SIMULATION ON THE EFFICIENCY OF GENE-CAPTURE PROBES

By

JINGQIAN LIU

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Biomedical Engineering

Written under the direction of

Biju Parekkadan

And approved by

New Brunswick, New Jersey

May, 2019

ABSTRACT OF THE THESIS

SIMULATION ON THE EFFICIENCY OF GENE-CAPTURE PROBES

By

JINGQIAN LIU

Thesis Director:

Biju Parekkadan

The multiplexed cloning of long DNA sequences can be of great value for biotechnology applications such as long-read genome sequencing or the creation of libraries of open reading frames (ORFs) from genomes for expression screening. Long adapter single-stranded oligonucleotide (LASSO) probes have shown promise as a tool to enable the capture and cloning of long DNA fragments by engineering an adapter region of user-defined length flanked by PCR primer regions, ultimately creating a long probe to bind long DNA regions. The success of LASSO in cloning target ORFs relies on the design of its long adapter sequence to be non-specific to a genomic region while also providing an optimal length for flexibility of a stable probe-target complex for downstream processing. The reason behind the adapter-length depend capture enrichment was explored in this thesis. Herein, we proposed two hypotheses, which are respectively related to the probes' secondary structures as well as probe-target interaction free energy. Mfold, a tool for secondary structure prediction and Monte Carlo simulation based on a coarse-grained DNA model were applied to verify the two hypotheses. According to the computational predictions, the latter explanation, which was associated with the probe-target interaction free energy, was considered to account for the capture enrichment efficiency. Our results suggest that the length of the adapter is a factor that contributes to the free energy of target-probe interaction, thereby determines the efficiency of capture. The results also reveal different preferences to various adapter lengths when the target shifts from 400bp to 1500bp. For long target genes, longer adapters are more favorable, as simulations and experiments show.

ACKNOWLEDGEMENTS

I would like to convey my gratefulness to my advisor Dr. Biju Parekkadan for providing me the opportunity to take this project, in which I learned a lot and finally confirmed the area that I'm going to pursue for my further academic career. I cannot be more thankful for the patience and support he provided all through this time. And equal thanks to my co-advisors Dr. Wilma Olson and Dr. Vikas Nanda, who offered great amount of helps to this project. I'd like to specifically thank Dr. Olson for raising the initial solution for this problem which illustrated the direction of this research and also Dr. Vikas for giving helpful suggestions which made the project went smoothly.

I also like to express my thanks to Dr. Li, Shuxiang for his guidance on using the molecular simulation software, learning the basic principles of simulation and dealing with the technical problems happening during this process. And thanks to Syukri Shukor and Dr. Viswanadham Sridhara for kindly providing and explaining the experimental data which built the experimental foundation for this study.

What's more, thanks to my colleagues Shilpaa Mukundan, Alexandra Burr and Lauren Timmins for their useful suggestions and comments after every presentation I did in group meetings.

Last but not the least, I would like to express the deepest appreciation to my parents who not only provide me financial support, but also give me constant encouragement from time to time.

Table of Content

LIST OF ILLUSTRATIONS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF ABBREVIATIONS

PCR	Polymerase Chain Reaction
ORF	Open Read Frame
LASSO	Long Adapter single strand oligonucleotide
MIP	Molecular Inversion probe
bp	Base Pair
nt	Nucleotide
MC	Monte Carlo
VMMC	Visual Movement Monte Carlo
PAGE	Polyacrylamide Gel Electrophoresis
LA	Long Adapter
AL	Adapter length
TI	Thermodynamic Integration
FEP	Free Energy Perturbation
WHAM	Weighted Histogram Analysis Method
RPKM	Reads per Kilobase of transcript, per Million mapped reads

CHAPTER I INTRODUCTION

1.1 DNA Read-write gap

1.1.1 DNA sequencing

The order of nucleic acids in the double strand helix of DNA contains all the hereditary information of living organisms. Deciphering the sequence, therefore, has always been one of the primary goals in understanding how genetics influences organism function. The first milestone of DNA sequencing technology was famously established by Frederick Sanger in 1977, by using "chain-terminating" method to determine the sequence of DNA of bacteriophage ϕX174. (1) The technique was highlighted by the high accuracy and its read length which could reach 1000bp. Nevertheless, its high cost and low throughput have seriously limited its broad application in genome sequencing. (2) To pursue improvements, the second-generation of DNA sequencing then sprung up as a result of the effort that was led by several companies including Illumina, Roche and ABI. (3) The technique was based on massive parallel analysis and alignment of short fragment reads of a genome sequence and achieved the high throughput which allows the simultaneous sequencing of millions of DNA molecules. The capability of reading long pieces of DNA has been challenging for the field, though is of great interest, to minimize errors in aligning short reads that may have redundant sequences as well as reducing the sequencing depth (thus, processing time and cost) needs to fully catalogue a genome. The third generation of sequencing which mainly dependent on single molecule sequencing without DNA amplification has largely resolved the problem of reading long DNA chains. (2) With improved performance in read length and acceptable error rate, these single molecule approaches may become the mainstream of DNA sequencing in the future. With considerable technology improvements, there has been a rapidly dropping rate of DNA sequencing cost per megabase, which owes to the fast development of DNA sequencing equipment and protocols. In 2004 after 14 years the Human Genome Project was launched, the project was declared complete. Now, the speed of decoding natural DNA sequencing remains 15 petabases per year, (4) which is even facilitated by the development of the third generation of sequencing. These improvements have generated great excitement about understanding genetics more rapidly, though it has created a huge disparity between the amount of genetic content and our ability to functionally test the biomolecules that a genome encodes for.

1.1.2 DNA synthesis

If we consider DNA sequencing as "reading" of the genomic information, DNA synthesis could be described as "writing". Generally, the construction of DNA could either be accomplished by *de novo* synthesis with a known sequence or based on reproduction of a template DNA sequence.

De novo DNA construction starts from the synthesis of oligos (short DNA molecules) and ends with sealing these fragments. The most common chemical oligo synthesis method utilized today is still based on the Phosphoramidite-based method developed by Marvin Caruthers in the early time of the 1980s. The method basically includes a four-step cycle in which nucleotide was added one at a time. (5) After entering 1990s, spatially localized parallel chemical synthesis on a surface (6) was brought to people's version, contributing to the introduction of array-based oligo syntheses. With this development, nowadays, the cost synthesis of DNA oligos has largely decreased to $0.00001–0.001 per nucleotide (varying among different platforms). (7) Nevertheless, this ideal efficiency only happens to the construction of short oligos. It turns out that a large-scale DNA de novo synthesis still has technical difficulties that occur when splicing these raw substrates to long DNA pieces ranging over hundreds of bases. High costs are always caused by the needs of cloning and frequent sequencing verification which due to the high error rate of this process. (7) Therefore, a cost-efficient *de novo* DNA synthesis technique for long targets still demands to be demonstrated. **(Figure 1.1.2)**

Similar to the current status of *de novo* DNA construction, attempts have also been made to address the issue of template-based DNA synthesis by cloning of specific genome regions. For a single template, PCR is a sufficient method to clone a DNA region. Ideally, there is interest to copy hundreds to thousands of DNA regions in a single reaction, a method known as multiplex cloning, to enable the rapid creation of large libraries for expression testing. Traditional PCR is not effective for multiplex cloning because of the interaction between the primers that always prevent them from combining with targets. Short selector oligonucleotides could be applied for amplification of DNA pieces in hundreds-base length. (8) (9) Molecular inversion probes (MIPs; structure and applications will be described later in this thesis) was generated to prevent the problem of primers' interaction and to improve the specific recognition to target DNA fragments. (10) MIPs have shown a favorable performance in capturing and amplifying target sequences ranging from 0-200bp, becoming the commercial reagent for single nucleotide polymorphism (SNP) analysis of genomes. Long-target capture and amplification, however, have not been established by MIPs for its utility in multiplex cloning of biomolecule-encoding regions that span >200bp in length.

