrays on an array slide. After loading, the slide containing the arrays is placed functional side down onto the gasket slide containing samples, and secured in a metal cassette to ensure samples do not leak across arrays. The assembled hybridization cassette is incubated using slow rotation in a temperature controlled oven with a rotor to ensure even distribution of samples across the array surface. Hybridization conditions are 65C for 17 hours unless explicitly stated otherwise.

3.3.4 Initial Washing

The hybridization cassette containing the array and gasket slide sandwich is removed from oven, and disassembled. The sandwiched gasket and array slides are separated while submerged in Agilent's proprietary DNA microarray washing buffer #1. The array slide is subsequently washed in fresh wash buffer 1 for 1 minute, and then placed in a slide dish for secondary staining.

3.3.5 Secondary detection

All sequences in the enriched library used as input are 5' biotinylated. The sequences that hybridized to the arrays were detecting by staining with fluorophores conjugated to streptavidin. Secondary detection occurred at 37C for 1 hour with gentle

mixing using a 1:2000 dilution of streptavidin-AF546 and streptavidin-AF647 conjugates.

3.3.6 Secondary Washing

After secondary staining, array slides were washed for 1 minute in wash buffer 1, followed by 1 minute in Agilent's proprietary DNA microarray washing buffer #2. Slides are removed from wash buffer 2 slowly ensuring the slide is fully dry upon removal from the wash.

3.3.7 Scanning

Arrays were scanned in an Innopsys Innoscan 100AL with excitation lasers of 532nm and 635nm. Scan resolution was 2um per pixel. A map containing probe sequences and positions was aligned on top of the TIFF slide images after scanning to generate a spreadsheet containing a sequence identifier and fluorescence at the two scanned wavelengths for each probe. These fluorescence values were used for all further analysis.

3.4 Results

3.4.1 Copy number variability and reproducibility from NGS

Robust biological assays rely on low variability and high reproducibility to be viable. To assess reproducibility and variability of sequencing as the readout for an assay relying on PLP we have to exhibit some form of control on copy number for the sequences we're aiming to monitor simultaneously. To do this, the L_{2000} library was serially diluted with full diversity non-enriched (NE) library to generate 8 titer points with equal overall DNA

concentration (4ng/uL) but with L_{2000} making up smaller proportions of the sample for each titer point. Titer ranges from 100% to 0% L_{2000}. Each of the 8 titer points was PCR amplified to incorporate the Illumina sequencing adapter and barcode sequence for data deconvolution. The 8 titer points were replicated in triplicate with a unique barcode sequence for each of the 24 reactions. Amplification was confirmed by running 5uL of each PCR product on a 4% agarose gel with a 50bp ladder. Removal of excess primer and nucleotides was performed using the AmPure XL magnetic bead purification protocol, samples were all eluted in 1M Tris-EDTA, and DNA concentration of the purified PCR product was determined using QuantIt. Figure 3.3 show the gel confirmation after PCR for one replicate group for the sample titers and the NTC. The other two groups of titer replicates show similar amplification with no secondary PCR products present in their gels as well (Supplemental Figure 1).

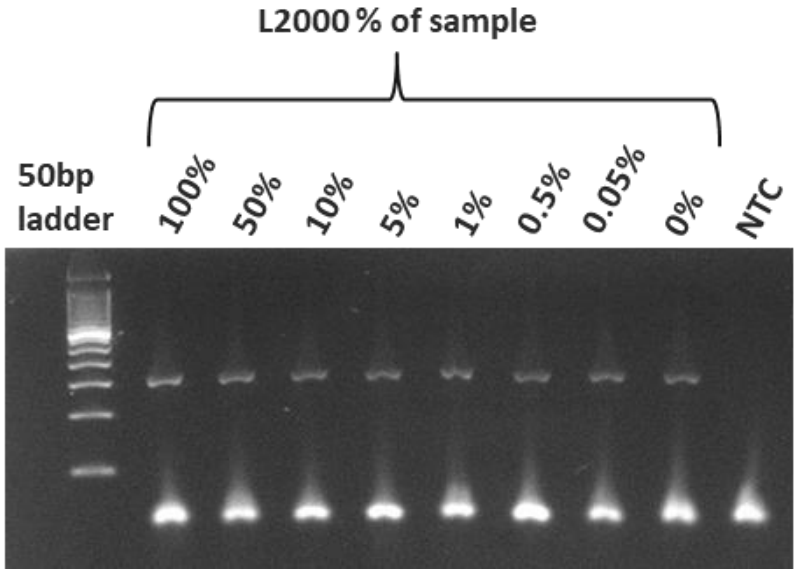


Figure 3.3: Post-PCR gel confirmation of amplification. No secondary amplification products in the samples, and no amplification in the no template control (NTC).

All 24 reactions were normalized to 0.415ng/uL and pooled with equal volume for the Illumina sequencing library prep protocol. The pool of 24 samples, 8 titer points in triplicate, was sequenced three times on successive days on the same NextSeq500, using the Illumina v2.5 500/550 75 cycle single read sequencing kit. Multiple sequencing runs of the same pool allows for calculating variability and reproducibility within a sequencing run, and across different runs. Table 3.1 shows experimental and measured values for percent of sample content taken up by the sequences from the L_{2000}, and average copy number per species of the target sequences. Experimental average copies per species was calculated by determining the number of molecules present in the 100% L_{2000} sample in the sample pool used for sequencing, multiplying that by the 22.3% (the % of sample captured during flowcell loading according to Illumina), and dividing that number by the number of sequences in the library. The experimental average copies per species for the remaining titer points are calculated by dividing the average copies per species for the 100% L_{2000} sample by the dilution factor from the titer. Moving forward each sample will be identified by the average copies per species of the L2000 present in each sample.

Table 3.1

% of sample's sequence content	Experimental	100%	50%	10%	5%	1%	0.5%	0.05%	0%
	Measured	91.2 ± 0.24%	54.4 ± 1.50%	12.4 ± 0.16%	6.2 ± 0.05%	1.3 ± 0.01%	0.63 ± 0.003%	0.066 ± 0.0022%	0.00 ± 0.00%
Average copies per sequence	Experimental	6547	3274	655	327	65	33	3	0
	Measured	6526 ± 192	3655 ± 203	800 ± 77	387 ± 46	86 ± 6	42 ± 3	4 ± 1	0 ± 0

Since each titer point was sequenced in triplicate on each flowcell, every sequence in the L2000 library will have a copy number from each of three unique barcodes. This allows for monitoring the variability across PCR reactions, as well as

variability between sequencing runs. Figure 3.4 shows violin plots for distribution of all 2,000 %CVs for each of the 8 titer points.

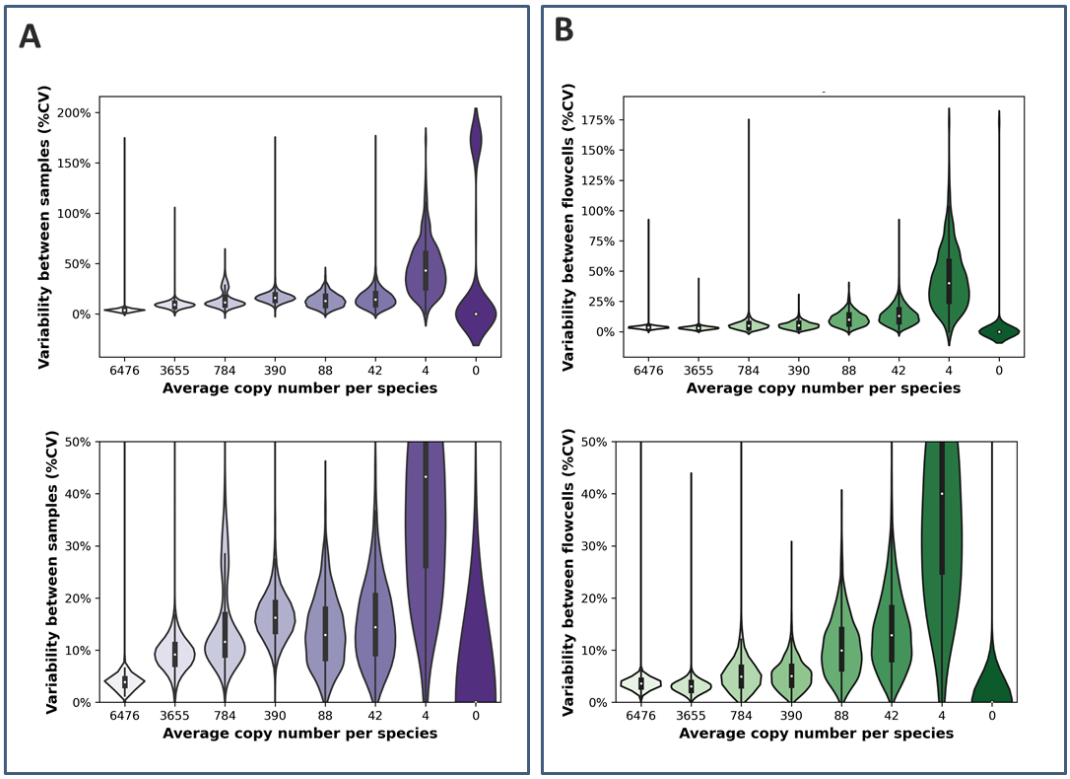


Figure 3.4: Violin plots depicting variability across samples and across sequencing runs. A) Distribution of 2000 %CVs between samples on a single flowcell for each of the 8 titer points. B) Distribution of 2000 %CVs between flowcells for each of the 8 titer points. The top graph in both panels shows the full data set, with the bottom graph only showing data within the 0% to 50% CV range.

3.4.2 Array hybridization dependence on sample concentration

To analyze binding dependence on concentration, 8 titer points of the library were each hybridized to 8 separate arrays on a single slide ranging from 10 ng to 0 ng of the L2000 library. Arrays were hybridized at 65C for 17 hours, and detected with 4nM each

of AF546 and AF648 streptavidin conjugates. Figure 3.5A shows a TIFF image of the scanned slide with the yellow intensity representing a join of the AF546 scan and the AF648 scan of the slide. Average fluorescence per feature was calculated for the L2000 probes (n = 8000) and the negative control probes (n = 4000) for each library input (Figure 3.5B). Variability of hybridization to the four replicates of each feature was quantified using %CV (standard deviation / mean). The 2,000 %CVs for the L2000 probes, and the 1,000 %CVs for the negative control probes are plotted as violin plots to visualize the distribution of variability within a sample type for each of the inputs of the library (Figure 3.5C).

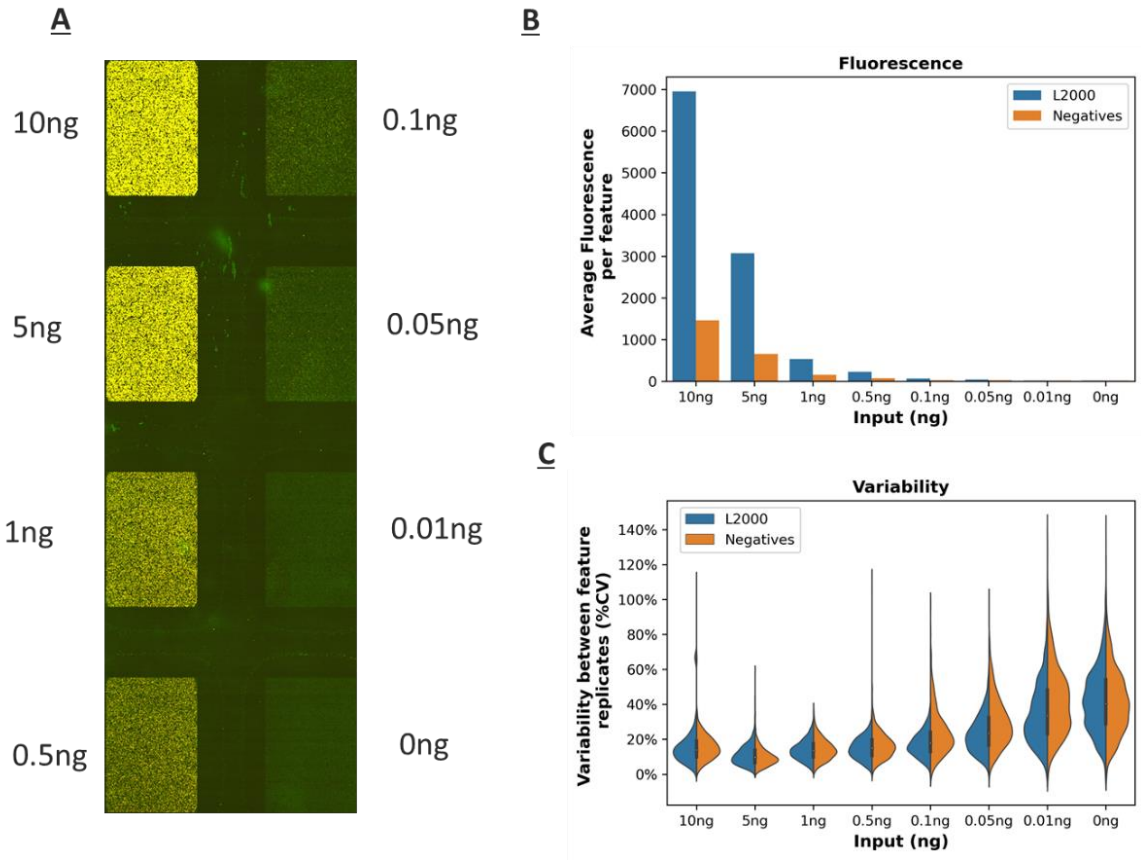


Figure 3.5: A) Raw Fluorescence TIFF image of a slide with 8 microarrays. Library inputs indicated next to each array. TIFF image contains the merged AF546 and AF648 scans. B) Bar plot of average fluorescence per feature for each of the 8 library inputs. C) Violin plot showing distribution of %CVs for each of the 8 library inputs. L2000 n = 2000, Negatives n = 1000. Population variability is indicated by the violin plot, hence no error bars on the bar plot.

3.4.3 Reproducibility and variance across probes, arrays, and slides

Reproducibility is key to any assay. To determine consistency across probes, arrays, and slides 1ng of the enriched library was denatured and hybridized to 9 arrays; 3 arrays across 3 separate slides. Hybridization occurred at 65C for 17 hours. Each probe on all 9 arrays were background subtracted and median normalized. Figure 3.4 shows a summary of variance in fluorescence across probes, arrays, and slides. %CV was calculated for individual probes, and the average %CVs for all the positive controls and negative controls for all 9 replicates represented in Figure 3.4A. The %CV between all positive probes and between all negative probes across 3 technical sample replicates on each slide were calculated and represented in Figure 3.4B. The %CV between all positive probes and all negative probes across arrays, and across slides is represented in Figure 3.4C.

3.5 Discussion

3.5.1 Variability and reproducibility of NGS

The reproducibility and variability of NGS was assessed by controlling the average copy numbers per species for a set of 2000 sequences within a diverse library, and sequencing in triplicate across 3 separate flowcells. Variability within a flowcell

remains below the accepted 10%CV threshold for biological assays when the copy number for a particular sequence is over 390. Figure 4A illustrates the tendency towards larger copy number variability between replicates on a single flowcell as the average copy number per species of the pool of sequences being monitored decreases. The same trend applies to variability between sequencing runs, with %CVs for all 2000 sequences being monitored remaining below the 10%CV threshold until the copy numbers are below 390 (Figure 4B).

3.5.2 Binding is concentration dependent

Binding to the arrays is concentration dependent. There is a strong positive linear relationship between library input and average fluorescence per positive probe on the arrays. The slope of the linear regression between library inputs and average fluorescence per probe has a magnitude of 62.343, and the R^2 is 0.9952. The Pearson correlation coefficient between average fluorescence per positive probe, and library input amount is 0.9975. The linear regression between average fluorescence per negative probes and library input is nearly horizontal. The magnitude of the slope for the linear regression for the negative controls is -1.9509, with an R^2 of 0.1387. The Pearson correlation coefficient between library input and average fluorescence per negative control probe is -0.3724 (figure 2). All of this lends evidence to the library binding to the arrays being concentration dependent, signal primarily coming from the positive control probes. Signal is being driven by the positive control probes on the array is expected, as all the negative controls are at least 10 bases different from any sequence in the entire library, we would expect to see minimal hybridization to those probes.

3.5.3 Denaturing library increases signal

Denaturing the library before processing ensures a larger amount of species in the library exist in single stranded forms. With more species existing as single stranded, the probability of hybridization increases. There was an average of 153% increase in average fluorescence per positive control probe across all library inputs. Denaturing had no consistent increasing or decreasing effect on the average fluorescence per negative control probe.

3.5.4 Library binding is highly reproducible

The %CV is the percent the standard deviation is of the average. %CV below 15% is widely accepted as the standard for biological assays. The average %CV for fluorescence of positive probes across probe replicates on a single array, across all technical replicates on a single slide, and across slides is less than 15%. Hybridization and fluorescence detection is highly reproducible with low variability (figure 3.5).