In summary, although the protocols for constructing short DNA fragments by either de novo synthesis or cloning have been well established, a reliable technique for long target construction still demands to be further developed.

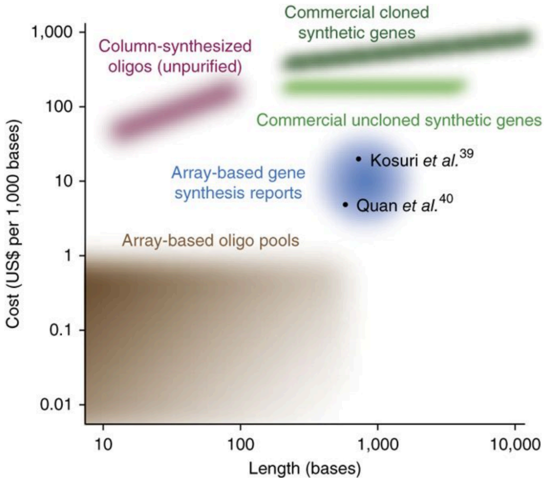

Figure 1.1.2 Cost of different de novo synthesis methods (Taken from (7))

1.1.3 Gap between DNA sequencing and DNA synthesis

As it has been demonstrated before, while our capability of reading DNA sequences was significantly improved, the techniques of “writing” these sequences are facing with a bottleneck. Owing to improved techniques on decoding genome sequence (11), plenty of information about DNA sequence has flooded in. Nevertheless, the establishment of a Central Dogma-based streamline which may transform all the “sequence” information to protein-level observation still needs a long way to go. The realization of this goal may rely on a reliable, efficient approach of DNA synthesis. It is presented in **Figure 1.1.3** that DNA synthesis is far more expensive than DNA sequencing. The dramatic asymmetry between flooded sequencing information and poor interpretation about their function has been caused by this lack of cost-efficient technique for the synthesis of long DNA fragments that include the protein-encoding genes. This asymmetry is what is usually called the “read-write” gap.

To fill the “read-write” gap is of great significance for it will provide a credible way to transform the sequence information into functional information. Although theoretical methods such as machine learning were utilized to realize this transformation (12-14), a reliable and detailed conclusion undoubtedly relies on an experimental study which could only be accessible with DNA synthesis. Therefore, requisition for an efficient and widely-applicable DNA synthesis method has become imminence for biological research. What's more, as we mentioned before, a pipeline of producing protein with corresponding ORF loaded on DNA fragment is another attractive benefit of developing new tools of DNA synthesis.

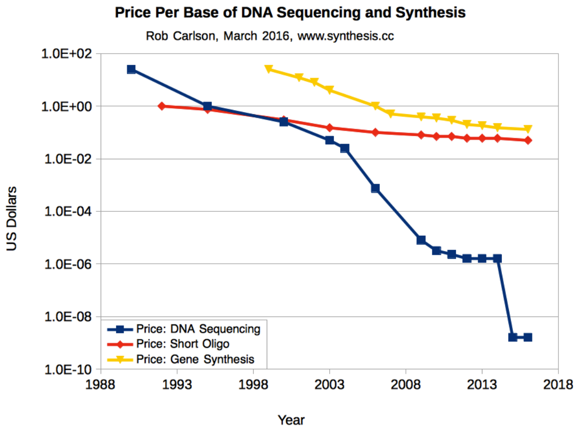

Figure 1.1.3 Contrast between cost of DNA sequencing and DNA synthesis (Taken from Rob Carlson, March 2016, www.synthesis.cc)

1.2 Long adapter single-stranded oligonucleotide probe (LASSO)

1.2.1 Molecular inversion probe

Long-adapter single-stranded oligonucleotide (LASSO) probe is a recently-developed tool used to realize the capture and cloning of long target DNA pieces. The application of this technique allows an efficient synthesis of ORF regions, which is always very long, thereby providing a promising way to transform the sequence information into functional information. To describe the structure of LASSO, we'd like to introduce its precursor, molecular inversion probes (MIP) which was addressed in the previous context. A MIP is a target-gene capture probe used in multiplex DNA cloning. Compared with traditional PCR primers, its design could prevent the probes from combining with each other and enhance its recognition to targets. It is a linear single oligonucleotide consisting of a linker sequence between two primer sequences which are complementary to a target DNA fragment. The linker sequence enables MIPs to form a padlock shape **(Figure 1.2.1)** in order to specifically hybridizing to flank target genomic region of interest. (10) In its application in gene typing , MIPs exhibited high specificity and efficiency in multiplex short DNA target (<200bp) enrichment. (15)

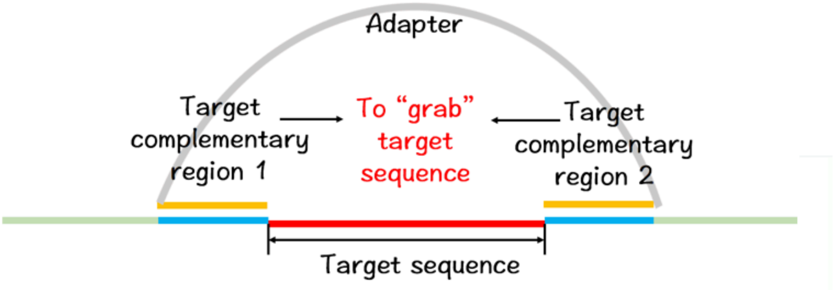
Adapter
Target complementary region 1
To "grab" target sequence
Target complementary region 2
Target sequence

Figure 1.2.1 Structure of MIP and LASSO. There are two complementary regions to recognize the target gene and an adapter to link the two regions together.

1.2.2 Structure and application procedures of LASSO

As the cloning capacity of MIPs diminishes in target sequences longer than 200bp, efforts were invested to improve the performance of MIPs in capturing and cloning long DNA targets. Notably, increasing probe linker sequence length enabled the capture of target DNA up to 400bp. (16, 17) According to this improvement made after extending the linker, our group developed a modified version of padlock probes, LASSO Probe (18), which addressed the limitations of traditional MIPs and PCR probes with further extended linker sequences.

Similar to MIP, LASSO was an oligonucleotide fragment integrated by two sticky arms (One is called extension arm while the other one is called ligation arm.) that are complementary to the two ends of target genome region and one extended linker (aka adapter) whose length ranges from 242 nucleotide (nt) to 788nt (**Figure 1.2.1**). To be more specific, three types of adapters were experimentally tested for constructing LASSO, including 242-nt, 442-nt, and 788-nt adapters which were gained from plasmid pCDH-CMV-MCS-EF1-Puro. (18)

The application of LASSO probes in capturing target genome fragments could be generally divided into several steps. **(Figure 1.2.2)** First, LASSO probes designed for multiple targets are injected into a pool with fragmented genomic DNA inside. With the two complementary regions, each probe recognizes its target gene and forms a local padlock shape. After that, the polymerase will join in the reaction and fill the gap from the extension arm to the ligation arm, which allows the formation of a complete circle with the target sequence included. Then, this linear DNA chain will be digested by an exonuclease so that the following amplification is based on target genes. Finally, the circles are converted to linear pieces and the target-probe complexes get amplified. In the final step, interactions between primers are negligible since the design of LASSO allow these primer-binding sites to be identical with each other.