3.5.6 Probes on the array are specific to the target library

The probes on the array were selected to be reverse compliments of specific sequences in the target library, but due to the diversity of aptamer libraries there is always a possibility of non-specific binding. The probes on this particular array are highly specific to the target library. Signal from a non-target enriched library and a non-enriched library can be interpreted as background signal compared to the signal of the positive probes from the target enriched library. The signal from the negative control probes from the target enriched library can similarly be interpreted as negligent.

3.6 Conclusion

High density DNA microarrays provide a robust platform for simultaneous direct detection of specific sequences in an enriched library of oligonucleotides. Hybridization of an enriched library to probes is highly specific and reproducible across arrays, and across slides, with low inter- and intra- assay variability. Signal strength has a linear relationship to input in ng of the target library. DNA microarrays can provide quantitative and qualitative information on sequence content of an enriched library without the possibility of introducing mutations into species through PCR and without the possibility of equal distribution of sequences across flowcells and experiments through detection on NGS platforms. The next steps using DNA microarrays moving forward would be to assess the viability of using this platform as detection and readout for diagnostic assays that utilize enriched oligonucleotides as their substrate.

3.7 Supplemental Figures

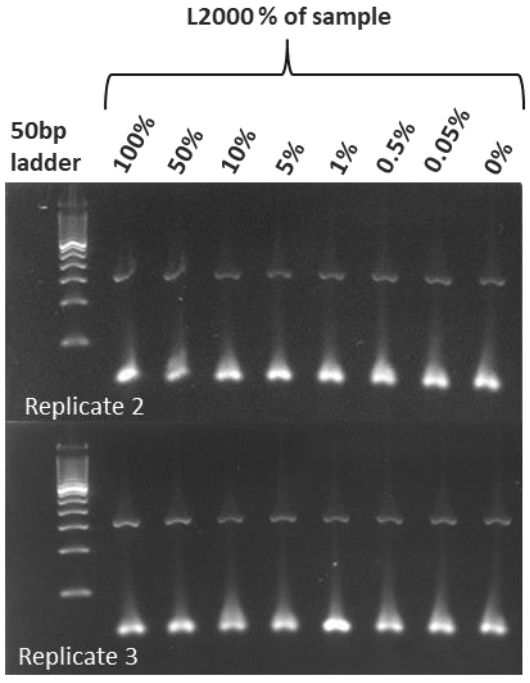


Figure S3.1: Gel confirmation of PCR amplification for the other two titer replicate groups.

CHAPTER 4

A DNA-Directed Light-Harvesting/Reaction Center System

Palash K. Dutta,[1,2] Symon Levenberg,[1,2] Andrey Loskutov,[2] Daniel Jun,[3] Rafael Saer,[3] J. Thomas Beatty,[3] Su Lin,[1,2] Yan Liu,[1,2] Neal W. Woodbury,*[,1,2] Hao Yan*[,1,2]

[1]Department of Chemistry and Biochemistry and [2]The Biodesign Institute, Arizona State University, Tempe, Arizona 85287, United States

[3]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada

4.1 Abstract

A structurally and compositionally well-defined and spectrally tunable artificial light-harvesting system has been constructed in which multiple organic dyes attached to a 3arm DNA nanostructure serve as an antenna conjugated to a photosynthetic reaction center isolated from *Rhodobacter sphaeroides* 2.4.1 (PDB 2J8C). The light energy absorbed by the dye molecules is transferred to the reaction center where charge separation takes place. The average number of DNA 3arm junctions per reaction center was tuned from 0.75 to 2.35. This DNA-templated multi-chromophore system serves as a modular light-harvesting antenna that is capable of being optimized for its spectral properties, energy transfer efficiency and photo-stability, allowing one to adjust both the size and spectrum of the resulting structures. This may serve as a useful test-bed for developing nanostructured photonic systems.

4.2 Introduction

During photosynthesis, light energy is collected by a large light-harvesting network and efficiently transferred to a reaction center (RC), which converts it to chemical energy via charge separation [41-44]. The quantum efficiency of the charge separation reaction by the photosynthetic reaction center is nearly unity. The architecture and spectral properties of the light-harvesting system that surrounds the reaction center have evolved to meet the constraints of a broad range of different light conditions and environments [41]. A number of researchers have attempted to mimic the natural photosynthetic apparatus by designing artificial light harvesting antenna systems [45-54] [55-60] [61-68] [69-72] for a variety of photonic applications [73].

To facilitate nanoscale photonic applications more broadly, the construction of artificial antenna systems that provide controllable light absorption, efficient energy transfer and improved photo-stability are desirable. Self-assembling proteins [55-60] and dendrimers [61-68] have been explored to create artificial antenna systems, but they lack a well-defined multi-chromophore geometry and stoichiometry. Synthetic porphyrin structures [69-72] have been investigated to create artificial antennas connected to electron transfer complexes, but these generally have an absorption cross-section that is spectrally relatively narrow. DNA nanotechnology can be used to generate programmable, self-assembled nanostructures [73-93] with multiple fluorophores at well-defined positions, and this approach has been used to create artificial light harvesting antenna systems. Double helical DNA structures, three-way junctions, seven helix bundles and several other DNA based antenna systems [48, 94-103] have been used to create artificial antennas with

unidirectional energy transfer along an excited state energy gradient between chromophores that mimics the stepwise energy transfer in some of the natural photosynthetic systems. However, thus far these assemblies have lacked the ability to convert the light energy to redox energy via charge separation.

Recently, we have studied different dye molecules directly conjugated to reaction centers and explored the effects of altering the dye spectral and excited state properties on the efficiency of energy transfer and charge-separation [104, 105]. In this report we go a step further and use a 3arm-DNA nanostructure to organize multiple dye molecules and specifically assemble these nanostructured complexes with reaction centers (Figure 4.1A), resulting in a geometrically programmable model system mimicking a natural photosynthetic apparatus.

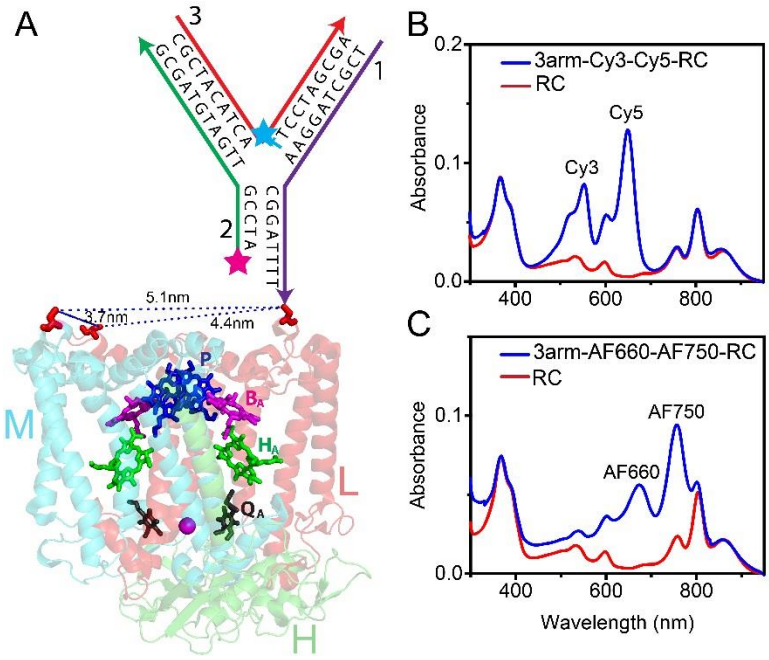


Figure 4.1. (A) Modified structure of the reaction center (RC) from the purple bacterium, *Rhodobacter sphaeroides* 2.4.1 (PDB 2J8C) with sequences of the 3arm-DNA construct

shown. The cofactors of the RC are colored and those active in electron transfer reactions involved in this report are designated by letters: P – bacteriochlorophyll pair, B_A – bacteriochlorophyll monomer, H_A – bacteriopheophytin, Q_A – ubiquinone. The arrows in the DNA structure point in direction of the 3' end of the DNA strands. The 3'-Amine modified Strand-1 (purple) of the 3arm-DNA is conjugated to one of the Cys residues (shown in red) on the surface of the RC via a SPDP (N-succinimidyl 3-(2-pyridyldithio) propionate) linker. The other two strands (Strand-2 and -3 in green and red, respectively) are allowed to hybridize to Strand-1 to form the 3arm-DNA junction. Inter-Cys distances on the RC are marked as dotted lines. The two stars on 3arm represent the positions of the two dye molecules, where the cyan star corresponds to either Cy3 or AF660, and the pink star corresponds to either Cy5 or AF750. It should be noted that because of the presence of three Cys residues on the surface of the protein, 1 or 2 or 3 copies of Strand-1 can be conjugated to the RC, and consequently up to three 3arm-DNA junctions (and three pairs of dyes) can be conjugated to the RC. For clarity, only one is shown here. (B) A representative absorption spectrum of RCs that have an average of 2.3 of the 3arm-DNA-Cy3-Cy5 nanostructures attached. (C) An absorbance spectrum of RCs that have an average of 2.1 of the 3arm-DNA-AF660-AF750 nanostructures attached. The absorbance spectra of panels B and C show enhanced absorbance cross-section in the spectral regions 450-700 nm or 500-800 nm, respectively, where the RC absorbance is relatively low. The spectrum of free RC is shown in both panels B and C (red trace) for comparison.

Two different pairs of DNA-conjugated chromophores are used in this study: Cy3 and Cy5, or Alexa Fluor 660 and Alexa Fluor 750. Cy3 acts as the donor and Cy5 as the acceptor in the first pair, and AF660 acts as the donor and AF750 as the acceptor in the

second pair. The fluorophores were chosen so that there is significant spectral overlap between emission of the dyes and the absorption of the RC to facilitate efficient energy transfer, and so that there is a substantial increase in the absorption cross-section in the spectral regions where the absorbance of the RC alone is low (Figures 4.1B-C and 4.3). A very simple 3arm-DNA structure was designed to assemble the two dye molecules in a geometrically defined manner and to avoid chemical modification of any DNA strands with more than one dye (to reduce cost and synthetic complexity) (Figure 4.1A). Two of the strands (Strand-2 and -3) in the 3arm-DNA contain the dye molecules, and the other one (Strand-1) is conjugated to the RC through a covalent cross-link.

The three dimensional structure of the RC complex from *Rhodobacter sphaeroides* 2.4.1 (PDB 2J8C) is depicted in Figure 1A, and it consists of three subunits H, M and L. There is a total of ten cofactors associated with the L/M transmembrane region of the structure, including a dimer of bacteriochlorophylls (P), two monomer bacteriochlorophylls (B_A and B_B), two bacteriopheophytins (H_A and H_B), two ubiquinone-10 molecules (Q_A and Q_B), one carotenoid and one nonheme iron (Fe^{2+}).[10] The special pair P is the primary donor of electrons in the light-driven electron transfer process, which subsequently transfers electron to Q_A via B_A and H_A, forming a long-lived charge-separated state $P^+Q_A^-$. When ubiquinone is bound in the Q_B site, electron transfer occurs from Q_A^- to Q_B forming $P^+Q_B^-$ [106-111] .

A genetically modified RC was used in these studies and contained a total of eight mutations, five of them to replace the five wild-type cysteines with serine or alanine, and the remaining three to replace three selected wild-type amino acids (asparagine or glutamic acid) with cysteine residues at specific locations on the surface of the RC that are close to

the primary electron donor, P [104] [112].Two of the new Cys residues are located on the surface of the L subunit (L72, L274) and the other one is on the surface of the M subunit (M100) (Figure 4.1A).

4.3 Results and Discussion

4.3.1 Assembly of Light-Harvesting/Reaction Center Complex

A 3'-Amine modified Strand-1 was conjugated to the introduced Cys residues of the RC in a 10:1 molar ratio by using a SPDP (N-succinimidyl 3-(2-pyridyldithio) propionate) cross-linker (see details in the Supporting Information). The reaction mixture was subsequently purified by fast protein liquid chromatography (FPLC) (Figure 4.2) (see Supporting Information for methods). The chromatograph shows four prominent peaks using absorbance at 260 nm and 280 nm, and three peaks using absorbance at 365 nm (Soret peak of RC). The fractions under each peak were collected and characterized. The UV-vis absorbance maxima for the first, second and third peaks in the chromatograph are at 271 nm, 268 nm and 266 nm, respectively. The blue shift of the absorbance peak together with a relative increase in the absorbance intensity (compared to the absorbance peak at 800 nm) indicate that the species contained in the peaks have different ratios of DNA conjugated to the RC, increasing from peak 1 to peak 3. (DNA:RC = 1:1, 2:1 and 3:1).

It is important to note that the single copy of Strand-1 conjugated to RC can be on any of the three Cys. Similarly, there are three ways that two copies of Strand-1 could be conjugated to the RC. This heterogeneity of the sample is reflected by the widths of the first and second chromatograph peaks. The third peak, in contrast, has the narrowest peak and highest ratio of A260/A365 among the first three and it represents a single species of RC with three copies of Strand-1 conjugated to all of the Cys residues. The last peak in

the chromatograph has no absorbance at 365 nm (the Soret absorbance band of the RC), indicating that it is excess free ssDNA with no RC attached.

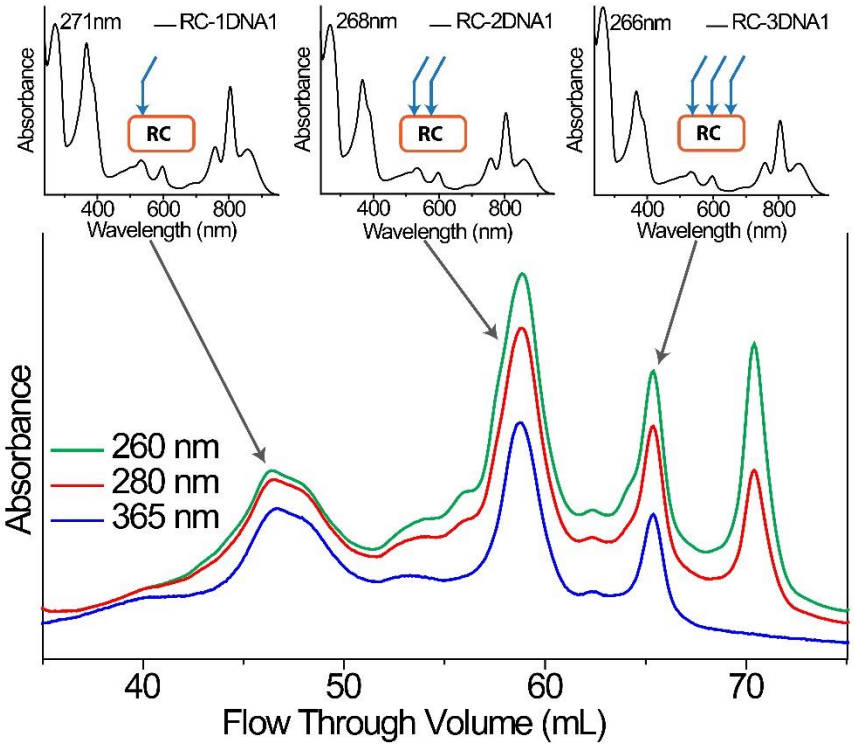


Figure 4.2: FPLC purification trace of DNA (Strand-1) conjugated RCs. Chromatographs at 260 nm (green), 280 nm (red) and 365 nm (blue) are shown. The absorbance bands at 260 nm and 280 nm are from both RC and DNA, whereas the absorbance bands at 365 nm are from the RC. The fractions from each of the peaks were collected separately and their respective absorbance spectra measured. Schematics corresponding to the absorbance spectra showing number of DNA strands conjugated per RC are given at the top of the figure.

Dye-labeled pre-annealed Strand-2 and -3 are then allowed to hybridize to the purified Strand 1-conjugated-RC to create 3arm-DNA-RC conjugates with one, two or three 3arm-DNA junctions on each RC (Scheme S4.3-S4.44) carrying different identities

and numbers of dyes. Cy3-modified Strand-3 and Cy5-modified Strand-2 were purchased from Integrated DNA Technologies (IDTDNA). AF660-modified Strand-3 and AF750-modified Strand-2 were synthesized by reacting amine-modified DNA (Strand-2 or -3, synthesized using a DNA synthesizer) with the succinimidyl ester of the corresponding dye (purchased from Invitrogen). The resulting conjugate was subsequently purified by reverse phase HPLC and characterized using matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectroscopy (see details in the Supporting Information, Figure S4.1).

Table 4.1. 3arm-to-RC ratio of different constructs

Dye	Sample	Abbreviation	3arm/RC [a]
Cy3/Cy5	3arm-Cy3-RC(1DNA)	1C	0.75 ±0.05
	3arm-Cy3-RC(2DNA)	2C	1.65 ±0.05
	3arm-Cy3-RC(3DNA)	3C	2.35 ±0.05
	3arm-Cy3-Cy5-RC(1DNA)	1CC	0.8 ±0
	3arm-Cy3-Cy5-RC(2DNA)	2CC	1.65 ±0.05
	3arm-Cy3-Cy5-RC(3DNA)	3CC	2.2 ±0.1

AF660/AF750	3arm-660-RC(1DNA)	1-6	0.85 ±0.15
	3arm-660-RC(2DNA)	2-6	1.6 ±0
	3arm-660-RC(3DNA)	3-6	2.15 ±0.05
	3arm-660-750-RC(1DNA)	1-6-7	0.9 ±0.1
	3arm-660-750-RC(2DNA)	2-6-7	1.65 ±0.05
	3arm-660-750-RC(3DNA)	3-6-7	2.0 ±0.1

[a]The molar ratios of the 3arm/RC were obtained by measuring the dye concentration and the RC concentration, calculated from their UV-vis absorbance spectra and known absorption coefficients, assuming a 100% dye labeling ratio on the HPLC purified DNA strands.

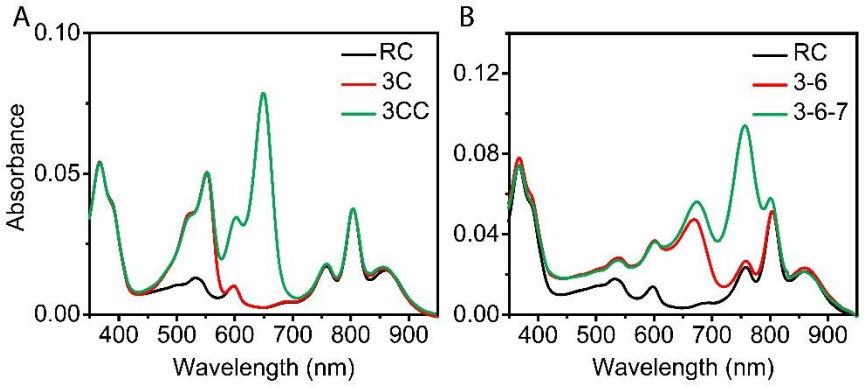


Figure 4.3: Absorption spectra of representative 3arm-DNA-dye-RC constructs. (A) Absorption spectra of **RC**, **3C** and **3CC** (B) Absorption spectra of **RC**, **3-6** and **3-6-7**.

The assembly of the 3arm-DNA-RC constructs containing only Cy3 and different DNA/RC ratios are named **1C**, **2C** or **3C** (Abbreviations as in Table 4.1). These were created by assembling Strand-2 (unmodified) and Cy3-modified Strand-3 with the FPLC fractions that contained conjugates of one, two or three Strand-1 conjugates per RC. The spectra of these structures show enhanced absorbance between 450-580 nm compared to the RC alone, due to the additional absorbance from Cy3 in this spectral region (Figures 4.3A and S4.5). 3arm-DNA nanostructure-to-RC ratios of 0.75 ±0.05, 1.65 ±0.05, and 2.35 ±0.05 were calculated based on the UV-vis absorbance spectra for **1C**, **2C** and **3C** (see note in Table 4.1 caption). Apparently, the yield of assembly for the fully loaded 3arm-DNA junction on the RC was ~75-80%. This <100% yield may be due to local steric effects near the protein surface that reduce the DNA hybridization yield. The similarly assembled 3arm-DNA-RC constructs containing 1, 2 and 3 copies of both Cy3 and Cy5 labeled DNA strands are named **1CC**, **2CC** and **3CC** (Table 4.1), and the spectral analysis revealed that they have 3arm-DNA nanostructure-to-RC ratios of 0.8 ±0, 1.65 ±0.05, and 2.2 ±0.1, respectively (Figures 4.3A and S4.6). Apparently adding the second dye molecules (covalently modified on the 5' end of Strand-2) did not affect the DNA hybridization yield. When both Cy3 and Cy5 are present (as in **1CC**, **2CC** and **3CC**), they absorb significantly between 450 and 700 nm. Similarly, the 3arm-DNA-RC constructs containing different numbers of AF660 only (abbreviated as **1-6**, **2-6** and **3-6**) and different numbers of both AF660 and AF750 (abbreviated as **1-6-7**, **2-6-7** and **3-6-7**) provide strong absorbance between 500 and 800 nm (Figures 4.3B and S4.7-S4.8). The 3arm DNA-to-RC ratios for the different constructs are listed in Table 4.1.

4.3.2 Excitation Energy Transfer Efficiency

The efficiency and kinetics of the FRET (Förster resonance energy transfer) process for each construct was investigated using both steady-state and time-resolved fluorescence spectroscopy (see Supporting Information for calculations). The free 3arm-DNA constructs with respective dye(s) attached (without the RC) were used as reference samples for these experiments (Figures S4.3-S4.4). Upon exciting **3C** at 510 nm, 30% of the Cy3 emission was quenched compared to that of 3arm-DNA-Cy3, presumably due to energy transfer from Cy3 to the RC (Figures 4.4A, S4.5C). In the case of **3CC**, there was an 84% decrease in Cy3 emission intensity compared to that of 3arm-DNA-Cy3 without the RC (Figures 4.4B, S4.6). Comparing **3C** and **3CC**, the greater decrease in fluorescence of Cy3 when Cy5 was present is attributed to the summation of multiple energy transfer pathways, which include a direct energy transfer from Cy3 to the RC and a stepwise energy transfer from Cy3 to Cy5 to the RC. Compared with the 3arm-DNA-Cy3-Cy5 alone with no RC, **3CC** (with both dyes in the same 3arm-DNA that is linked to the RC) shows a 45% decrease in total fluorescence intensity integrated from 520 nm to 850 nm upon Cy3 excitation (Figure S4.6). On the other hand, upon Cy5 excitation at 620 nm, the direct FRET efficiency of Cy5 to the RC in **3CC** is calculated to be 48%, using the emission of the 3arm-DNA-Cy3-Cy5 as a reference (Figure S4.6).

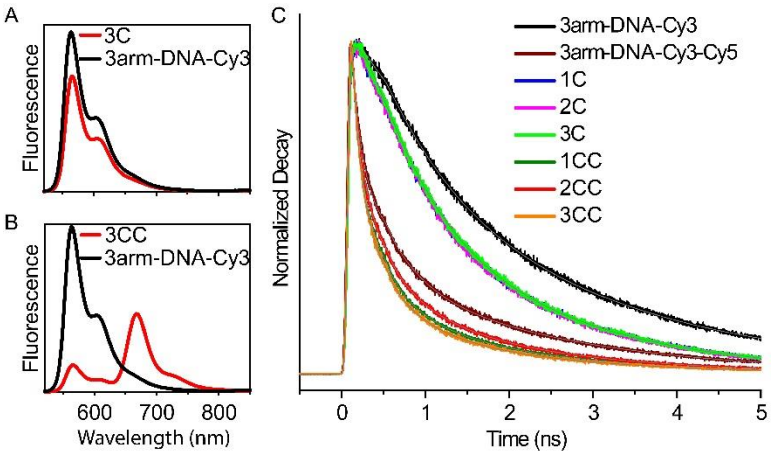


Figure 4.4: Fluorescence emission spectra of **3C** (A) and **3CC** (B) in comparison with emission spectra of 3arm-DNA-Cy3 (λ_{ex} = 510 nm). (C) Cy3 fluorescence decay profiles of free 3arm-DNA and 3arm-DNA conjugated to the RC in various ratios, with either constructs containing Cy3 alone (**1C**, **2C**, **3C**) or constructs with both Cy3 and Cy5 (**1CC**, **2CC**, **3CC**), monitored at 565 nm (λ_{ex} = 510 nm).

Similar experiments were performed on all the other 3arm-DNA-dye-RC constructs, and the energy transfer efficiency values obtained are shown in Figures 4.5, S4.5-S4.8 and S4.11. Samples with different ratios of 3arm-DNA-dye conjugate to RC (for example, compare **1C**, **2C** and **3C**, or **1-6**, **2-6** and **3-6**) all yielded similar energy transfer efficiency values between the individual dyes and the RC or between the dyes together and the RC. This is due to the fact that although there are multiple dye molecules on the assembled structures, the probability of exciting more than one dye molecule associated with a particular RC at any time is very low due to the continuous nature and low intensity of the excitation light. Moreover, as expected, the efficiency of energy transfer from AF650 to the RC (~55%) is higher than the efficiency of Cy3 transfer to the RC (~35%) (comparing Figure 4.5 and S4.11). This is presumably due to the greater spectral overlap between the emission of AF660 and the absorbance of the RC compared to Cy3. However, even though AF750 has a greater spectral overlap with RC than does Cy5, it has a lower energy transfer efficiency to RC (~42%) than Cy5 does (~52%), and this results in a higher overall energy transfer efficiency of the Cy3-Cy5 pair to the RC (~83%) than the AF660-AF750 pair (~75%). We have observed similar phenomena earlier [104], and the reason for the lower energy transfer efficiency of AF750 to RC is the shorter intrinsic lifetime of AF750 compared to that of Cy5.

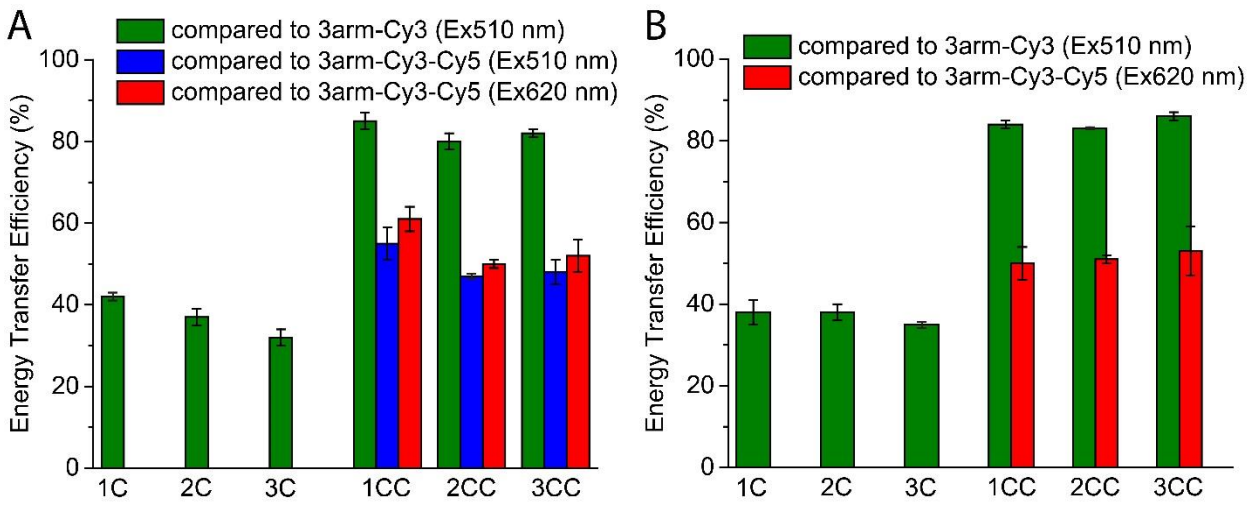


Figure 4.5: Energy transfer efficiency of 3arm-DNA conjugated RC calculated from (A) steady-state data and (B) from lifetime data. The green bars show energy transfer efficiency calculated by comparing fluorescence from the RC containing complex with that from the 3arm-DNA containing only Cy3 (without RC). The blue and red bars are the energy transfer efficiency values calculated with excitation of Cy3 and Cy5 respectively, using the 3arm-DNA containing both the dyes (Cy3-Cy5) without the RC attached as the fluorescence reference. The FRET efficiencies (E) from steady-state fluorescence data were calculated according to the following equation: $E = 1 - \frac{I_{DA}/A_{DA}}{I_D/A_D}$, where I_{DA} and I_D are the integrated area of fluorescence from the donor with and without an acceptor. A_{DA} and A_D are the absorbance of the donor at excitation wavelength with and without an acceptor. The energy transfer efficiencies ($E_{lifetime}$) from lifetime data were calculated according to the following equation: $E_{lifetime} = 1 - \frac{\tau_{ave,DA}}{\tau_{ave,D}}$, where $\tau_{ave,DA}$ and $\tau_{ave,D}$ are the average lifetime of the donor (Table 4.2) with and without an acceptor.

Time-resolved fluorescence analysis was performed using time-correlated single-photon counting (TCSPC) (Figures 4.4C, S4.9-S4.10) excited by a pulsed laser. The decay

traces of individual dye labeled 3arm-DNA (only one dye on the 3arm-DNA without the RC) could be fitted adequately with biexponential decay kinetics [96] [104] (Tables 4.2 and S4.1-3). The amplitude-weighted average lifetimes were 1.79 ns for Cy3, 1.65 ns for Cy5, 1.68 ns for AF660, and 0.64 ns for AF750. In contrast, fitting the fluorescence decays for each of the 3arm-DNA-dye-RC constructs required three or four exponential components (Tables 4.2 and S4.1-3). For example, considering the decay profiles of Cy3 in various samples (λ_{ex} = 510 nm and λ_{em} = 565 nm in Figure 4.4), a substantial increase in the fluorescence decay rate is observed for the constructs with the RC, e.g. the average lifetimes of **3C** and **3CC** are ~1.17 ns and ~0.25 ns, respectively. This follows the same trend as the steady-state energy transfer measurements and again implies that a significant amount of energy transfer takes place from the dye to the RC. Similar decay patterns were observed for the set of constructs with Alexa Fluor dyes (Figure S4.10). Based on the lifetime data for the dyes alone (without RCs) or one dye with the RC, the rate constants for the various component processes can be determined as described in the supplemental information. For example, the fluorescence decay rate constant for Cy3 alone (in the absence of Cy5 or RC) is measured to be 0.55 ns^{-1}, the rate constant for energy transfer from Cy3 to the RC is calculated to be 0.39 ns^{-1}, and the rate constant for energy transfer from Cy3 to Cy5 is calculated to be 1.45 ns^{-1}. If one uses the rate constants for these individual processes to calculate the decay lifetime of

Table 4.2: Fitting parameters for the Cy3 lifetime data in different constructs, monitored at 565 nm (□$_{ex}$ = 510 nm). The results from two replicates of each sample are shown.

[a]Average lifetime is calculated as $\tau_{ave} = \sum_i A_i\tau_i / \sum_i A_i$, where A_i is the amplitude of the i^{th} component and □$_i$ is the corresponding lifetime.

sample	□1 ns (amplitude %)	□2 ns (amplitude %)	□3 ns (amplitude %)	□4 ns (amplitude %)	□^{2}	average lifetime (ns)[a]
3arm-DNA-Cy3	0.63(34.9) 0.64(35.5)	2.41(65.1) 2.45(64.5)	- -	- -	1.18 1.17	1.788 1.807
3arm-DNA-Cy3-Cy5	0.06(59.8) 0.07(52.7)	0.40(22.9) 0.52(23.6)	2.15(17.3) 2.19(23.7)	- -	1.17 1.16	0.499 0.678
1C	0.12(14.1) 0.09(12.1)	0.68(45.0) 0.67(42.7)	1.80(40 9) 1.9(45.2)	- -	1.03 1.01	1.059 1.156
2C	0.12(12.8) 0.09(11.5)	0.71(43.3) 0.66(46.5)	1.89(43.9) 1.8(42.0)	- -	1.06 1.07	1.152 1.073
3C	0.10(12.5) 0.11(13.2)	0.70(42.3) 0.75(45.4)	1.90(45.2) 1.96(41.4)	- -	1.05 1.14	1.167 1.167
1CC	0.04 (50.9) 0.03(51.5)	0.15(28.7) 0.15(29.1)	0.59(11.5) 0.54(11.1)	1.86(8.9) 1.85(8.3)	1.05 1.04	0.297 0.272
2CC	0.04(49.0) 0.03(47.0)	0.14(28.1) 0.14(29.8)	0.53(12.7) 0.56(12.9)	1.81(10.2) 1.85(10.3)	1.00 1.07	0.311 0.318
3CC	0.04(51.4) 0.03(53.6)	0.14(29.2) 0.14(27.3)	0.51(11.4) 0.48(11.7)	1.79(8.0) 1.75(7.4)	1.07 1.02	0.263 0.240

Cy3 in the fully assembled complex (**1CC)**, it is predicted to be 0.42 ns, whereas the experimentally measured average lifetime is 0.28 ns. Similarly, experimentally observed lifetime of AF660 is 0.90 ns in **1-6-7**, and the predicted decay lifetime of AF660 in **1-6-7** is 0.92 ns (based on the measurements of the decay lifetime of AF660 alone, the energy transfer rate constants from AF660 to AF750 and from AF660 to RC). The approximate agreement of the experimentally measured decay times for the full nanostructures and the predicted values based on the kinetic constants for individual component reactions

indicates that the experimental measurements are internally consistent with each other, and consistent with an overall picture of step-wise energy transfer.