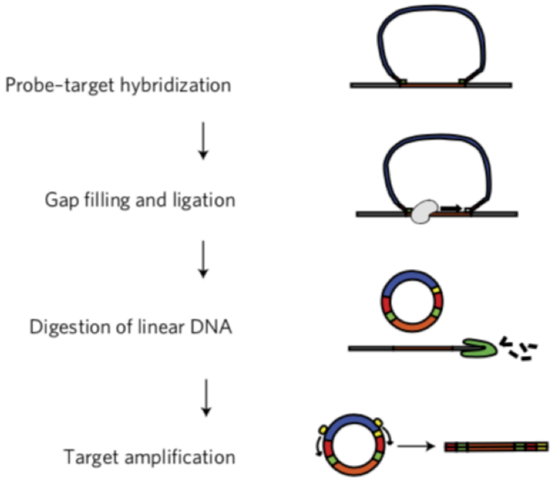

Figure 1.2.2 Procedure of using LASSO (Taken from 16)

1.2.3 Construction of LASSO

Two predecessors **(Figure 1.2.3b)** need to be prepared before the construction of LASSO (**Figure 1.2.3a**), including the pre-LASSO and the long adapter. Pre-LASSO could be either dsDNA oligos or ssDNA gained through oligo synthesis. It is combined by five parts, namely ligation arm, extension arm, two primer annealing sites that are going to function in the following producing procedures and one piece of sequence used for fusing the two predecessors. The long adapter is obtained from a plasmid template. It consists of one primer annealing site, the conserved linker region as well as a corresponding complementary sequence for the fusion site.

After the preparation of pre-LASSO and the adapter, these two predecessors will be fused together by overlap-extension PCR (**Figure 1.2.3**c) and then go through intramolecular ligation to accomplish cyclization (**Figure 1.2.3d**). Finally, the DNA circle will be transformed into a linear one with the desired sequence through inverted PCR. (**Figure 1.2.3e**)

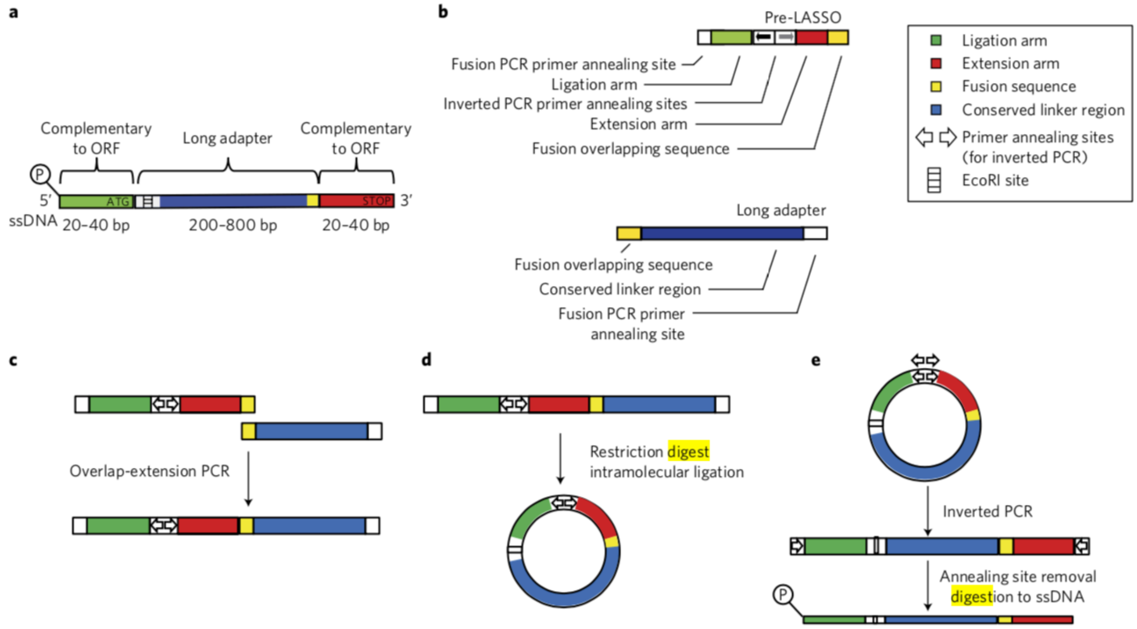
a
b
Pre-LASSO
Fusion PCR primer annealing site
Ligation arm
Inverted PCR primer annealing sites
Extension arm
Fusion overlapping sequence
🟩 Ligation arm
🟥 Extension arm
🟨 Fusion sequence
🟦 Conserved linker region
⇦ ⇨ Primer annealing sites (for inverted PCR)
EcoRI site
Complementary to ORF
Long adapter
Complementary to ORF
P
5′
ATG
STOP
3′
ssDNA
20–40 bp
200–800 bp
20–40 bp
Long adapter
Fusion overlapping sequence
Conserved linker region
Fusion PCR primer annealing site
c
Overlap-extension PCR
d
Restriction digest intramolecular ligation
e
Inverted PCR
Annealing site removal digestion to ssDNA
P

Figure 1.2.3 Synthesis of LASSO probes. a, Final LASSO probe. b, Two predecessors for LASSO synthesis, including a pre-LASSO probe and a long adapter. c, PCR reaction to fuse the adapter and the pre-LASSO probe. Now the complex is in arm-arm-adapter structure. d, Circular fusion

PCR product generated by intramolecular circularization reaction, which step is critical for the complex to rearrange itself into arm-adapter-arm sequence in the next step. e, Inverted PCR to digest the circular DNA on the other site to create linear LASSO with ligation and extension arm on the two sites (Figure and legend Taken from (16))

1.2.4 Efficiency discrepancy for different LASSO design

The high efficiency of LASSO in capturing and amplifying target DNA fragments has been verified in experiments. Its coverage of application is also significantly enlarged compared with MIP and other derivatives. Enrichment of targets is shown to be largely improved especially for the DNA target longer than 400bp. Proof-of-concept studies detected highly multiplexed cloning of >1000 prokaryotic open reading frames (ORF, which is always regarded as the encoding gene that corresponds to proteins) ranging from 400 - 4000 bps in a single reaction. (16) These results all suggest LASSO's potential of being a promising tool of long-target DNA synthesis.

Different capture efficiencies that due to the application of different long adaptor, however, was interestingly found in our experiments, which will be discussed in detail in the following chapters. It was shown that LASSOs with different adapter length (AL) demonstrated different performances in target capture. To figure out the implicit mechanism behind this phenomenon, we proposed two hypotheses, intervention of DNA secondary structures and the free energy barrier for probe-target interaction. The two hypotheses will be explored and explained in the next two chapters. The secondary structures were predicted but showed light complexity at experimental temperature. Thus, it is not considered to be the main reason that causes different capture efficiencies. Furthermore, the free energies of probe-target interactions are estimated by a coarse-grained DNA model, oxDNA (19). The computational results suggest the correlation between the probe-target interaction free energy and the efficiency of target capture.

Since the application of LASSO has shown preliminary success and demonstrated preferred performance compared with traditional methods of DNA cloning, it is reasonable to believe that further development of LASSO may bring the solution for the filling of the read-write gap and pave the way of large-scale DNA synthesis. Therefore, this study which aims to uncover the mechanism of the discrepancy of different LASSO is of great significance for it potentially contributes to the optimization of the LASSO and the improvement of the DNA cloning efficiency.