Further evidence of a stepwise energy transfer process is provided by the initial rise of the Cy5 and AF750 emission in the TCSPC experiment in the two dye complexes, upon excitation of Cy3 and AF660, respectively (Figures S4.9-S4.10). The Cy3 or AF670 in all cases shows an instantaneous increase of the emission upon direct excitation (which is convoluted with the instrument response function), the rise of the Cy5 or AF750's emission without RC is much slower than the instrument response. This is due to the energy transfer from the initial donor (Cy3 or AF660) to the intermediate dye (Cy5 or AF750) on the sub-nanosecond time scale, which results in an initial increase in the excited-state population of the intermediate. In the presence of the RC, Cy5 or AF750 show a much faster decay. A comparison of the average lifetimes of the dyes in the 3arm-DNA-RC constructs vs. that in the 3arm-DNA structures (without RC) result in estimated energy-transfer efficiencies from the dyes to the RC (Figures 4.5B and S4.11B) which are in reasonable agreement with the results obtained from the steady-state fluorescence intensity measurements (Figures 4.5A and S4.11A). Like the steady-state measurements, similar energy-transfer efficiencies are observed for samples with different numbers of DNA-dye constructs per RC. Again, in the case of time-resolved measurements, higher energy-transfer efficiency is observed for constructs that contain Cy5 compared to AF750, even though the fluorescence spectrum of AF750 overlaps better with the absorbance of the RC than does Cy5. This can be explained by the fact that AF750 has a shorter excited state lifetime (0.64 ns) than Cy5 (1.64 ns), which gives the excited state of Cy5 a greater probability of transferring energy to the RC before decaying to the ground state by other pathways.

Similar results were obtained previously when dye molecules with different lifetimes were conjugated directly to the RC [104].

4.3.4 Enhancement of Reaction Center Charge Separation

Because charge separation in the RC has an almost unity yield, the amount of charge separation that takes place correlates with the energy transfer efficiency [104]. The relative amount of charge separation in the RC was investigated by measuring the light-minus-dark difference absorbance spectra of the different dye-DNA-RC complexes. The light-minus-dark difference spectra were obtained by substracting the absorbance spectrum of a sample taken in the dark from the absorbance spectrum taken under continuous illumination at 550 nm (Cy3 absorbance peak, 10 nm bandwidth). The light intensity at 550 nm was kept low enough to ensure the light-minus-dark signals changed linearly with the light energy absorbed. Under low light conditions, no RC is excited more than once during the ~100 ms lifetime of $P^+Q_A^-$, avoiding artifacts due to photopumping. A 1.3 fold absorbance change at 862 nm (reflecting P^+ formation) was observed for **3C** compared to the RC alone, implying enhanced charge separated state formation due to the increased absorbance cross section at 550 nm, confirming that photons absorbed by Cy3 result in energy transfer to RC cofactors (Figure 4.6). Similarly, **3CC** shows a 1.8 fold enhancement in P^+ formation over unconjugated RCs. The enhanced P^+ formation in **3CC** compared to **3C** presumably results from the higher efficiency of the overall stepwise energy transfer from Cy3 to Cy5 to the RC, compared to direct transfer from Cy3 to the RC (Figure 4.5). The insertion of Cy5 between Cy3 and the RC results in two relatively efficient transfer steps (better spectral overlap and shorter distance) compared to the single Cy3 to RC transfer. As with the energy transfer efficiency results obtained from both the steady state

and the time resolved fluorescence measurements, the relative intensity of P^+ formation is similar for samples with different numbers of 3arm-DNA nanostructures conjugated to each RC (i.e., similarity among the samples **1C**, **2C** and **3C** or among the samples **1CC**, **2CC** and **3CC**).

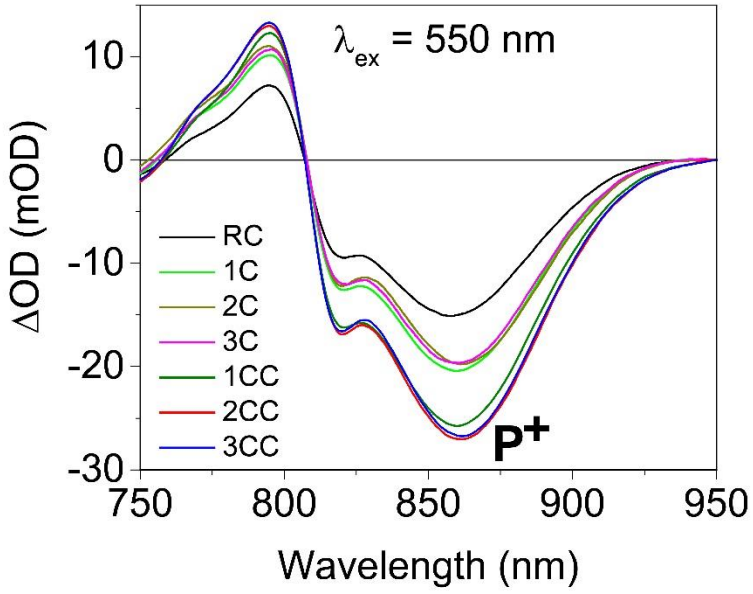


Figure 4.6: Light-minus-dark difference absorbance spectra of RCs with and without conjugation to a 3arm-DNA nanostructure-dye complex.

In the natural system, the RC operates in conjunction with the cytochrome bc_1 complex, cytochrome c_2, and a quinone pool, to convert light energy into a proton motive force [113]. In this process, the oxidized initial electron donor of the RC, P^+, that is formed upon light-driven electron transfer is subsequently reduced by cytochrome c_2, which docks to the periplasmic face (P side) of the RC. In our artificial antenna system, the 3arm-DNA structures are located on the P side of RC, and so one might expect that this conjugation of DNA close to the docking site of cytochrome would hinder cytochrome binding as well as the electron transfer process from cytochrome to P^+. To explore this possibility, a 10-fold

molar excess of reduced cytochrome *c* [114] and a 100-fold molar excess of decylubiquinone were added into a solution of 3arm-DNA-dye-RC constructs, and the absorbance intensity change at 550 nm (an absorbance decrease at this wavelength reflects the oxidation of cytochrome *c*) was measured, while either exciting the RC directly or the dye directly [104, 105] [115].Using 800 nm excitation (direct excitation of the RC), where both the Cy3 and Cy5 have no absorbance, the wild type RC, the Cys-modified RC, and the RC conjugated with the DNA-dye construct all showed similar rates of cytochrome *c* oxidation (Figure 4.7A). Apparently, DNA conjugation does not hinder the rate of cytochrome electron transfer to the RC, at least at these concentrations. However, upon 650 nm excitation (Cy5 excitation peak), the DNA-dye conjugated RC showed a much faster rate of oxidation then did the Cys-modified RC or wild type RC, both of which have very low absorbance at 650 nm (Figure 4.7B). It is interesting to note that under the conditions of this kinetic measurement, the oxidation rate of cytochrome *c* depends on the number of dye molecules in the construct. This presumably results from the enhanced absorbance cross-section of the light harvesting antenna that increases the number of photons absorbed per unit time by the 3arm-DNA-dye-RC complex. The cytochrome *c* oxidation experiment is real time and reports the accumulative result (i.e. integration of the change over time). Since the spectrum of reduced cytochrome *c* overlaps strongly with that of Cy3, making difficult to quantitate the number of photons absorbed by Cy3, similar measurements using 550 nm excitation were not attempted.

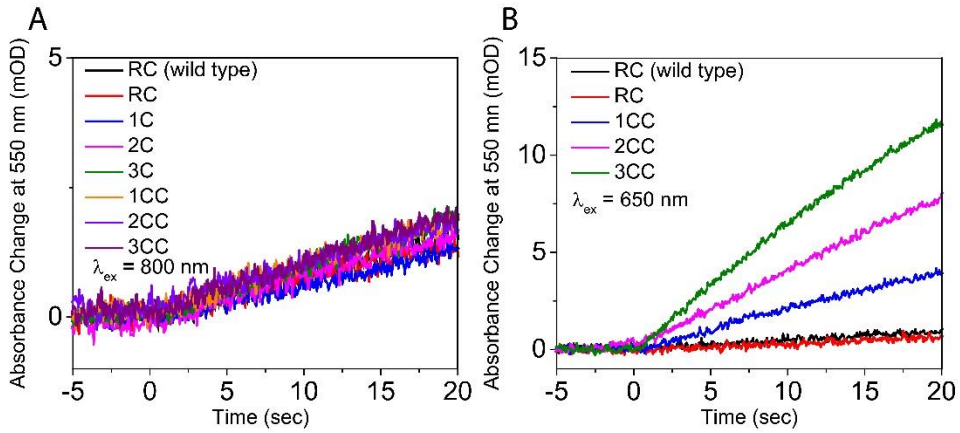


Figure 4.7: Cytochrome *c* oxidation monitored at 550 nm (where the difference in absorbance between reduced and oxidized cytochrome *c* is maximal) after exciting the RC directly at 800 nm (A) or Cy5 directly at 650 nm (B).

4.4 Conclusion

A DNA nanostructure with dyes attached at specific positions was conjugated to a RC to serve as a geometrically defined light harvesting antenna. This extended the absorbance cross section of the complex into a spectral range where the RC has only weak absorbance. A combination of factors including the spatial placement, spectral properties and excited state kinetic properties of the dyes used are important in determining the efficiency of the antenna in energy transfer. At low light flux, the rate of photon capture by the complex is proportional to the number of dye molecules in the complex that absorb at the excitation wavelength; thus increasing the number of DNA-dye constructs attached to the reaction center increases the functional cross section but does not greatly change the energy transfer efficiency. The complexes explored in this work provide useful model systems for future applications in nanophotonics.

4.5 Supplemental Material

Supporting Information: Methods, calculations, gel electrophoresis, DNA sequences, additional spectral data, and DNA synthesis and modification characterization. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION:

Corresponding Authors

Hao.Yan@asu.edu

nwoodbury@asu.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS:

We thank Douglas Daniel for his help in setting up the spectroscopy for light-minus-dark experiment. This research was supported by Multidisciplinary University Research Initiative (MURI) program (award W911NF-12-1-0420) funded by Army Research office (ARO) to H.Y. and N.W.W. and a grant from the Canadian Natural Sciences and Engineering Research Council to J.T.B.

4.5.1 Reaction Center Protein Preparation

4.5.1.2 Reaction center mutations

Among a total of eight mutations in the RC, five of them serve to replace the five wild type cysteines with serine and alanine, and the remaining three mutations introduce cysteines on the P side of the RC, by replacing wild type amino acids (glutamic acid or asparagine) with Cys on the surface near P. The mutations are as follows: (H)C156A,

(H)C234S, (M)E100C, (L)C92S, (L)C108S, (L)C247S, (L)E72C and (L)N274C. In addition, the engineered RC contains a six-histidine tag at the C-terminus of the H subunit, to facilitate purification with a Ni-sepharose affinity column [116].

4.5.1.3 RC isolation and purification

RCs were isolated from *R. sphaeroides* 2.4.1 [112] containing a pRK-based expression plasmid encoding the modified RC puf operon. 2 L of modified LB medium, containing 810 □M $MgCl_2$, 510 □M $CaCl_2$ and 4 mM NaCl, was used to grow cells at 30°C for 3.5 days. The cells were pelleted and resuspended in 50 mM phosphate buffer (pH 8) containing 150 mM NaCl. The cells were then lysed by passing through a French press, followed by addition of small amount of DNase. After removal of any unbroken cells and large cell debris via centrifugation (9000 g for 10 minutes), the remaining supernatant was treated with imidazole (final concentration 5 mM) and the RC protein was solubilized by adding N,N-Dimethyldodecylamine N-oxide (LDAO, final concentration 0.4% by volume). After 20 min incubation at 22°C, the solution was centrifuged at 14000g followed by Ni-sepharose column purification. The eluted RC was dialyzed overnight at 4°C against dialysis buffer (15 mM Tris, 0.025% LDAO, 150 mM NaCl, 1 mM EDTA, pH 8) using 50 kD molecular weight cutoff membrane (Amicon), to remove imidazole and excess LDAO. The concentration of the purified RC was measured using absorbance at 804 nm (ε ~288000 $M^{-1}cm^{-1}$) [117].

4.5.2 RC-DNA Conjugation and Purification

4.5.2.1 SPDP labeling of DNA

An amine-modified DNA (Strand 1, 5'-TCGCTAGGAACGG ATTTT-NH_2-3') of ~400 □M in 1×PBS, pH 7.6 was treated with 20 fold excess of 50 mM SPDP (N-succinimidyl 3-(2-pyridyldithio) propionate) in dimethyl sulfoxide (DMSO), followed by addition of 1M $NaHCO_3$ (~1/10 of total volume of DNA-SPDP mixture, to adjust pH) and the mixture was shaken gently for 3 hours at room temperature. The DNA-SPDP conjugate was purified with Nap-10 desalting column (GE Healthcare) and then washed 3 times with 1×PBS using 3kD molecular weight cut-off filter (Amicon) to remove the excess SPDP.

4.5.2.2 Reduction of the disulfide bond in RC

Before conjugation, the RC was treated with 8 fold excess of 50 mM TCEP-HCl (Tris(2-carboxyethyl)phosphine hydrochloride) for 30 min at 4°C, followed by washing with 1×PBS, 0.025% LDAO, pH 8 using 50kD molecular weight cut-off filter (Amicon) to remove excess TCEP-HCl.

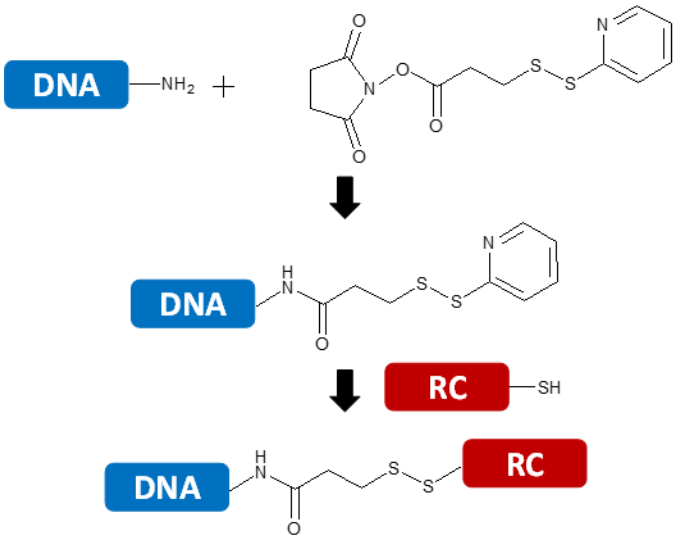
4.5.2.3 SPDP mediated cross-linking of DNA and RC

A 10 fold excess of DNA-SPDP conjugate was mixed with TCEP-HCl treated RC and left for ~6 hours at 4°C with gentle mixing (Scheme S1). Then the mixture was treated with 10 mM phosphate buffer with high salt (1.5 M NaCl, 0.025% LDAO, pH 8), followed by washing 3 times with 10 mM phosphate, 0.025% LDAO, pH 8 buffer to remove the NaCl.

4.5.2.4 Purification of the RC-DNA conjugates

The sample was then run through an anion exchange column (Mono Q 4.6/100 PE, product code-17-5179-01) using a fast protein liquid chromatography (FPLC) system (AKTA purifier). The desired fractions containing the RC-DNA conjugates with different protein:DNA ratios were washed with dialysis buffer as described previously. The composition of the equilibration buffer used was 10 mM phosphate, 0.025% LDAO, pH 8 and the elution buffer consisted of 10 mM phosphate, 1M NaCl, 0.025% LDAO, pH 8.

Scheme S4.1: RC-DNA conjugation using SPDP as bi-specific cross-linker.



Scheme S4.2: DNA-Alexa Fluor dye conjugation

4.5.3 DNA-dye conjugation and purification

Cy3 and Cy5 labeled strands (HPLC purified) (5'-CGCTACATCA/iCy3/TCCTAGCGA-3' and 5'-/5Cy5/ATCCGTTGATGTAGCG-3') were purchased from IDTDNA and used as received. Alexa Fluor dye (AF660 and AF750) labeled DNA strands were prepared as follows.

4.5.3.1 Synthesis of amine-modified DNA and purification

Amine modified DNAs for dye conjugation were synthesized on a DNA synthesizer (ABI 394 DNA/RNA Synthesizer, Applied Biosystems) via standard protocols by using CPGs (1 □mole scale) with a coupling time of 5 min for amine modified phosphoramidite (amino-modifier C6 dT phosphoramidite for Strand 3 and 5'-amino-modifier C6 phosphoramidite for Strand 2; both purchased from Glen Research). The oligonucleotide was cleaved from the resin by treatment with 1:1 volume mixture of NH_4OH (28% in water) and methylamine (40% in water) for 2 hours at 50°C, and then purified using HPLC (Agilent Technologies 1200 series) with a Phenomenex-C18 column (Solvent A: 100 mM triethylammonium acetate, pH 7; Solvent B: acetonitrile; Flow rate: 4 mL/min). The fractions containing the desired oligonucleotides were collected and lyophilized. After being redissolved in water, the lyophilized fractions were precipitated in 70% cold ethanol. The pellet of oligonucleotide was washed with 70% ethanol and dried under vacuum, and then dissolved in 0.1 M sodium tetraborate buffer ($Na_2B_4O_7.10H_2O$, pH 8.5) to a final concentration of ~200 □M.