CHAPTER 2 THE ADJUSTMENT TO LASSO ADAPTER

2.1 The efficiency discrepancy caused by different adapter length

A critical characteristic of LASSO that contributes to this highlight performance is the adjustability of the ALs. The AL that was around 50nt in MIPs was extended to 242nt, 442nt and 788nt in three types of design which aims to figure out the most optimized length for targets. As experimental results showed, LASSO presented higher efficiency in capturing long DNA targets. However, longer adapters do not necessarily mean higher efficiency. Discrepancies of capturing efficiency occurred among these three types of probes, which are obviously related to the various ALs. They could be observed in the sequencing results which are used to test the products of simultaneously capturing thousands E. coli ORFs by MIPs and LASSOs with different ALs. Enrichment efficiencies of the probes were presented using both target and off-target data. We found that although the LASSOs with 242nt-long adapter (LA) and 242nt-LA demonstrated better performance than MIP, the capturing efficiency began to largely drop when LA grew to 788nt. (**Figure 2.1-1**)

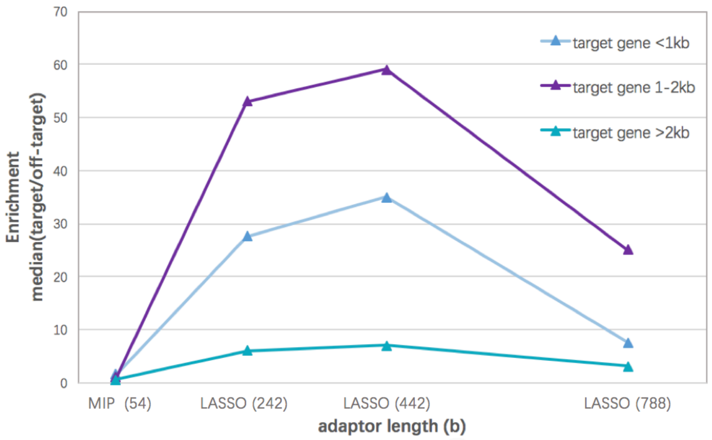

Figure 2.1-1 Sequencing results for the amplification results of MIP and LASSOs (Data offered by Dr. Viswanadham Sridhara)

Figure 2.1-1 has demonstrated an overall performance of MIPs and LASSOs when thousands of target ORFs are involved. After this, we'd like to detect LASSO efficiencies in a more reliable scale. To further verify the different capture efficiency, experiments for single target is conducted to achieve a clearer observation rather than a pooled effect. Products of capture, namely the target-probe complexes, were subject to PAGE. (**Figure 2.1-2A**) PAGE DNA quantification in **Figure 2.1-2B** demonstrated a two-tailed distribution of capturing efficiency for different ALs on each LASSO target. P-values were respectively measured as 0.0068, 2.5×10^{-7}, and 4.1×10^{-6} within the group of 0.6-, 1.0-, and 2.0-kb targets. Furthermore, the preference of AL also changes as target length varies. As the results show, targets 0.6 and 2.0kb in length prefer 442-nt adapters while 1.0kb targets gain more captures with a 242-nt adapter. Furthermore, this is also evidence showing the efficiency differences happen or at least happen before or exactly in the target capture process, which provides the basis for our hypothesis regarding to the performance differences. As far, both library experiment and single target test demonstrate the efficiency differences between different LASSOs with different ALs.

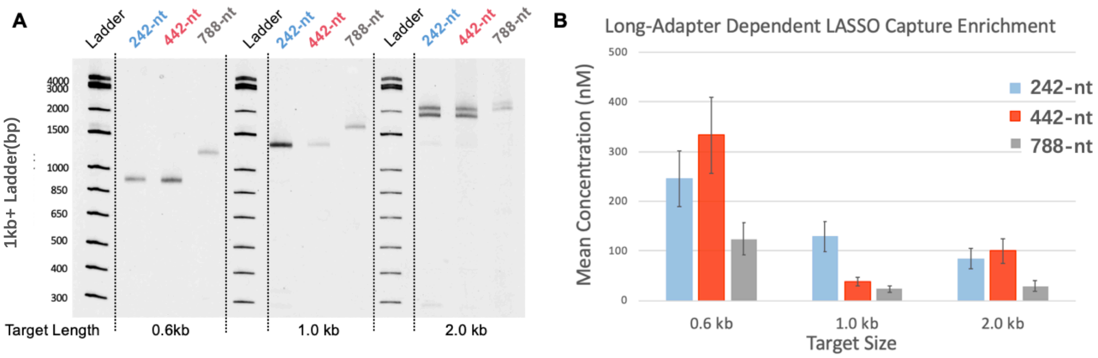

Figure 2.1-2 Capture results for targets of 0.6kb, 1.0kb, 2.0kb. (A) PAGE visualization of amplified capture products for different LA lengths and target sizes (B) Single-target capture quantification from four PAGE experiments, each quantified with standard curves derived from DNA ladder serial dilutions with known band concentrations. (Figure offered by Syukri Shukor)

2.2 Hypotheses regarding to the amplification efficiency difference

Two hypotheses were initially proposed to explain the discrepancies between probes with different AL: 1) secondary structures may form in a long LASSO probe and further prevent the probe from capturing the target. 2) In padlock probe technologies, the interaction between the target and the probe are necessary for further hybridization on the two arms and form a padlock shape. In this process, the free energy barrier needed to be overcome by both probe and target might vary between probes with different ALs. Both hypotheses have been taken into consideration in our computational work and the results will be described and discussed in the following two chapters.

CHAPTER 3 SECONDARY STUCTURE PREDICTION FOR LASSO PROBE

3.1 Principle of DNA secondary structure prediction

Nucleic acid secondary structures are structures formed due to the interaction between bases inside the DNA/RNA pieces, including helixes, stem loops, pseudoknots etc. Their formations always play an important role in the functional effect of nucleic acid. For example, the transferring function of tRNA is supported by its cloverleaf structure (20) and some of the protein-DNA interactions also rely on the secondary structures on the site of DNA (21). In our study, the secondary structures of LASSOs may probably cause the intervention to the binding of target and probe and thus lead to poor efficiency of target capture. And here we'd like to test this hypothesis through prediction to secondary structures of LASSOs.

Nucleic acid secondary structure prediction has been under study for decades of years. The model based on minimizing free energy was raised at the early time of the 1970s by Tinoco et al, which assigned free energy values to different shapes of local structures and aimed to find the most optimized interaction with minimum free energy. The main idea is that paired bases lower the free energy while unpaired bases contribute to higher free energy. (22) Another mainstream method for secondary structure prediction is the comparative sequence analysis method which starts from an alignment of related sequences and estimates the secondary structure through the analysis of the structures of this series of sequences. (23) Generally, the minimum free energy method is usually applied to the predictions towards a single sequence whereas the latter one is preferred when a set of homologous sequences are provided.

3.2 Introduction of Mfold

Considering the reliance, the tolerance to the sequences' length, the capability of predicting various structures and the comparably favorable speed, we selected Mfold (24) as the prediction tool in our study. Mfold is an online web server for nucleic acid secondary structure prediction which is based on the method of minimum free energy and further introduced dynamic programming to accelerate the process. It provides more adaption to complicated motif such as pseudoknots and allows the structure estimation to long sequence over 200nt, which cannot be achieved by some other counterparts.

In its computation, it regards the total free energy as the sum all the free energy of every structural motif where each of them is considered independent. (24) It is able to predict the optimal and the suboptimal structures by adjusting the base pairs formed according to a given or automatically set free energy increment. (25) It is widely applied to secondary structures prediction of DNA and RNA.

3.3 Prediction of LASSO secondary structure under experimental condition

The computation was conducted with the input information representing experimental condition (T=65 ℃ , $[Mg^{2+}] = 10mM$, $[Na^{+}] = 100mM$) **Figure 3.3** shows the secondary structure prediction results for LASSOs for the E.coli gene nemR, which demonstrates extremely complex secondary structures under 37℃ but largely reduced motifs under experimental temperature.

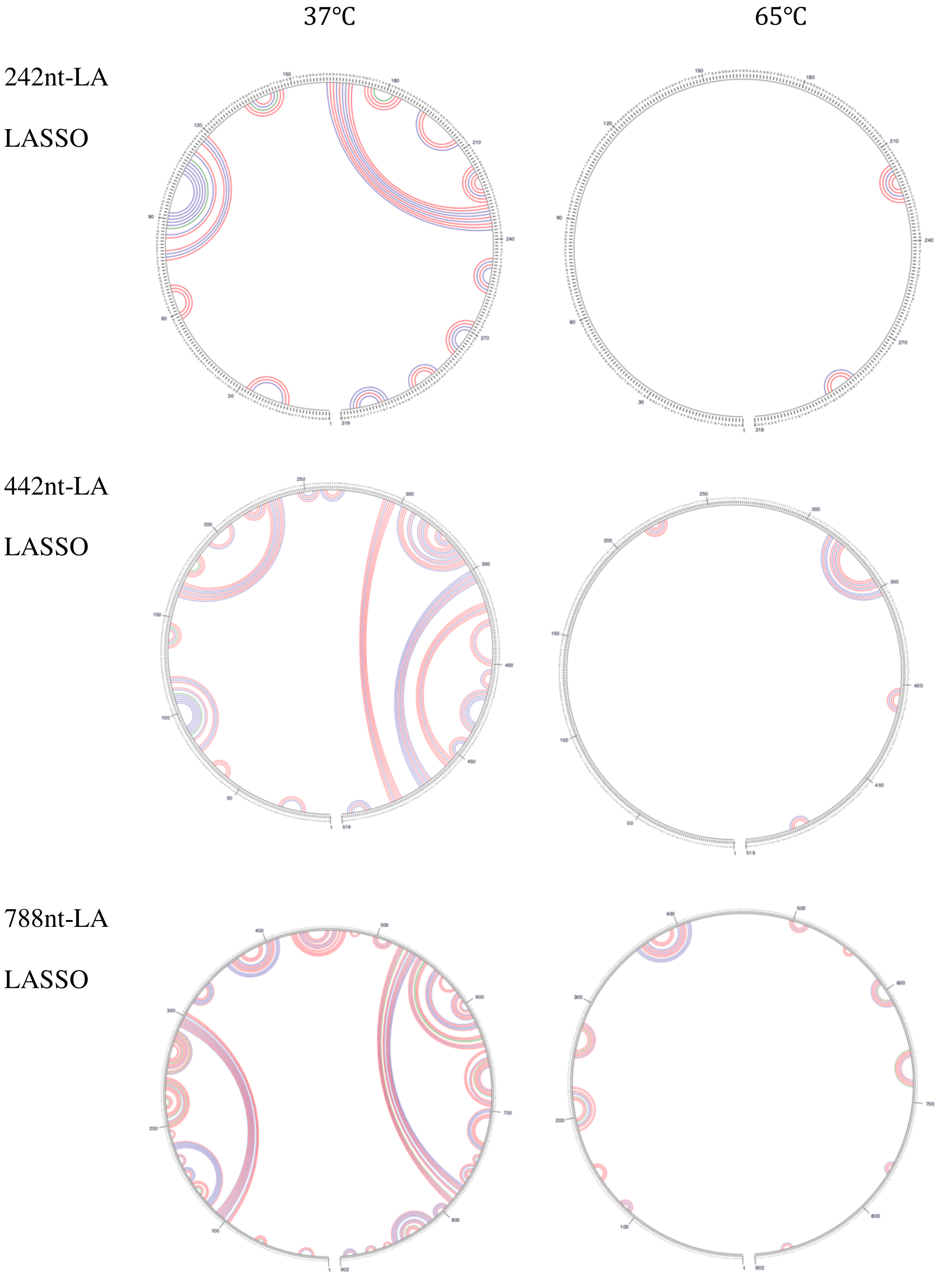

Figure 3.3 Secondary structure predictions for LASSOs of the E.coli gene nemR under 37℃ and 65℃

3.4 Discussion

The contrast between structure prediction at 37℃ and under the experimental conditions of LASSO target capture suggests that the secondary structure, which may intervene in target-probe interaction at the lower temperature, is nearly eliminated by the high temperature of the experiment. Most of the predicted motifs only involve bases within a short range, thereby preventing dramatic deformation of the whole probe. Also, the percentage of secondary structures does not show apparent differences as the AL is changed. Through this observation, we conclude that secondary structure formation is not the main factor that underlies the discrepancy of efficiency.

CHAPTER 4 LOCAL PROBE-TARGET HYBRIDIZATION FREE ENERGY

4.1 Molecular simulation and free energy calculation

Molecular simulation is a computational method that aims to simulate the dynamic behavior or physical/chemical properties of molecules. Two methods are usually applied for molecular simulation, including Molecular Dynamics (MD) simulation and Monte Carlo (MC) simulation. Through the MD method, the trajectory of molecular movement could be obtained by solving Newton's equation of motion. (26) MC simulation is much simplified for it only considers the energy change between the configurations. In one MC step, a proposed configuration will be accepted if the energy change is acceptable without massive computation required for solving equations. (27) Plenty of configurations are generated by Markov chains and used for the analysis of the behavior of the molecules. Metropolis algorithm which provides proposal distribution (28) are always applied for effective sampling.

Free energy is always used to predict if a chemical reaction could happen spontaneously. A negative value of free energy change is necessary for a reaction to be spontaneous while a positive value represents a reaction that needs to overcome free energy barrier. Thermodynamic integration (TI)(29), free energy perturbation (FEP) (30) and umbrella sampling (31) are three main approaches to quantify free energy. In this study, the free energy of probe-target interaction will be quantified by MC simulation associated with umbrella sampling since it is already in conjunct with MC algorithm in oxDNA.

Other kinds of computational works have been done before to simulate the process of traditional PCR, such as calculating final yield using model based on PCR efficiency and amount of circles

(32), visual PCR model which further included the “recognition capability” of primers (33). What’s more, to pursue higher enrichment efficiency, software of designing MIP probes are developed upon learning algorithms (34, 35). Nevertheless, as we know, a study modelling for understanding target DNA capture on a molecule level is still lacking.

4.2 Method of simulation

4.2.1 oxDNA

oxDNA (19) is a software designed for the study of DNA self-assembly. It especially emphasizes the thermodynamics of the transition process of DNA melting and hybridization. Before oxDNA, many coarse-grained models, in which degrees of freedom are much more reduced compared with all-atom models in order to simplify the computation, were established. For example, statistical models without much consideration to the structural and dynamics details were found able to reproduce the data obtained from the experiments with 4-16bp DNA strands. (36) (37) These models, however, are unable to describe the structural and topological details of DNA. Except for statistical models, lattice model which remains more structural information (38) and rigid base-pair models (39) (40) have been developed as alternatives of statistical coarse-grained models. These models, although hold advantages such as containing structural details or sequence specificity, are not capable to describe the transition between single and double strands of DNA, which is the primary issue that oxDNA aimed to address.

There are mainly two methods to develop a coarse-grained model. One of them is to set the parameters of the force field given the results of all-atom simulation or the crystal structures, which is generally called the idea of “bottom-up”. Another way is to design a force field in order to explain the experiments. This is what we call “top-down” method. The latter one was adopted in the development of oxDNA to provide effective interactions. (41)

The interaction potentials inside oxDNA consider excluded volume (a volume in a space is not accessible to other molecules because of the presence of the first molecule (42)), backbone connectivity, hydrogen bonding and base stacking (Van der Waals interaction between bases, known as an important force to stabilize the DNA double helix structure (43)). (41) The potentials and the interaction sites are respectively represented in the equation and **Figure 4.2.1**(41). “nn” represents that all the nucleotides in the DNA strands should be considered in the calculation of the first part of the equation. It is worth mentioning here that all the interactions in oxDNA happen between two nucleotides, which means no three-body interactions are involved. (41)

$$V = \sum_{nn}(V_{backbone} + V_{stack} + V'_{exc}) + \sum_{other\ pairs}(V_{HB} + V_{cross_{stack}} + V_{coaxial_{stack}} + V_{exc})$$

(Taken from (41))

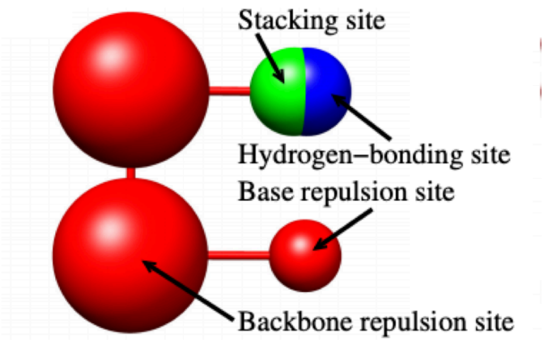

Figure 4.2.1 Interaction sites in oxDNA force field (Taken from (41))

Since oxDNA has been developed, studies on DNA hybridization(44-46) and DNA cyclization(47, 48) have been successfully conducted with oxDNA. Its coarse-grained model enables the simulation for large system under large time scale (49) and thus provides an appropriate tool for our study where two long strands of DNA are included. In our simulations, we adapted reaction coordinates to fit the case of LASSO probes and followed the same procedures used in prior work with oxDNA.