4.5.3.2 Dye-DNA conjugation and purification

A 10-fold excess of Alexa Fluor dye (Invitrogen, amine reactive Alexa Fluor 660 and -750) from a ~15 mM stock solution (dissolved in DMSO) was added to the DNA

solution described above and incubated overnight with gentle shaking at room temperature (Scheme S2). The DNA was then precipitated using 3 M NaCl and ethanol, and pelleted. The pellet was dissolved in water followed by HPLC purification (as described above for the amine modified DNA). The fraction containing the Dye-DNA conjugate was collected and lyophilized.

4.5.3.3 Characterization of the dye-DNA conjugate

MALDI-mass spectrometry (Applied Biosystem Voyager System 4320 and Bruker Microflex) analysis was carried out before and after the dye conjugation, using 3-hydroxypicolinic acid as the matrix (Figure S4.1).

4.5.4 3arm-RC preparation

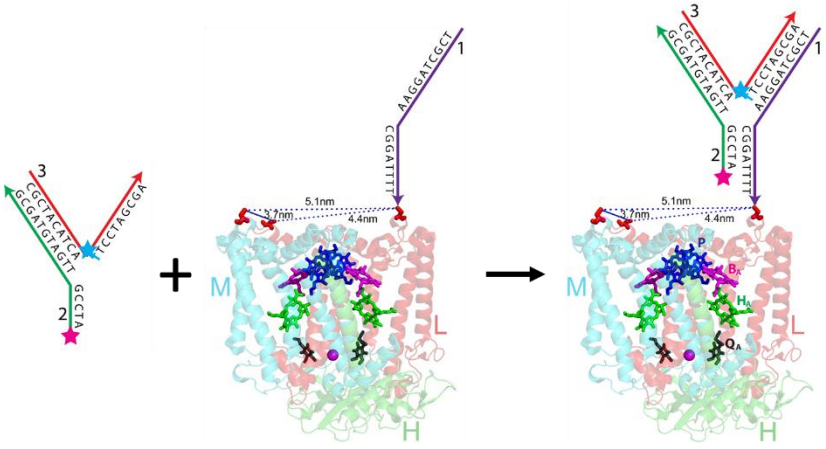
4.5.4.1 Free 3arm-DNA constructs

Free 3arm-DNA constructs were prepared by mixing stoichiometric quantities of three DNA strands in TAE/Mg^{2+} buffer (40 mM Tris, 20 mM Acetic acid, 2 mM EDTA, 12.5 mM Mg^{2+}, pH 8) and subsequent annealing from 90°C to 10°C. After annealing the structures were purified by 8% native PAGE (polyacrylamide gel electrophoresis) and transferred into Tris buffer (15 mM Tris, 20 mM Mg^{2+}, 150 mM NaCl, 1 mM EDTA, pH 8). The stoichiometric formation of the 3arm-DNA constructs were confirmed by native PAGE (Figure S2).

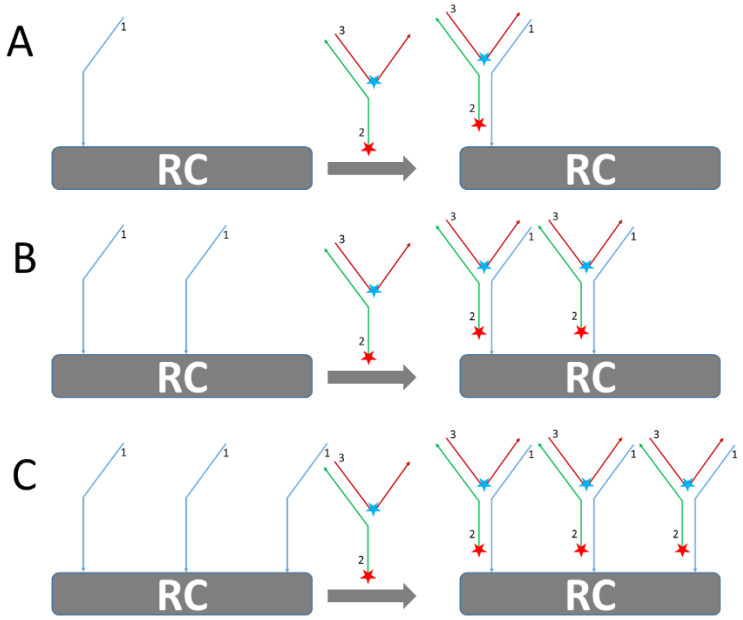
4.5.4.2 3arm-RC conjugate

First, strand-2 and -3 were annealed in the Tris buffer described above from 90°C to 10°C and then mixed with DNA conjugated RC (strand-1 conjugated with RC), with 1.5 fold molar excess followed by annealing from 30°C to 10°C over a 12 hr period (Scheme S4.3-S4). The mixture was then purified using a 50kD molecular weight cut-off filter (Amicon) using Tris buffer described above plus 0.025% LDAO, to remove the excess DNA strands.

Scheme S4.3: Schematic showing preparation of 3arm DNA-RC conjugate.



Scheme S4.4: Schematic representation of formation of the RC conjugate with different ratios of 3arm DNA. (A) For 1CC or 1-6-7. (B) For 2CC or 2-6-7. (C) For 3CC or 3-6-7.



4.5.5 Spectroscopic Analysis

4.5.5.1 Absorbance and fluorescence spectroscopy

Absorbance spectra were measured using a quartz cell with 1 cm path length in a Jasco V-670 spectrophotometer. Steady state fluorescence spectra were measured in a Nanolog Fluorometer (Horiba Jobin Yvon), with a quartz cuvette of 1 cm path length. All the steady state emission spectra were corrected for the wavelength dependence of the response of the detection system.

4.5.5.2 Time-correlated single-photon counting measurements

Fluorescence lifetime measurements were analyzed by time-correlated single-photon counting (TCSPC). A fiber supercontinuum laser (Fianium SC450) was used as the excitation source, with a repetition rate of 20 MHz. The laser output was sent through an Acousto-Optical Tunable Filer (Fianium AOTF) to obtain excitation pulses at wavelengths

of 510 nm, 600 nm, 620 nm and 740 nm. Fluorescence emission was collected at a 90° geometry setting and detected using a double-grating monochromator (Jobin-Yvon, Gemini-180) and a microchannel plate photomultiplier tube (Hamamatsu R3809U-50). The polarization of the emission was 54.7° relative to that of excitation. Data acquisition was done using a single photon counting card (Becker-Hickl, SPC-830). The typical instrument response function had a full width half maximum of 50 ps, measured using light scattered from the sample at the excitation wavelength. The data were fitted using a locally written software package ASUFIT.

4.5.5.3 Calculation of FRET efficiency, average lifetime of dye molecules and decay rate constants

FRET efficiencies (E) were calculated according to the following equation:

$$E = 1 - \frac{I_{DA}/A_{DA}}{I_D/A_D} \qquad (1)$$

Where I_{DA} and I_D are the integrated area of fluorescence from the donor with and without an acceptor. A_{DA} and A_D are the absorbance of the donor at the excitation wavelength with and without an acceptor.

The average lifetime was calculated using the following equation.

$$\tau_{ave} = \frac{\sum_i A_i \tau_i}{\sum_i A_i} \qquad (2)$$

Where A_i is the amplitude of the i^{th} exponential component in the fit and τ_i is the corresponding lifetime.

The energy transfer efficiency calculated from the lifetime measurements was determined as:.

$$E_{lifetime} = 1 - \frac{\tau_{ave,DA}}{\tau_{ave,D}} \qquad (3)$$

Where $\tau_{ave,DA}$ and $\tau_{ave,D}$ are the average lifetimes of the donor with and without an acceptor obtained from the TCSPC data.

The average lifetime (τ1) determined for Cy3 in the 3arm DNA-Cy3 molecules is 1.79 ns (Table 2).

$$\tau 1 = \frac{1}{k_{r,Cy3} + k_{nr,Cy3}} \qquad (4)$$

Where $k_{r,Cy3}$ and $k_{nr,Cy3}$ are the radiative and nonradiative decay rate constants of Cy3. Thus, $k_{r,Cy3} + k_{nr,Cy3} = 0.55\ ns^{-1}$.

In the case of 3arm DNA-Cy3-Cy5, the measured average lifetime of Cy3 (τ2) is 0.50 ns (Table 2).

$$\tau 2 = \frac{1}{k_{Cy3-Cy5} + k_{r,Cy3} + k_{nr,Cy3}} \qquad (5)$$

Where $k_{Cy3\text{-}Cy5}$ is the rate constant for Cy3 to Cy5 energy transfer. By combining (4) and (5), $k_{Cy3\text{-}Cy5}$ can be determined as 1.45 ns^{-1}.

In the case of **1C**, the average lifetime of Cy3 (τ3) is 1.06 ns (Table 2).

$$\tau 3 = \frac{1}{k_{Cy3-RC} + k_{r,Cy3} + k_{nr,Cy3}} \qquad (6)$$

Where $k_{Cy3\text{-}RC}$ is the rate constant for Cy3 to RC energy transfer. Combining (4) and (6) gives 0.39 ns^{-1} for $k_{Cy3\text{-}RC}$.

Similarly for **1CC**, the average lifetime of Cy3 (τ4) is

$$\tau 4 = \frac{1}{k_{Cy3-RC}+k_{Cy3-Cy5}+k_{r,Cy3}+k_{nr,Cy3}} \qquad (7)$$

Based on the values determined for the microscopic rates in the denominator of this expression, one would expect τ4 to be 0.42 ns. The experimentally observed lifetime of Cy3 in **1CC** is 0.28 ns (Table 2).

Values of $k_{AF660\text{-}RC}$, $k_{AF660\text{-}750}$, and ($k_{r,AF660}$+$k_{nr,AF660}$) can be calculated in an analogous manner and are 0.20 ns^{-1}, 0.30 ns^{-1}, 0.59 ns^{-1}. Based on these values, the calculated lifetime of AF660 in **1-6-7** should be 0.92 ns, in close agreement with the experimental value of 0.90 ns.

The fact that the microscopic rate constants estimated and the observed average lifetimes are internally consistent supports the kinetic model used and the resulting energy transfer efficiencies determined.

4.5.5.4 Cytochrome *c* oxidation experiment

Before measuring the cytochrome *c* oxidation kinetics, bovine heart cytochrome *c* was reduced by treating with a 10-fold molar excess of sodium ascorbate in 10 mM sodium phosphate buffer (pH 6.9) [114], followed by purification with Nap-25 column (GE Healthcare). The oxidation kinetics of cytochrome *c* in presence of the 3arm DNA-RC were measured by monitoring the change in the absorbance at 550 nm in the presence of a 650

nm or 800 nm excitation beam (Figure S4.12). The 800 nm excitation beams was generated by passing white light (Dolan-Jenner MH-100 Metal Halide Fiber optic illuminator) through an 800 nm band pass filter (FB800-40, FWHM 40 nm). The 650 nm excitation beam was generated by passing the white light beam through both an RG610 (long pass) and a IF650 (band pass, FWHM 10 nm) filter. The sample contained 0.1 □M RCs (RC-wild type, RC, **1C**, **2C**, **3C**, **1CC**, **2CC** and **3CC**), 100-fold molar excess of decylubiquinone (extinction coefficient at 409 nm in ethanol = 343 $M^{-1}cm^{-1}$) and 10-fold molar excess of reduced cytochrome *c* in the dialysis buffer described in section 2 in part I.

4.5.5.5 Light-minus-dark measurements

The light-minus-dark measurements were performed by measuring the absorbance spectra of a sample taken in the dark (dark spectra) and in presence of 550 nm (bandwidth ~10 nm) continuous light (light spectra), and then subtracting the dark spectra from the light spectra. The samples were illuminated with 550 nm light for 6 minutes prior to the measurement. The path of the excitation light was perpendicular to the path of the probe light from the UV-Vis absorbance spectrophotometer. The excitation light at 550 nm was obtained by passing a white light source (Dolan-Jenner MH-100 Metal Halide Fiber optic illuminator) through two filters (BG 38 and IF550, 10 nm band pass). For all measurements, samples contained a 50-fold excess of 1,10-phenanthroline compared to the RC concentration.

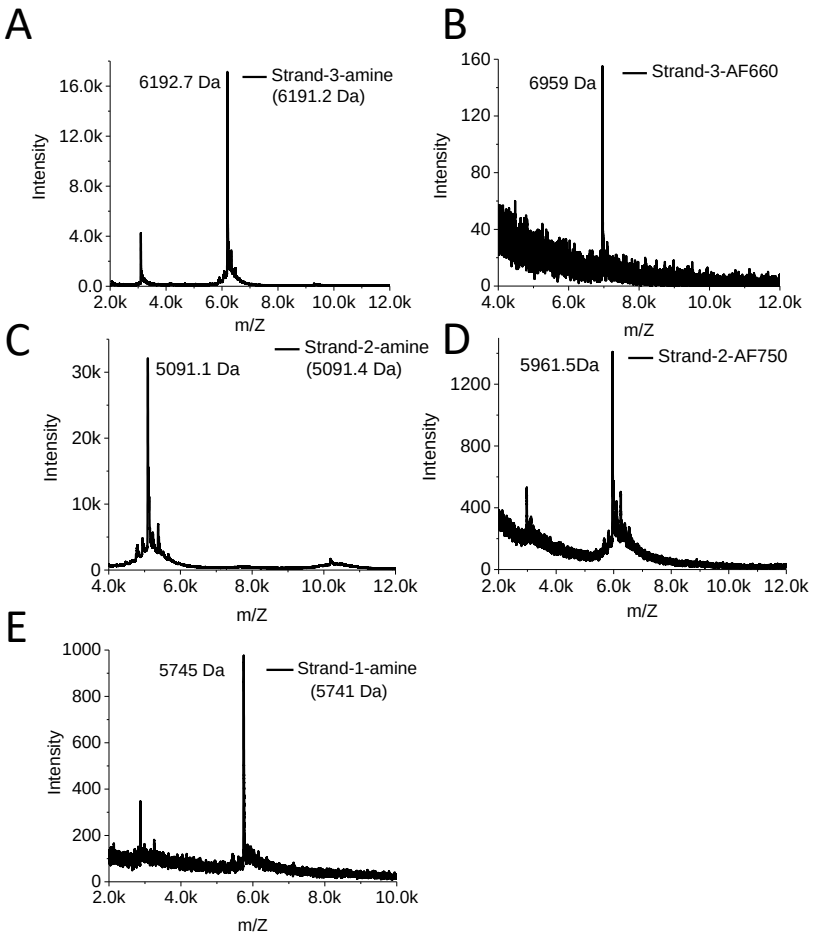


Figure S4.1: MALDI-TOF spectra of (A) amine modified Strand-3, (B) Alexa Fluor 660 conjugated Strand-3, (C) amine modified Strand-2, (D) AF750 conjugated Strand-2, and (E) amine modified Strand-1.

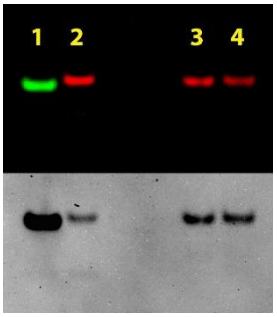


Figure S4.2: Images of an 8% nondenaturing polyacrylamide gel electrophoresis (PAGE). (Top) Gel image measured with a Typhoon™ Trio multifunction imager (Amersham Biosciences) exciting at 532 nm and 633 nm with emission at 580 nm and 670 nm, respectively. (Bottom) Ethidium bromide stained gel image. (1), (2), (3) and (4) represent the purified 3arm labeled with Cy3, Cy3-Cy5, AF660 and AF660-AF750, respectively.

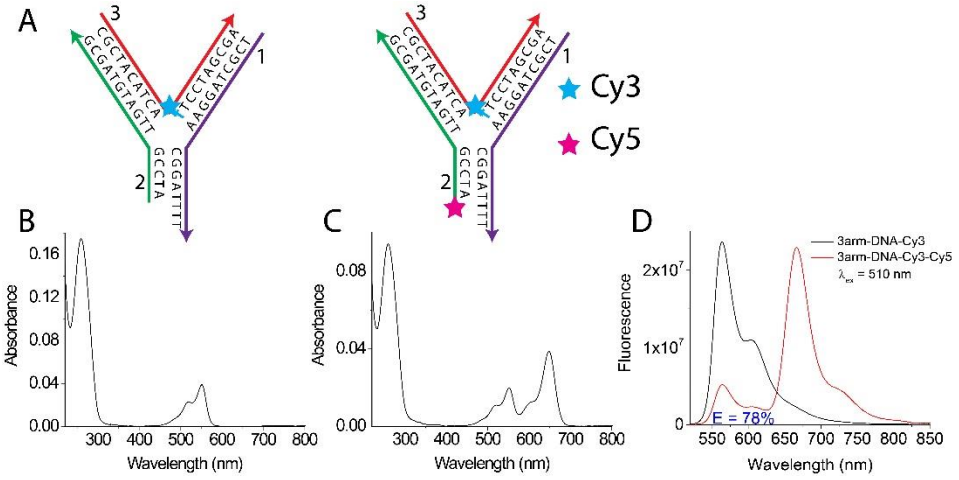


Figure S4.3: (A) Schematic of 3arm-DNA structure with Cy3 only (3arm-DNA-Cy3) and with both Cy3 and Cy5 (3arm-DNA-Cy3-Cy5). (B)-(C) Absorption spectra of 3arm-DNA-Cy3 and 3arm-DNA-Cy3-Cy5. (D) Corresponding fluorescence emission spectra with excitation at 510 nm. The spectra were corrected by adjusting for the independently determined wavelength-dependent detector response and normalized by dye absorbance at

510 nm. A 78% energy transfer efficiency was observed from Cy3 to Cy5 organized within the 3arm-DNA nanostructure.

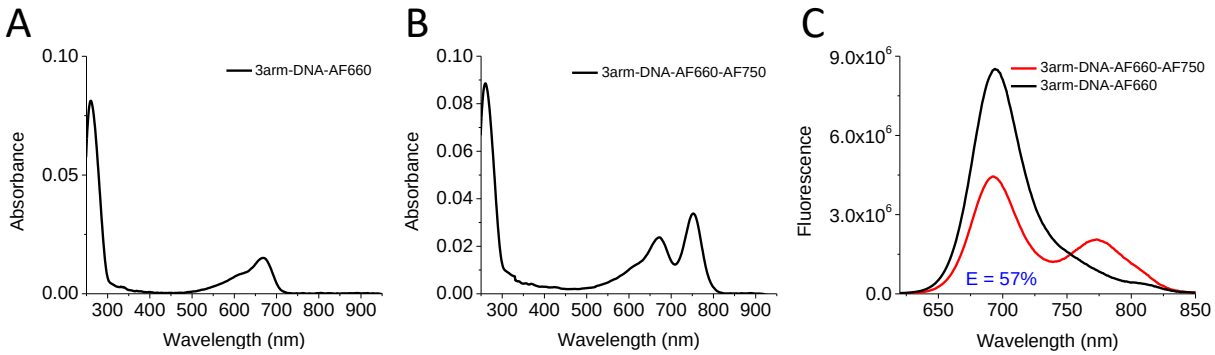


Figure S4.4: (A)-(B) Absorption spectra of 3arm-DNA-AF660 and 3arm-DNA-AF660-AF750. (C) Corresponding fluorescence emission spectra with excitation at 600 nm. The spectra were corrected for the wavelength dependence of the detector sensitivity and normalized by dye absorbance at 600 nm. A 57% energy transfer efficiency was observed from AF660 to AF750 organized within the 3arm DNA nanostructure.

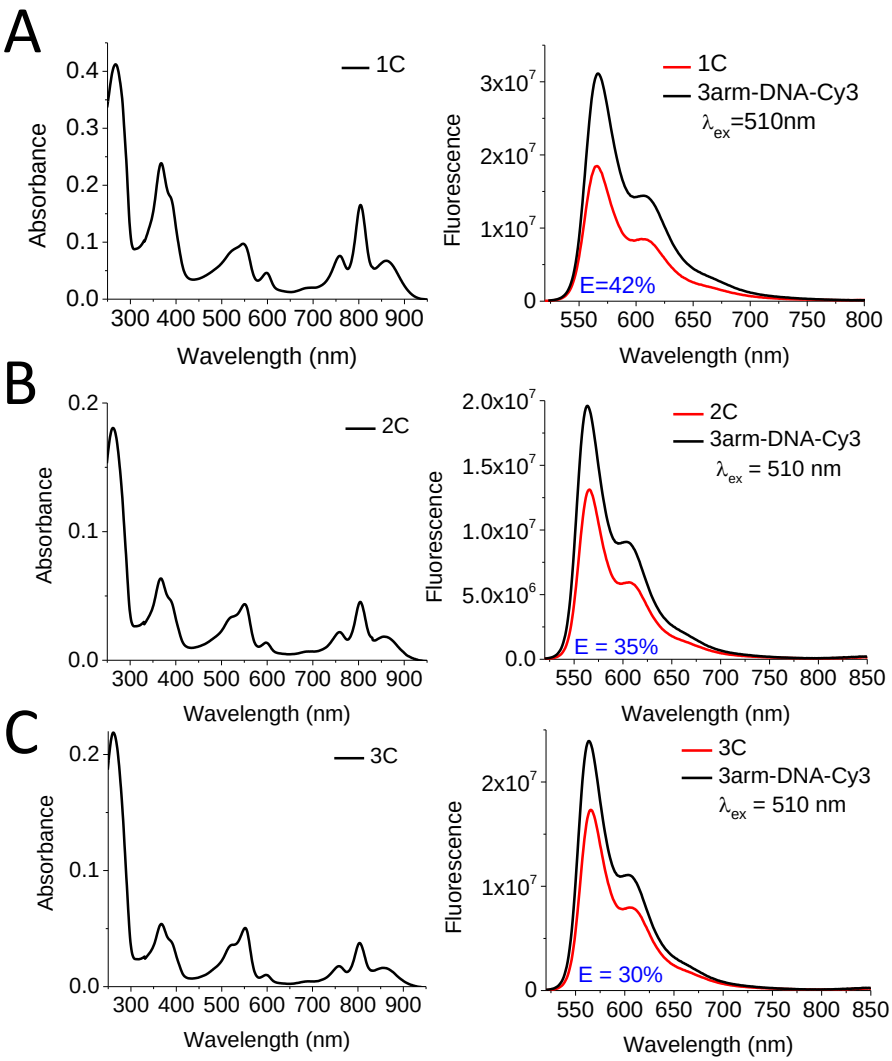


Figure S4.5: Absorbance spectra (left) and fluorescence spectra (right) of RCs with different numbers (1-3) of the 3arm-DNA-Cy3 complexes attached per RC. The energy transfer efficiency (E) values between the Cy3 and the RC are shown in the fluorescence spectra in blue.

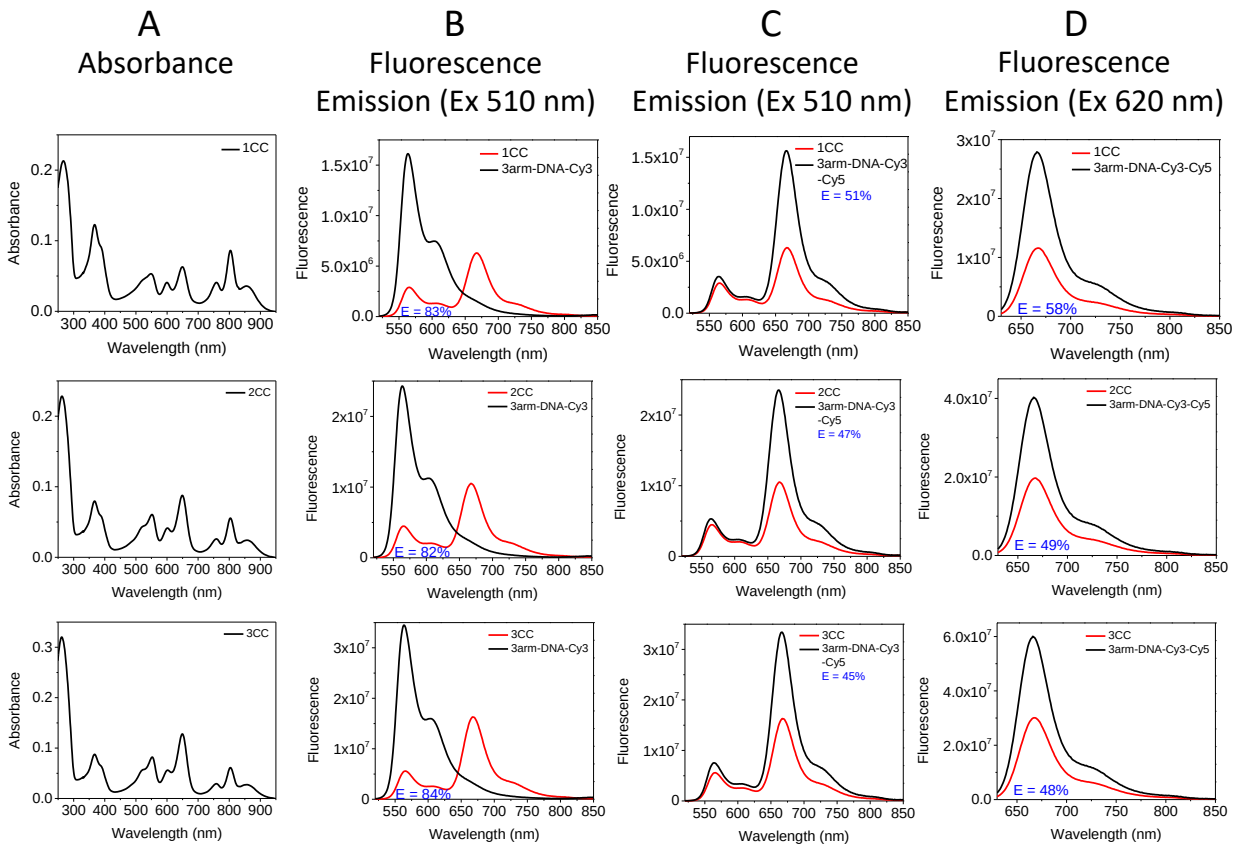


Figure S4.6: Absorbance spectra (panel A) and fluorescence spectra (panel B, C and D) of RCs with different numbers of 3arm-DNA-Cy3-Cy5 complexes per RC. The energy transfer efficiency (E) values are shown on the fluorescence spectra in blue.

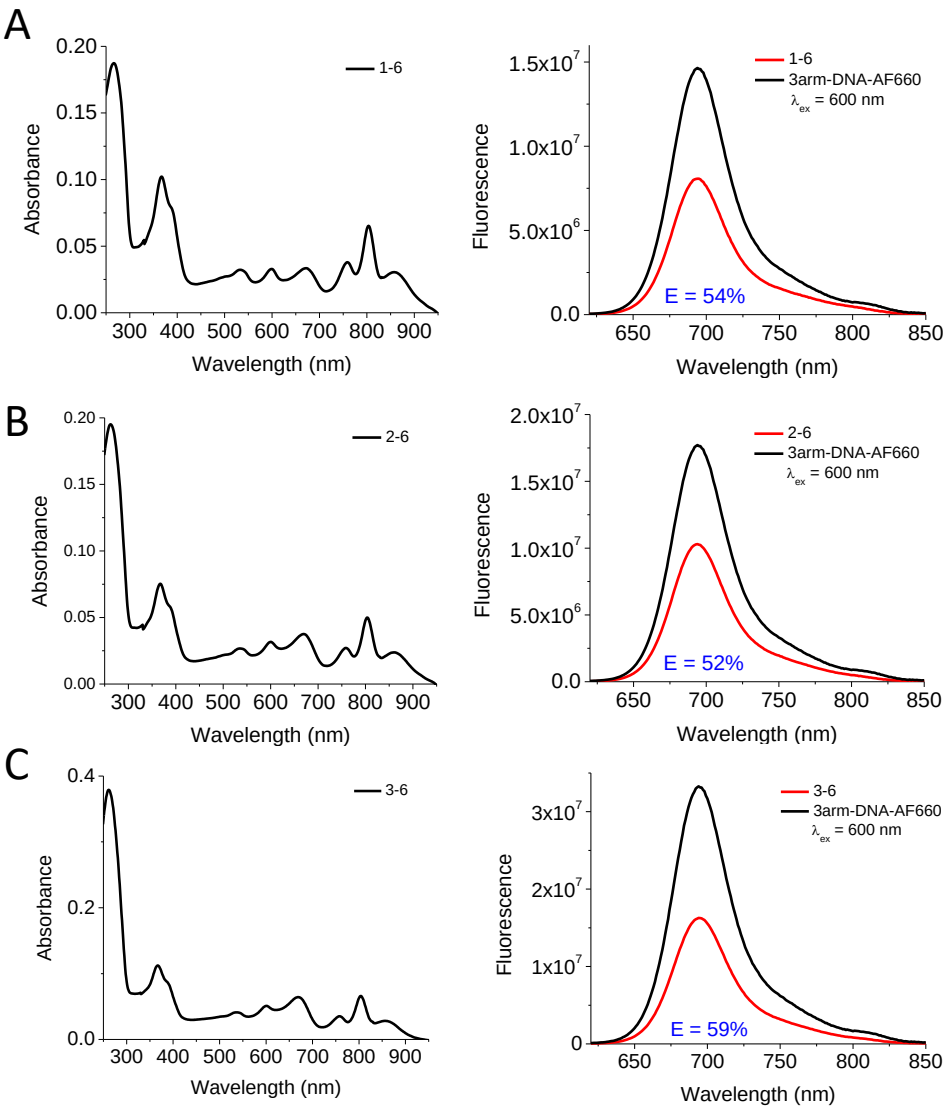


Figure S4.7: Absorbance spectra (left) and fluorescence spectra (right) of RCs with different numbers of 3arm-DNA-AF660 complexes per RC. The energy transfer efficiency (E) values are shown on the fluorescence spectra in blue.

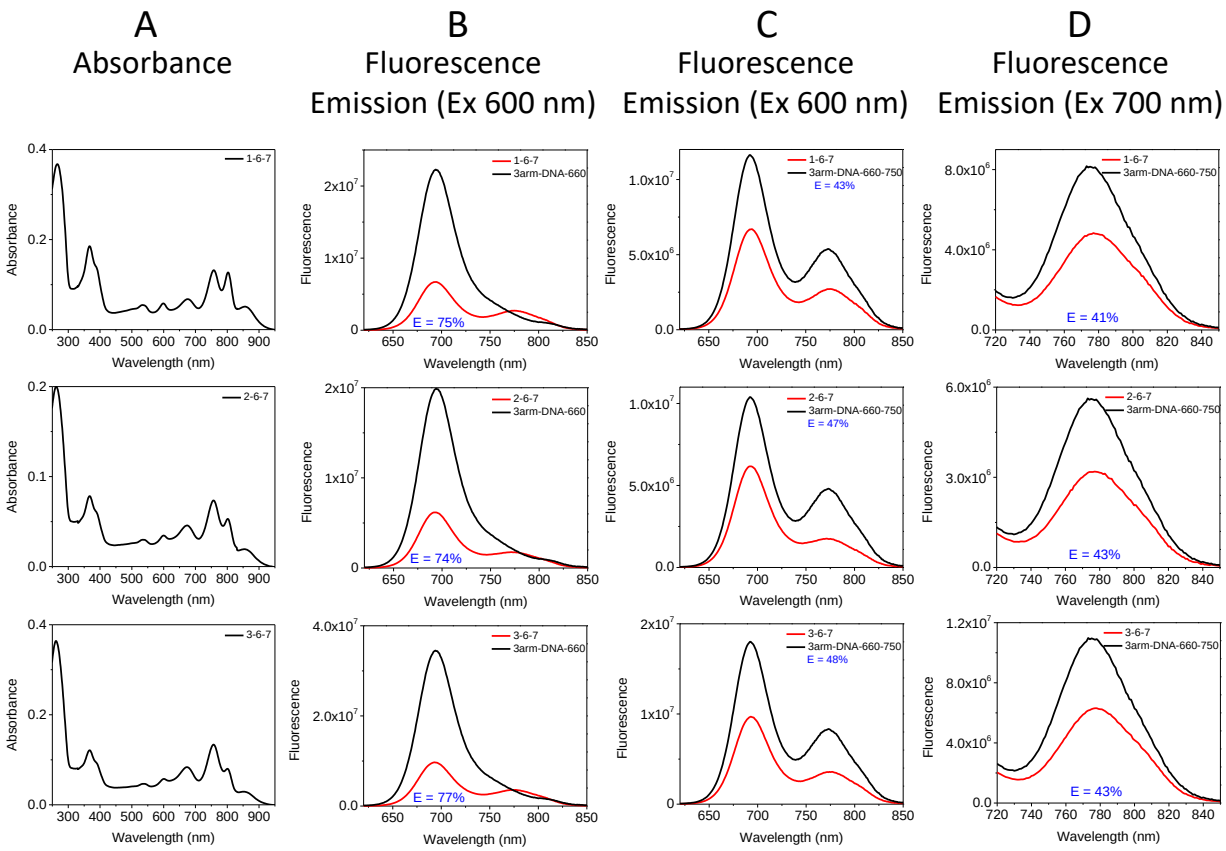


Figure S4.8: Absorbance spectra (panel A) and fluorescence spectra (panel B, C and D) of RCs with different numbers of 3arm-DNA-AF660-AF750 per RC. The energy transfer efficiency (E) values are shown on the fluorescence spectra in blue.