4.2.2 Techniques for efficient sampling

Visual movement Monte-Carlo (VMMC) (50) method was adopted to avoid low rates of acceptance in order to allow more effective movements rather than keeping rejecting movements which lead to low efficiency of simulation. Instead of accepting or rejecting a proposed movement of one particle which is always applied in traditional MC algorithm, VMMC algorithm further forms a "moving cluster" which is composed by the neighbor particles that potentially move with the randomly picked seed. The process of forming a cluster is not a literal proposed move, which is therefore named visual movement. The diagram of the algorithm is shown in **Figure 4.2.2**. First of all, a seed will be randomly chosen from the system. In the next step, a neighbor particle will go through a visual movement and it will be determined to move with the seed if there is a strong interaction between this particle and the seed. Otherwise, it will stay at the original location if the energy change is acceptable. The other particles will be subject to the same procedure afterwards. Finally, a cluster will be generated and its movement will be determined by the algorithm.

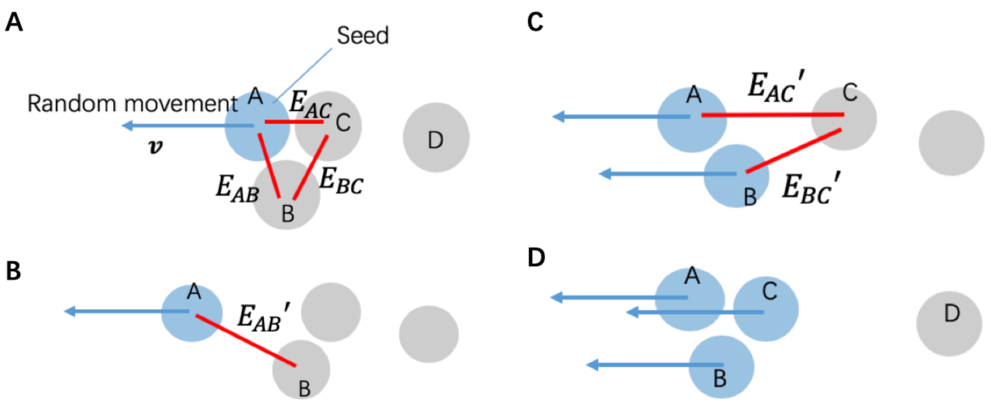

Figure 4.2.2 Diagram of VMMC algorithm. (A) seed A is randomly picked from the system of particles and a random movement is proposed. (B) The algorithm determines whether the neighboring particle B moves with A. (C) If movement of B is accepted, algorithm determines whether C moves with A and B. (D) The final cluster composed of A, B and C is generated and their movement is proposed.

Another technique used to promote an efficient sampling is a biased sampling called umbrella sampling created by Torrie and Valleau in 1977 (31), since the hybrid state of target and probe is actually a rare event that is hard to be sampled under a normal canonical ensemble. The idea of traditional umbrella sampling is to manually add a biasing potential in order to get desirable states samples, which is always accompanied with weighted histogram analysis method (WHAM) (51) to generate a whole free energy profile from several pieces of local free energy geography.

In oxDNA, the conception of umbrella sampling was introduced in a form of assigning biased weights to different states. (41) The unbiased sampling could be obtained from biased sampling and the weights assigned to these states without getting the distribution deformed. This has been described in their report (41). For example, use r^N, Ω^N to represent the coordinate space and vector space of an ensemble and $< A(r^N, \Omega^N) >$ to present any thermodynamic average. We can have eq 4.2.2-1:

$$< A >= \frac{\int dr^N d\Omega^N A(r^N,\Omega^N) e^{-\beta V(r^N,\Omega^N)}}{\int dr^N d\Omega^N e^{-\beta V(r^N,\Omega^N)}}$$ eq 4.2.2-1 (Taken from (41))

By contrast, if we assigned a weight $w(Q)$ to a state represented by Q, and rewrite eq 4.2.2-1 with introduction of the weight, we can have eq 4.2.2-2:

$$< A >= \frac{\int dr^N d\Omega^N \left(\frac{A(r^N,\Omega^N)}{w(Q)}\right) w(Q) e^{-\beta V(r^N,\Omega^N)}}{\int dr^N d\Omega^N \left(\frac{1}{w(Q)}\right) w(Q) e^{-\beta V(r^N,\Omega^N)}}$$ eq 4.2.2-1 (Taken from (41))

Apparently, after adding bias $w(Q)$, factor $w(Q)e^{-\beta V(r^N,\Omega^N)}$ will be what we have for distribution probability while $\left(\frac{A(r^N,\Omega^N)}{w(Q)}\right)$ will be the samples we get from the simulation. With a brief process of normalizing with factor $\frac{1}{w(Q)}$, the unbiased sampling $A(r^N, \Omega^N)$ could be obtained.

Similar method was also proposed by Berg and Neuhaus (52) as a sampling technique to access rare events and transition states. Without this biased sampling, the energy different between the frequent evet and rare event is always rejected because of the large energy difference. Therefore, the rare event could be hardly sampled which leads to difficulty in calculating free energy difference. With umbrella sampling, a biased weight will be assigned to the rare event and thereby improved the sampling frequency of rare event. Specific algorithm will be discussed in the following part.

4.2.3 Order Parameter

To determine if a configuration is an interaction or non-interaction state, and to assign weights to certain states in a more convenient way, two-dimension order parameters are introduced to describe the issue. Order parameters could be treated at a simplified reaction coordinate to describe the reaction. In our study, the two-dimension order parameters are the two minimum distances between bases on the probe arms (extension arm and ligation arm) and the complementary regions on the target. (**Figure 4.2.3-1**)

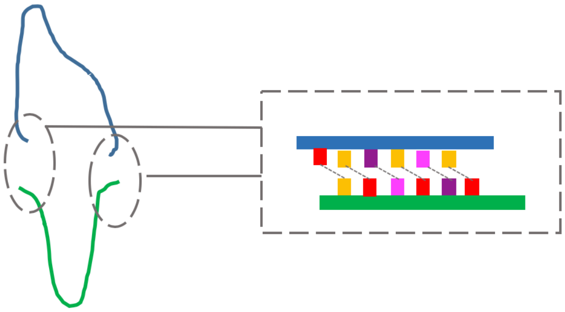

Figure 4.2.3-1 Distance between the probe arm and complementary region on target

For clarity, parameters Q are adopted to represent the minimum distance between the bases on the extension/ligation arm and the complementary bases on the targets. (Chart 4.2.3-2) The parameter Q represent distance which lies in the intervals of 0-1.70, 1.70-3.41, 3.41-5.11, 5.11-8.52, 8.52-12.78, 12.78-17.04, 17.04-34.07, 34.07-68.11, 68.11-102.22, >102.216nm, corresponding to Q values that range from 0 to 10.