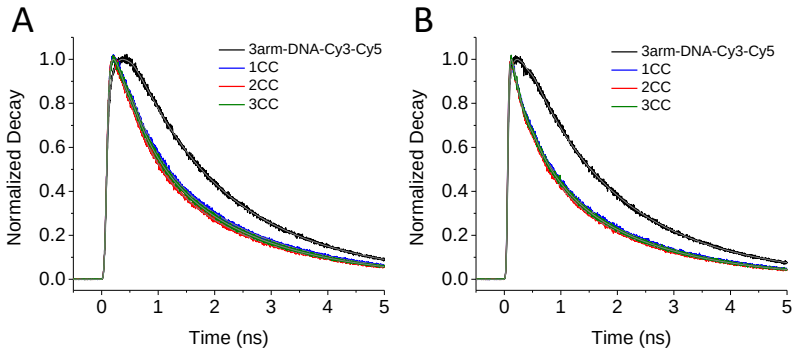


Figure S4.9: Time resolved emission of 3arm-DNA-Cy3-Cy5 with and without the RC. (A) Cy5 emission was monitored at 668 nm after exciting Cy3 at 510 nm. (B) Cy5 emission monitored at 668 nm after exciting Cy5 at 620 nm.

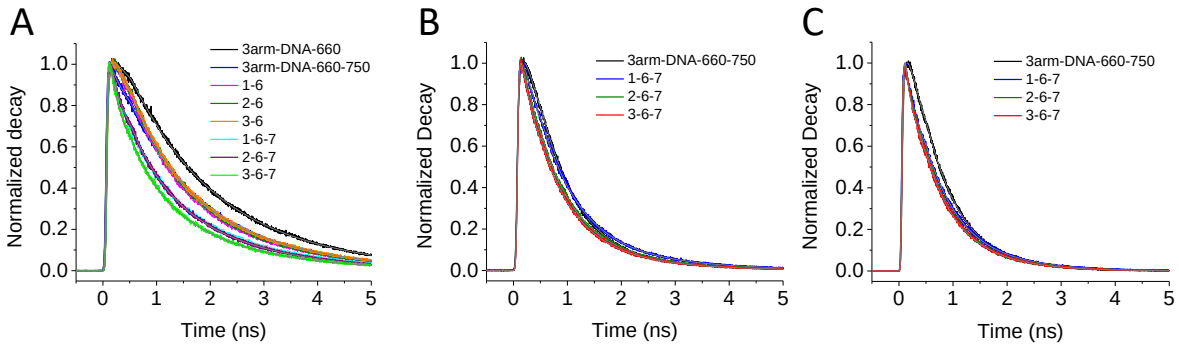


Figure S4.10: Time resolved emission of 3arm-DNA-AF660 and 3arm-DNA-AF660-AF750 samples with and without RCs. (A) AF660 emission was monitored at 698 nm after exciting AF660 at 600 nm. (B) AF750 emission was monitored at 780 nm after exciting AF660 at 600 nm. (C) AF750 emission was monitored at 780 nm after exciting AF750 at 740 nm.

Table S4.1: Fitting parameters for Cy5 lifetime data, monitored at 668 nm (λ_{ex} = 620 nm). The results from two replicates of each sample are shown.

sample	τ1 ns (amplitude %)	τ2 ns (amplitude %)	τ3 ns (amplitude %)	χ^2	average lifetime (ns)
3arm-DNA-Cy3-	0.77(22.8)	1.92(77.2)		1.13	1.658
Cy5	0.84(26.2)	1.92(73.8)		1.16	1.637
1CC	0.10(41.9)	0.48(26.7)	1.89(31.4)	1.06	0.763
	0.10(36.8)	0.51(24.8)	1.92(38.4)	1.07	0.900
2CC	0.11(40.3)	0.50(25.7)	1.85(34.0)	1.02	0.802
	0.10(37.8)	0.48(26.0)	1.86(36.2)	1.03	0.836
3CC	0.11(45.0)	0.45(27.9)	1.84(27.1)	1.03	0.674
	0.10(37.5)	0.51(24.5)	1.87(38.0)	1.02	0.873

Table S4.2: Fitting parameters for AF660 lifetime data, monitored at 698 nm (λ_{ex} = 600 nm). **1-6** to **3-6** represent AF660 labeled 3arm-DNA conjugated to RCs that have 3arm-DNA to RC ratios between 1 and 3. **1-6-7** to **3-6-7** represent both AF660- and AF750-labeled 3arm-DNA conjugated to RCs that have 3arm-DNA to RC ratios between 1 and 3. The results from two replicates of each sample are shown.

sample	**τ1 ns (amplitude %)**	**τ2 ns (amplitude %)**	**τ3 ns (amplitude %)**	**χ^2**	**average lifetime (ns)**
3arm-DNA-660	1.10(27.4)	1.90(72.6)		1.09	1.681
3arm-DNA-660-750	0.08(25.5)	0.90(19.1)	1.68(55.4)	1.08	1.123
1-6	0.61(39.5)	1.7(60.5)		1.06	1.267
	0.62(32.4)	1.69(67.6)		1.15	1.343
2-6	0.65(38.5)	1.73(61.5)		1.11	1.314
	0.65(35.6)	1.65(64.4)		1.10	1.294
3-6	0.63(37.6)	1.76(62.4)		1.06	1.335
	0.64(37.2)	1.64(62.8)		1.14	1.268
1-6-7	0.08(36.3)	0.58(25.8)	1.7(37.9)	1.07	0.823

	0.09(28.0)	0.68(24.2)	1.65(47.8)	1.04	0.978
2-6-7	0.08(33.6)	0.57(25.8)	1.6(40.6)	1.00	0.823
	0.09(29.1)	0.65(22.4)	1.56(48.5)	1.07	0.928
3-6-7	0.07(39.3)	0.52(27.1)	1.57(33.6)	1.12	0.696
	0.08(36.7)	0.59(23.2)	1.54(40.1)	1.13	0.784

Table S4.3: Fitting parameters for AF750 lifetime data, monitored at 780 nm (λ_{ex} = 740 nm).

sample	**τ1 ns (amplitude %)**	**τ2 ns (amplitude %)**	**τ3 ns (amplitude %)**	**χ^2**	**average lifetime (ns)**
3arm-DNA-660-750	0.48(31.1)	0.72(68.9)		1.08	0.645
1-6-7	0.08(25.4)	0.56(59.6)	1.04(15.0)	1.07	0.510
2-6-7	0.07(31.6)	0.46(37.0)	0.81(31.4)	1.15	0.446
3-6-7	0.08(39.0)	0.47(28.8)	0.78(32.2)	1.19	0.418

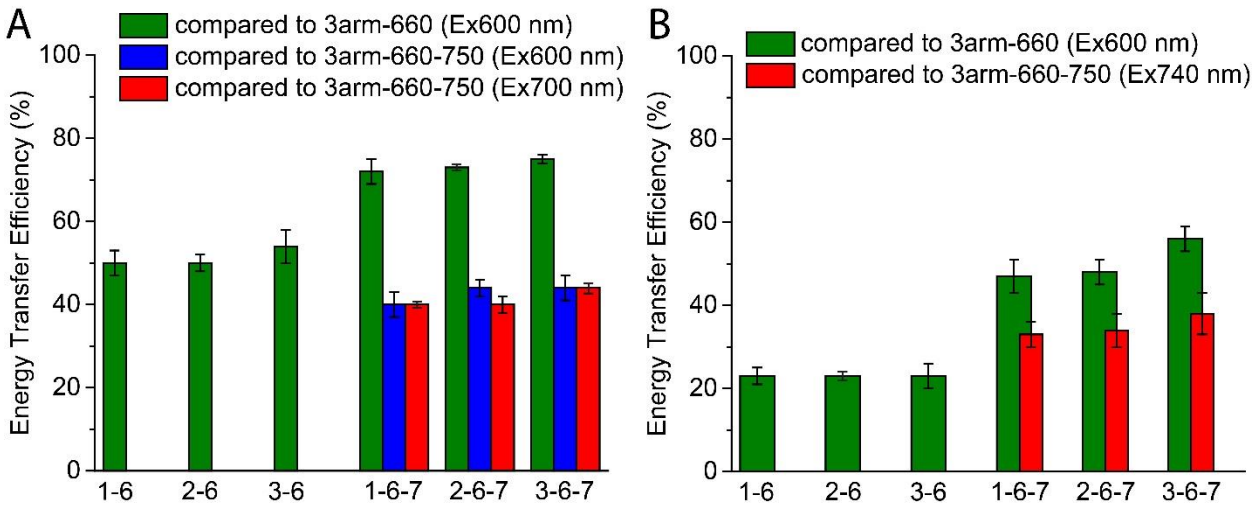


Figure S4.11: Energy transfer efficiencies calculated from (A) steady-state data and (B) lifetime data for **1-6**, **2-6**, **3-6**, **1-6-7**, **2-6-7** and **3-6-7**.

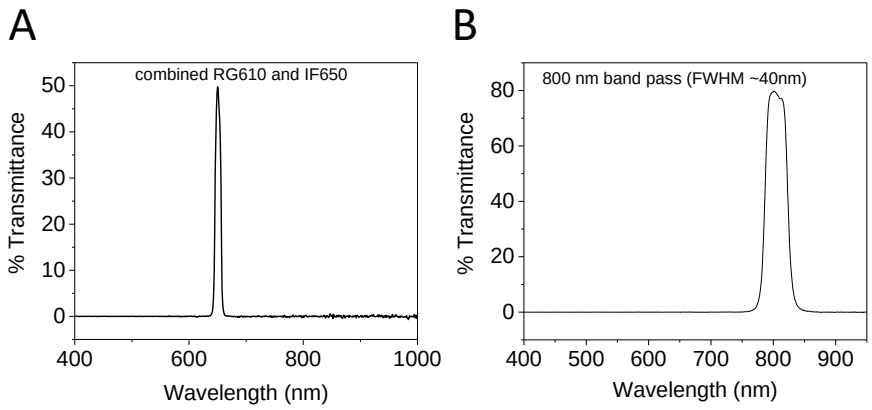


Figure S4.12: Transmittance spectra of filters used in the cytochrome *c* oxidation experiments, (A) for excitation at 650 nm and (B) for excitation at 800 nm.

CHAPTER 5

CONCLUSION AND OUTLOOK

The foundation laid in this body of work investigation of machine learning's utility in the biophysical characterization of aptamer-target binding interfaces can be expanded.

Through the course of my research I've explored the structural and functional topology of synthetic DNA. The simple binary rules of base pairing were leveraged to build nanostructures that can function as a light harvesting system that mimics the photosynthetic membranes from *R. sphaeroides*. SELEX utilizes complex rules of structure-function relationships of nucleotide sequence for molecular recognition to identify high affinity ligands. Poly-ligand profiling of various patient samples can differentiate between complex biological states solely by monitoring sequence and frequency of random synthetic oligonucleotides that bind to the system. I've demonstrated the capabilities of machine learning to move beyond the known rules of DNA to make new structure-function associations between amino acid sequence and aptamer binding.

Neural networks informed on results of screening aptamers against inherently unstructured peptides have displayed potential to predict structurally relevant binding data by showing multiple highly probably non-adjacent interaction sites within a protein's sequence. An interesting path to explore in this field would be to increase the dimensionality of the dataset used to inform the neural network. Rather than building a separate model for each binding signature, informing a single model on multiple binding

signatures and assessing any potential improvements in predictive capabilities. This could also show potential identifying a target from a list of proteins based on the binding signature of an aptamer to random, high density peptide microarrays.

REFERENCES

1. Blind, M. and M. Blank, *Aptamer Selection Technology and Recent Advances.* Mol Ther Nucleic Acids, 2015. **4**: p. e223.

2. Catuogno, S. and C.L. Esposito, *Aptamer Cell-Based Selection: Overview and Advances.* Biomedicines, 2017. **5**(3).

3. Sharma, T.K., J.G. Bruno, and A. Dhiman, *ABCs of DNA aptamer and related assay development.* Biotechnol Adv, 2017. **35**(2): p. 275-301.

4. Domenyuk, V., et al., *Plasma Exosome Profiling of Cancer Patients by a Next Generation Systems Biology Approach.* Sci Rep, 2017. **7**: p. 42741.

5. Hermann, T. and D.J. Patel, *Adaptive recognition by nucleic acid aptamers.* Science, 2000. **287**(5454): p. 820-5.

6. Komarova, N., D. Barkova, and A. Kuznetsov, *Implementation of High-Throughput Sequencing (HTS) in Aptamer Selection Technology.* Int J Mol Sci, 2020. **21**(22).

7. Slatko, B.E., A.F. Gardner, and F.M. Ausubel, *Overview of Next-Generation Sequencing Technologies.* Curr Protoc Mol Biol, 2018. **122**(1): p. e59.

8. Wen, X., Zhong, S., *3D Genome: From Technology to Visualization*. 2018: GitBooks.

9. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi, *Next-Generation Sequencing: From Basic Research to Diagnostics.* Clinical Chemistry, 2009. **55**(4): p. 641-658.

10. Buermans, H.P.J. and J.T. den Dunnen, *Next generation sequencing technology: Advances and applications.* Biochimica Et Biophysica Acta-Molecular Basis of Disease, 2014. **1842**(10): p. 1932-1941.

11. Schutze, T., et al., *Probing the SELEX process with next-generation sequencing.* PLoS One, 2011. **6**(12): p. e29604.

12. Zhou, J. and J. Rossi, *Aptamers as targeted therapeutics: current potential and challenges.* Nat Rev Drug Discov, 2017. **16**(6): p. 440.

13. Mayer, G., *The chemical biology of aptamers.* Angew Chem Int Ed Engl, 2009. **48**(15): p. 2672-89.

14. Dapic, V., et al., *Biophysical and biological properties of quadruplex oligodeoxyribonucleotides.* Nucleic Acids Research, 2003. **31**(8): p. 2097-2107.

15. Vinkenborg, J.L., G. Mayer, and M. Famulok, *Aptamer-Based Affinity Labeling of Proteins.* Angewandte Chemie-International Edition, 2012. **51**(36): p. 9176-9180.

16. Tonapi, S.S., et al., *Translocation of a Cell Surface Spliceosomal Complex Induces Alternative Splicing Events and Lymphoma Cell Necrosis.* Cell Chem Biol, 2019. **26**(5): p. 756-764 e6.

17. Zhang, Y., B.S. Lai, and M. Juhas, *Recent Advances in Aptamer Discovery and Applications.* Molecules, 2019. **24**(5).

18. Campos-Fernandez, E., et al., *Post-SELEX Optimization and Characterization of a Prostate Cancer Cell-Specific Aptamer for Diagnosis.* ACS Omega, 2020. **5**(7): p. 3533-3541.

19. Chen, M., et al., *Development of Cell-SELEX Technology and Its Application in Cancer Diagnosis and Therapy.* Int J Mol Sci, 2016. **17**(12).

20. Domenyuk, V., et al., *Poly-ligand profiling differentiates trastuzumab-treated breast cancer patients according to their outcomes.* Nat Commun, 2018. **9**(1): p. 1219.

21. Opazo, F., et al., *Modular Assembly of Cell-targeting Devices Based on an Uncommon G-quadruplex Aptamer.* Molecular Therapy-Nucleic Acids, 2015. **4**.

22. Keefe, A.D., S. Pai, and A. Ellington, *Aptamers as therapeutics.* Nature Reviews Drug Discovery, 2010. **9**(7): p. 537-550.

23. Boltz, A., et al., *Bi-specific aptamers mediating tumor cell lysis.* J Biol Chem, 2011. **286**(24): p. 21896-905.

24. Taguchi, A.T., et al., *Comprehensive Prediction of Molecular Recognition in a Combinatorial Chemical Space Using Machine Learning.* ACS Comb Sci, 2020. **22**(10): p. 500-508.

25. Zhang, M., et al., *Application of Machine Learning Approaches for Protein-protein Interactions Prediction.* Med Chem, 2017. **13**(6): p. 506-514.

26. You, Z.H., et al., *Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis.* BMC Bioinformatics, 2013. **14 Suppl 8**: p. S10.

27. Szostak, J.W., *Introduction: Combinatorial chemistry.* Chemical Reviews, 1997. **97**(2): p. 347-348.

28. Legutki, J.B., et al., *Scalable high-density peptide arrays for comprehensive health monitoring.* Nat Commun, 2014. **5**: p. 4785.

29. Lupu, L., et al., *Molecular Epitope Determination of Aptamer Complexes of the Multidomain Protein C-Met by Proteolytic Affinity-Mass Spectrometry.* ChemMedChem, 2020. **15**(4): p. 363-369.

30. Nonaka, Y., K. Sode, and K. Ikebukuro, *Screening and Improvement of an Anti-VEGF DNA Aptamer.* Molecules, 2010. **15**(1): p. 215-225.

31. Yoshida, W., et al., *Selection of DNA aptamers against insulin and construction of an aptameric enzyme subunit for insulin sensing.* Biosensors & Bioelectronics, 2009. **24**(5): p. 1116-1120.

32. Huang, R.H., et al., *A Structural Explanation for the Antithrombotic Activity of ARC1172, a DNA Aptamer that Binds von Willebrand Factor Domain A1*. Structure, 2009. **17**(11): p. 1476-1484.

33. Ditzler, M.A., et al., *Broad-spectrum aptamer inhibitors of HIV reverse transcriptase closely mimic natural substrates.* Nucleic Acids Res, 2011. **39**(18): p. 8237-47.

34. Schneider, D.J., et al., *High-affinity ssDNA inhibitors of the reverse transcriptase of type 1 human immunodeficiency virus.* Biochemistry, 1995. **34**(29): p. 9599-610.

35. Diener, J.L., et al., *Inhibition of von Willebrand factor-mediated platelet activation and thrombosis by the anti-von Willebrand factor A1-domain aptamer ARC1779.* J Thromb Haemost, 2009. **7**(7): p. 1155-62.

36. Dolot, R., et al., *Crystal structures of thrombin in complex with chemically modified thrombin DNA aptamers reveal the origins of enhanced affinity.* Nucleic Acids Res, 2018. **46**(9): p. 4819-4830.

37. Bock, L.C., et al., *Selection of Single-Stranded-DNA Molecules That Bind and Inhibit Human Thrombin.* Nature, 1992. **355**(6360): p. 564-566.

38. Cheung, Y.W., et al., *Structural basis for discriminatory recognition of Plasmodium lactate dehydrogenase by a DNA aptamer.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(40): p. 15967-15972.

39. Sutanto, A.R. and D.-K. Kang. *A Novel Diminish Smooth L1 Loss Model with Generative Adversarial Network*. 2021. Cham: Springer International Publishing.

40. Duchi, J., Hazan, E., Singer, Y., *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.* Journal of Machine Learning Research, 2011. **12**: p. 2121-2159.

41. Blankenship, R.E., *Molecular Mechanisms of Photosynthesis, 2nd Edition*. 2nd ed. 2014: Wiley-Blackwell.

42. Mcdermott, G., et al., *Crystal-Structure of an Integral Membrane Light-Harvesting Complex from Photosynthetic Bacteria.* Nature, 1995. **374**(6522): p. 517-521.

43. Scholes, G.D., et al., *Lessons from nature about solar light harvesting.* Nat Chem, 2011. **3**(10): p. 763-74.

44. Wraight, C.A. and R.K. Clayton, *The absolute quantum efficiency of bacteriochlorophyll photooxidation in reaction centres of Rhodopseudomonas spheroides.* Biochim Biophys Acta, 1974. **333**(2): p. 246-60.

45. Boeneman, K., et al., *Quantum Dot DNA Bioconjugates: Attachment Chemistry Strongly Influences the Resulting Composite Architecture.* Acs Nano, 2010. **4**(12): p. 7253-7266.

46. Boeneman, K., et al., *Self-Assembled Quantum Dot-Sensitized Multivalent DNA Photonic Wires.* Journal of the American Chemical Society, 2010. **132**(51): p. 18177-18190.

47. Choi, M.S., et al., *Bioinspired molecular design of light-harvesting multiporphyrin arrays.* Angewandte Chemie-International Edition, 2004. **43**(2): p. 150-158.

48. Hannestad, J.K., P. Sandin, and B. Albinsson, *Self-Assembled DNA Photonic Wire for Long-Range Energy Transfer.* Journal of the American Chemical Society, 2008. **130**(47): p. 15889-15895.

49. Jullien, L., et al., *Multichromophoric cyclodextrins .4. Light conversion by antenna effect.* Journal of the American Chemical Society, 1996. **118**(23): p. 5432-5442.

50. Nakamura, Y., N. Aratani, and A. Osuka, *Cyclic porphyrin arrays as artificial photosynthetic antenna: Synthesis and excitation energy transfer.* Chemical Society Reviews, 2007. **36**(6): p. 831-845.

51. Sautter, A., et al., *Ultrafast energy-electron transfer cascade in a multichromophoric light-harvesting molecular square.* Journal of the American Chemical Society, 2005. **127**(18): p. 6719-6729.

52. Schenning, A.P.H.J., et al., *Porphyrin wheels.* Journal of the American Chemical Society, 1996. **118**(36): p. 8549-8552.

53. Terazono, Y., et al., *Multiantenna Artificial Photosynthetic Reaction Center Complex.* Journal of Physical Chemistry B, 2009. **113**(20): p. 7147-7155.

54. Wurthner, F. and A. Sautter, *Energy transfer in multichromophoric self-assembled molecular squares.* Organic & Biomolecular Chemistry, 2003. **1**(2): p. 240-243.

55. Endo, M., M. Fujitsuka, and T. Majima, *Porphyrin light-harvesting arrays constructed in the recombinant tobacco mosaic virus scaffold.* Chemistry-a European Journal, 2007. **13**(31): p. 8660-8666.

56. Ma, Y.Z., et al., *Energy transfer dynamics in light-harvesting assemblies templated by the tobacco mosaic virus coat protein.* Journal of Physical Chemistry B, 2008. **112**(22): p. 6887-6892.

57. Miller, R.A., A.D. Presley, and M.B. Francis, *Self-assembling light-harvesting systems from synthetically modified tobacco mosaic virus coat proteins.* Journal of the American Chemical Society, 2007. **129**(11): p. 3104-3109.

58. Miller, R.A., et al., *Impact of Assembly State on the Defect Tolerance of TMV-Based Light Harvesting Arrays.* Journal of the American Chemical Society, 2010. **132**(17): p. 6068-6074.

59. Nam, Y.S., et al., *Virus-Templated Assembly of Porphyrins into Light-Harvesting Nanoantennae.* Journal of the American Chemical Society, 2010. **132**(5): p. 1462-+.

60. Scolaro, L.M., et al., *Supramolecular binding of cationic porphyrins on a filamentous bacteriophage template: Toward a noncovalent antenna system.* Journal of the American Chemical Society, 2006. **128**(23): p. 7446-7447.

61. Adronov, A. and J.M.J. Frechet, *Light-harvesting dendrimers.* Chemical Communications, 2000(18): p. 1701-1710.

62. Balzani, V., et al., *Light-harvesting dendrimers.* Current Opinion in Chemical Biology, 2003. **7**(6): p. 657-665.

63. Choi, M.S., et al., *Dendritic multiporphyrin arrays as light-harvesting antennae: Effects of generation number and morphology on intramolecular energy transfer.* Chemistry-a European Journal, 2002. **8**(12): p. 2668-2678.

64. Cotlet, M., et al., *Probing intramolecular Forster resonance energy transfer in a naphthaleneimide-peryleneimide-terrylenediimide-based dendrimer by ensemble and single-molecule fluorescence spectroscopy.* Journal of the American Chemical Society, 2005. **127**(27): p. 9760-9768.

65. Hahn, U., et al., *Light-harvesting dendrimers: Efficient intra- and intermolecular energy-transfer processes in a species containing 65 chromophoric groups of four different types.* Angewandte Chemie-International Edition, 2002. **41**(19): p. 3595-3598.

66. Imahori, H., *Giant multiporphyrin arrays as artificial light-harvesting antennas.* Journal of Physical Chemistry B, 2004. **108**(20): p. 6130-6143.

67. Schenning, A.P.H.J., E. Peeters, and E.W. Meijer, *Energy transfer in supramolecular assemblies of oligo(p-phenylene vinylene)s terminated poly(propylene imine) dendrimers.* Journal of the American Chemical Society, 2000. **122**(18): p. 4489-4495.

68. Weil, T., E. Reuther, and K. Mullen, *Shape-persistent, fluorescent polyphenylene dyads and a triad for efficient vectorial transduction of excitation energy.* Angewandte Chemie-International Edition, 2002. **41**(11): p. 1900-+.

69. Aratani, N., D. Kim, and A. Osuka, *Discrete Cyclic Porphyrin Arrays as Artificial Light-Harvesting Antenna.* Accounts of Chemical Research, 2009. **42**(12): p. 1922-1934.

70. Guldi, D.M., *Fullerene-porphyrin architectures; photosynthetic antenna and reaction center models.* Chemical Society Reviews, 2002. **31**(1): p. 22-36.

71. Guldi, D.M., *Molecular porphyrin-fullerene architectures.* Pure and Applied Chemistry, 2003. **75**(8): p. 1069-1075.

72. Harriman, A. and J.P. Sauvage, *Strategy for constructing photosynthetic models: Porphyrin-containing modules assembled around transition metals.* Chemical Society Reviews, 1996. **25**(1): p. 41-&.

73. Eisele, D.M., et al., *Utilizing redox-chemistry to elucidate the nature of exciton transitions in supramolecular dye nanotubes.* Nature Chemistry, 2012. **4**(8): p. 655-662.

74. Aldaye, F.A., A.L. Palmer, and H.F. Sleiman, *Assembling materials with DNA as the guide.* Science, 2008. **321**(5897): p. 1795-1799.

75. Deng, Z.T., et al., *Robust DNA-Functionalized Core/Shell Quantum Dots with Fluorescent Emission Spanning from UV-vis to Near-IR and Compatible with DNA-Directed Self-Assembly.* Journal of the American Chemical Society, 2012. **134**(42): p. 17424-17427.

76. Hung, A.M., et al., *Large-area spatially ordered arrays of gold nanoparticles directed by lithographically confined DNA origami.* Nature Nanotechnology, 2010. **5**(2): p. 121-126.

77. Lin, C., et al., *DNA tile based self-assembly: building complex nanoarchitectures.* Chemphyschem, 2006. **7**(8): p. 1641-7.

78. Liu, H.J., et al., *DNA-Templated Covalent Coupling of G4 PAMAM Dendrimers.* Journal of the American Chemical Society, 2010. **132**(51): p. 18054-18056.

79. Maune, H.T., et al., *Self-assembly of carbon nanotubes into two-dimensional geometries using DNA origami templates.* Nature Nanotechnology, 2010. **5**(1): p. 61-66.

80. Rinker, S., et al., *Self-assembled DNA nanostructures for distance-dependent multivalent ligand-protein binding.* Nat Nanotechnol, 2008. **3**(7): p. 418-22.

81. Rothemund, P.W.K., *Folding DNA to create nanoscale shapes and patterns.* Nature, 2006. **440**(7082): p. 297-302.

82. Seeman, N.C., *Nucleic acid junctions and lattices.* J Theor Biol, 1982. **99**(2): p. 237-47.

83. Seeman, N.C., *An overview of structural DNA nanotechnology.* Mol Biotechnol, 2007. **37**(3): p. 246-57.

84. Seeman, N.C., *Structural DNA nanotechnology: growing along with Nano Letters.* Nano Lett, 2010. **10**(6): p. 1971-8.

85. Sharma, J., et al., *Control of Self-Assembly of DNA Tubules Through Integration of Gold Nanoparticles.* Science, 2009. **323**(5910): p. 112-116.

86. Sharma, J., et al., *DNA-tile-directed self-assembly of quantum dots into two-dimensional nanopatterns.* Angewandte Chemie-International Edition, 2008. **47**(28): p. 5157-5159.

87. Stephanopoulos, N., et al., *Immobilization and One-Dimensional Arrangement of Virus Capsids with Nanoscale Precision Using DNA Origami.* Nano Letters, 2010. **10**(7): p. 2714-2720.

88. Voigt, N.V., et al., *Single-molecule chemical reactions on DNA origami.* Nature Nanotechnology, 2010. **5**(3): p. 200-203.

89. Winfree, E., et al., *Design and self-assembly of two-dimensional DNA crystals.* Nature, 1998. **394**(6693): p. 539-544.

90. Yan, H., et al., *DNA-templated self-assembly of protein arrays and highly conductive nanowires.* Science, 2003. **301**(5641): p. 1882-4.

91. Yan, H., et al., *A robust DNA mechanical device controlled by hybridization topology.* Nature, 2002. **415**(6867): p. 62-65.

92. Zheng, J.P., et al., *From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal.* Nature, 2009. **461**(7260): p. 74-77.

93. Zheng, J.W., et al., *Two-dimensional nanoparticle arrays show the organizational power of robust DNA motifs.* Nano Letters, 2006. **6**(7): p. 1502-1504.

94. Albinsson, B., J.K. Hannestad, and K. Borjesson, *Functionalized DNA nanostructures for light harvesting and charge separation.* Coordination Chemistry Reviews, 2012. **256**(21-22): p. 2399-2413.

95. Borjesson, K., et al., *Membrane-Anchored DNA Assembly for Energy and Electron Transfer.* Journal of the American Chemical Society, 2009. **131**(8): p. 2831-2839.

96. Dutta, P.K., et al., *DNA-Directed Artificial Light-Harvesting Antenna.* Journal of the American Chemical Society, 2011. **133**(31): p. 11985-11993.

97. Garo, F. and R. Haner, *A DNA-Based Light-Harvesting Antenna.* Angewandte Chemie-International Edition, 2012. **51**(4): p. 916-919.

98. Kumar, C.V. and M.R. Duff, *DNA-Based Supramolecular Artificial Light Harvesting Complexes.* Journal of the American Chemical Society, 2009. **131**(44): p. 16024-+.

99. Probst, M., S.M. Langenegger, and R. Haner, *A modular LHC built on the DNA three-way junction.* Chemical Communications, 2014. **50**(2): p. 159-161.

100. Stein, I.H., C. Steinhauer, and P. Tinnefeld, *Single-Molecule Four-Color FRET Visualizes Energy-Transfer Paths on DNA Origami.* Journal of the American Chemical Society, 2011. **133**(12): p. 4193-4195.

101. Tong, A.K., et al., *Triple fluorescence energy transfer in covalently trichromophore-labeled DNA.* Journal of the American Chemical Society, 2001. **123**(51): p. 12923-12924.

102. Woller, J.G., K. Borjesson, and B. Albinsson, *Self Assembled Porphyrin-DNA Antenna Complex.* Biophysical Journal, 2011. **100**(3): p. 137-137.

103. Woller, J.G., J.K. Hannestad, and B. Albinsson, *Self-Assembled Nanoscale DNA-Porphyrin Complex for Artificial Light Harvesting.* Journal of the American Chemical Society, 2013. **135**(7): p. 2759-2768.

104. Dutta, P.K., et al., *Reengineering the Optical Absorption Cross-Section of Photosynthetic Reaction Centers.* Journal of the American Chemical Society, 2014. **136**(12): p. 4599-4604.

105. Milano, F., et al., *Enhancing the Light Harvesting Capability of a Photosynthetic Reaction Center by a Tailored Molecular Fluorophore.* Angewandte Chemie-International Edition, 2012. **51**(44): p. 11019-11023.

106. Allen, J.P., et al., *Structure of the reaction center from Rhodobacter sphaeroides R-26: the protein subunits.* Proc Natl Acad Sci U S A, 1987. **84**(17): p. 6162-6.

107. Kirmaier, C. and D. Holten, *Primary photochemistry of reaction centers from the photosynthetic purple bacteria.* Photosynth Res, 1987. **13**(3): p. 225-60.

108. Kirmaier, C., et al., *The Nature and Dynamics of the Charge-Separated Intermediate in Reaction Centers in Which Bacteriochlorophyll Replaces the Photoactive Bacteriopheophytin .1. Spectral Characterization of the Transient State.* Journal of Physical Chemistry, 1995. **99**(21): p. 8903-8909.

109. Parson, W.W. and A. Warshel, *Mechanism of Charge Separation in Purple Bacterial Reaction Centers*, in *The Purple Phototrophic Bacteria*, C.N. Hunter, et al., Editors. 2009, Springer Netherlands: Dordrecht. p. 355-377.

110. Woodbury, N.W. and J.P. Allen, *The Pathway, Kinetics and Thermodynamics of Electron Transfer in Wild Type and Mutant Reaction Centers of Purple Nonsulfur Bacteria*, in *Anoxygenic Photosynthetic Bacteria*, R.E. Blankenship, M.T. Madigan, and C.E. Bauer, Editors. 1995, Springer Netherlands: Dordrecht. p. 527-557.

111. Zinth, W. and J. Wachtveitl, *The first picoseconds in bacterial photosynthesis - Ultrafast electron transfer for the efficient conversion of light energy.* Chemphyschem, 2005. **6**(5): p. 871-880.

112. Mahmoudzadeh, A., et al., *Photocurrent generation by direct electron transfer using photosynthetic reaction centres.* Smart Materials & Structures, 2011. **20**(9).

113. Okamura, M.Y. and G. Feher, *Proton-Transfer in Reaction Centers from Photosynthetic Bacteria.* Annual Review of Biochemistry, 1992. **61**: p. 861-896.

114. Spinazzi, M., et al., *Assessment of mitochondrial respiratory chain enzymatic activities on tissues and cultured cells.* Nature Protocols, 2012. **7**(6): p. 1235-1246.

115. Gerencser, L., G. Laczko, and P. Maroti, *Unbinding of oxidized cytochrome c from photosynthetic reaction center of Rhodobacter sphaeroides is the bottleneck of fast turnover.* Biochemistry, 1999. **38**(51): p. 16866-16875.

116. Goldsmith, J.O. and S.G. Boxer, *Rapid isolation of bacterial photosynthetic reaction centers with an engineered poly-histidine tag.* Biochimica Et Biophysica Acta-Bioenergetics, 1996. **1276**(3): p. 171-175.

117. Salafsky, J., J.T. Groves, and S.G. Boxer, *Architecture and function of membrane proteins in planar supported bilayers: A study with photosynthetic reaction centers.* Biochemistry, 1996. **35**(47): p. 14773-14781.