In this way, using two Q values, Q_{ex} and Q_{lig} respectively for extension arm and ligation arm could well describe the state of the probe and target. Here we represent these two values by (Q_{ex}, Q_{lig}). For example, (10,10) ($Q_{ex} = 10,\ Q_{lig} = 10$) represents a complete open state for the probe and target ends are far for each other and both the two end-to-end distance is larger than 102nm. On the contrary, (0,0) ($Q_{ex} = 0,\ Q_{lig} = 0$) depicts an interaction state where both the two ends are close enough.

Q_{ee}	**End to end distance (nm)**
0	0-1.7036
1	1.7036-3.4072
2	3.4072-5.1108
…	…
10	>102

Chart 4.2.3-2 Mapping relationship between Q and distance between the probe arm and complementary region on target

4.2.4 Algorithm of the simulation

During every simulation step, the algorithm will first determine which state is the input initial configuration in using the manually set order parameter file. Next, there will be a movement proposed by VMMC algorithm, which step included a random seed selection and the generation of a cluster. The energy change is calculated simultaneously with the proposed movement of the cluster. To sample the interaction states, possibility of state-to-state transition p is further processed by the weights assigned for the initial and proposed states. With the improved transition possibility value, the rare event is easier to be accepted and therefore reach an event sampling. Then the movement will be accepted or rejected. At last, if the revised p value is larger than the random number generated by the computer, the new configuration will be accepted and stored in the histogram used for the free energy calculation. (**Figure 4.2.4**)

To give an example of how a rare state is sampled, given weight of state 0 (corresponding to order parameter $(Q_{ex_0}, Q_{lig_0}))$ as w_0 and the weight of state 1 (corresponding to order parameter

$(Q_{ex_1}, Q_{lig_1}))$ as w_1, the probability of a state transition from state 0 to state 1 is adapted by the factor $\frac{w_1}{w_0}$. Therefore, state 1 can be easily sampled even if the energy change from state 0 to state 1 is unacceptable. After sampling is finished, the diagram which stores the frequency of each state will be transformed into an unbiased distribution to eliminate the effect of the chosen weight.

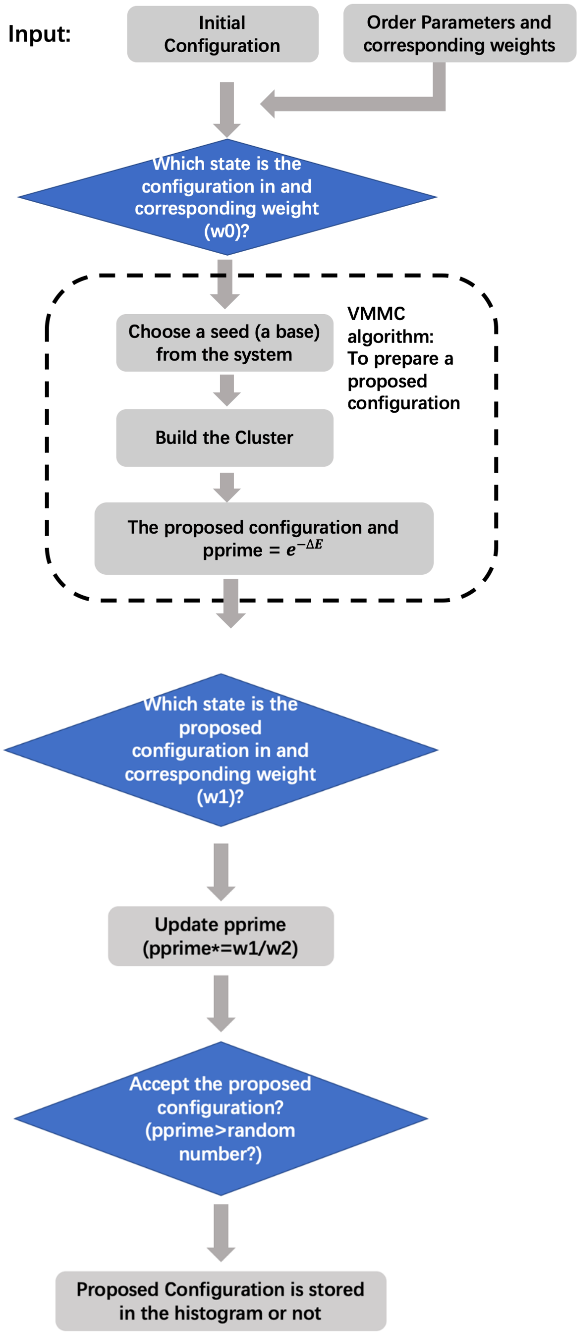
Input:
Initial Configuration
Order Parameters and corresponding weights
Which state is the configuration in and corresponding weight (w0)?
Choose a seed (a base) from the system
VMMC algorithm: To prepare a proposed configuration
Build the Cluster
The proposed configuration and pprime = $e^{-\Delta E}$
Which state is the proposed configuration in and corresponding weight (w1)?
Update pprime (pprime*=w1/w2)
Accept the proposed configuration? (pprime>random number?)
Proposed Configuration is stored in the histogram or not

Figure 4.2.4 Flow chart of the simulation algorithm

4.2.5 Experimental condition

Six target genes with their designed probes are considered as our samples. The length of these genes ranges from 400 base to 1500 base to ensure the coverage of this study. Simultaneously, a read value over 20 should be always guaranteed in order to avoid the potentially invalid experimental data used for comparison. The dimension of box side is set according to the size of target and probe in order that both are well contained inside the box. In addition, to reflect the salt environment of 10mM $[Mg^{2+}]$ and 100mM $[Na^{+}]$, the sodium concentration is set as 300mM (Since there is no magnesium concentration included in oxDNA, we used concentration of Na^{+} as a substitute consistent with known reports (53), which implied that DNA molecule demonstrated similar behavior under certain magnesium concentration and a much higher sodium concentration.)

4.2.6 Procedure of simulation and free energy difference calculation

The whole simulation could be sequentially divided into 3 stages: First, the weight assigned in the biased sampling should be adjusted iteratively until configurations under each state could be uniformly sampled and could easily transform from each other. This means the counts over each histogram bin should be comparable so that the whole histogram could be considered “flat”. Otherwise, the system may drop into a local potential well that was manually created, which could lead to inaccurate free energy profile. During this stage, a shell script was in conjunction with a python script to adjust the assigned weights according to the collective histogram data which includes the samples of every state. The adjustment is done every million steps. For example, after a round of sampling (one million step), we got h_0 samples in the bin of state 0 and h_1 samples in the bin of state 1. Simultaneously, state 0 was selected as a reference state to adjust the weight, which means the weight of state 0 $w_0$will keep unchanged. Then the weight of state 1 will be adapted to $w_1' = w_1 \cdot \frac{h_0}{h_1}$. The adjustment finishes until the minimum bin size is more than 80% of the average bin size. Second, the aforementioned weights are applied for equilibrium. Finally, sampling will begin to construct a free energy profile and further estimate the energy barrier that needs to cross for the capture of target ORF. Through the contrast between the two states, the free energy barrier for the target-probe interaction is calculated by eq 4.2.6.

$$\frac{\Delta G_{int}}{RT} = -\ln\left(\frac{P_{(Q_{ex}=0,Q_{lig}=0}}{P_{(Q_{ex}=10,Q_{lig}=10}}\right) \quad \text{eq } 4.2.6$$

4.3 VMMC simulation results

4.3.1 Calculation of free energy difference of probe-target interaction

In the scenario of capturing a DNA target, an improper probe length could lead to the mismatch of the end-to-end distances of the probe and target gene. Moreover, a longer probe may be ineffective due to the entropic cost of interacting with target genes. And the two factors may also contribute differently to the overall effect when the targets and probes vary. Going forward, we used computational simulations to understand this issue by estimating the free energy of probe-target interaction. (**Figure 4.3.1A**)

To quantitatively learned the free energy of target-probe interaction, MC simulation was conducted by oxDNA (19) . To avoid low acceptance rate and to further improve sampling, Visual Movement Monte Carlo (VMMC) (50) was applied. Simultaneously, umbrella sampling (31) was in conjunction with VMMC algorithm to access the rare states of target-probe interaction. With the help of these methods, an interaction state could be easily sampled with a frequency comparable to non-interaction states. The contrast of two states are shown in **Figure 4.3.1B**. An interaction state is defined as a scenario where both the two “arm distances” (distances between probe arm and the complementary region on target) are smaller than 1.7nm.

In the simulation, different weights were assigned to different states, which correspond to different values of the two-dimensional order parameters. A large weight is assigned to a rare state, which is generally rejected due to an unacceptable energy change, to increase the possibility of acceptance. Without such intervention, it is not possible to sample states of close probe-target interaction. After sampling, an unbiased sampling frequency can be gathered from the collective data of biased configurational samples and the weights originally assigned to the different states. An example for final unbiased sampling is shown in **Figure 4.3.1C**, which presents a slope-like distribution. The free energy is calculated from the two “corner” values, namely the two sampling frequency values of non-interaction states and states of interaction. The sampling condition on the cross section of the diagonal of the slope (marked by the red line in **Figure 4.3.1C**) is reflected in **Figure 4.3.1D**. The ΔG_{int} represents the calculated free energy of target-probe interaction, which was calculated from eq 4.2.6.

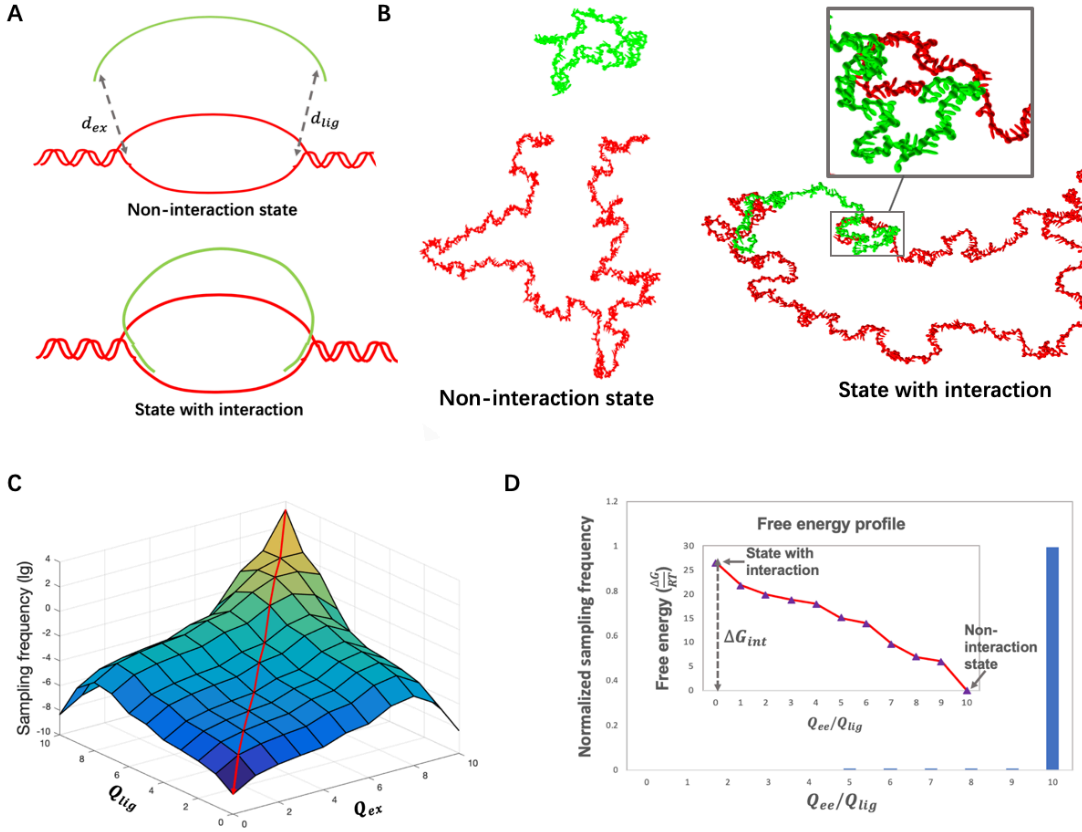

Figure 4.3.1 Calculation of probe-target interaction free energy (A) and (B): Diagram and oxDNA visualization showing non-interaction/interaction state of the target and gene. (A) A diagram showing the non-interaction/interaction states. d_{ex} and d_{lig} respectively represent the distances between the extension arm and its complementary region on the target gene and the ligation arm and its complementary region on the target gene. These two parameters are respectively represented by Q_{ex} and Q_{lig}. (B) Non-interaction/interaction states of the target and the probe visualized by cogli1, a tool to draw oxDNA configurations. The figure on the left shows a complete open state of the target (colored in red) and the probe (colored in green) in which both Q_{lig} and Q_{ex} equal 10. The figure on the right demonstrates an interaction state of the target and probe where both Q_{lig} and Q_{ex} equal 0. Each visualized nucleotide includes two beads, one representing the backbone and the other the base. There are two interaction sites on the base, including a stacking site as well as a hydrogen-bonding site.(41) (C) and (D): Free energy calculation for the E. coli gene nemR and molecular inversion probe (MIP). (C) An unbiased sampling frequency distribution over the whole space where the Q ranges from 0 to 10. (D) The free energy profile along the diagonal plane of part C (marked by the red line). The biased sampling frequency and the calculated free energy profile are shown. ΔG_{int}, the free energy of probe-target interaction is labeled in the figure.

4.3.2 Comparison with experiments

To evaluate whether the simulations account successfully for the experiments, the experimental data were represented in the form of relative free energy differences. The values of RPKM (Reads Per Kilobase per Million mapped reads)(54) were quantified using next-generation DNA sequencing after library cloning of E. Coli genes, aiming to represent the enrichment efficiency for certain probes.(55) It is reasonable to believe that these values could reflect the free energy difference. Therefore, RPKMs of the cloned copies for one target under the application of different probes were transformed into "relative free energy differences" by eq 4.3.2 so that a quantitative comparison could be conducted between simulation and experiment. During this process, we assumed that the experiment is equivalent to adding four probes simultaneously into a pool and that there is a competition among the four probes to interact with the same gene. A zero free energy level, which corresponds to the largest RPKM value, is regarded as the most favorable selection of AL and also as a zero free energy reference for simulations.

$$\Delta G_{probeA-probeB} = -\ln\left(\frac{RPKM_{probeA}}{RPKM_{probeB}}\right) \quad eq\ 4.3.2$$

Six groups of target genes and probes were subjected to VMMC simulation. The data were further transformed to relative free energies in order to compare simulations and experiments. **Figure 4.3.2 A-F** shows these results. Pearson correlation coefficients (56) were calculated to evaluate the correlation between the experimental results and the simulation predictions. Three out of six groups show relatively strong correlations (>0.7). Five out of six (~90%) provide predictions of AL preference consistent with experiment. The lowest free energy suggests the most favorable option of probes. As the target length increases, a transition in optimal AL from 54-nt (the MIP AL) to 242-nt and 442-nt could be observed. (compare **Figure 4.3.2A** vs **Figure 4.3.2E, F**) The relative interaction free energy of MIP increases when the target length changes from 400 to 600bp, which accounts for its low performance of long-target capture. This increase of interaction free energy of MIP provides a reason why LASSOs exhibit more favorable efficiency in capturing and cloning long targets from a computational aspect.

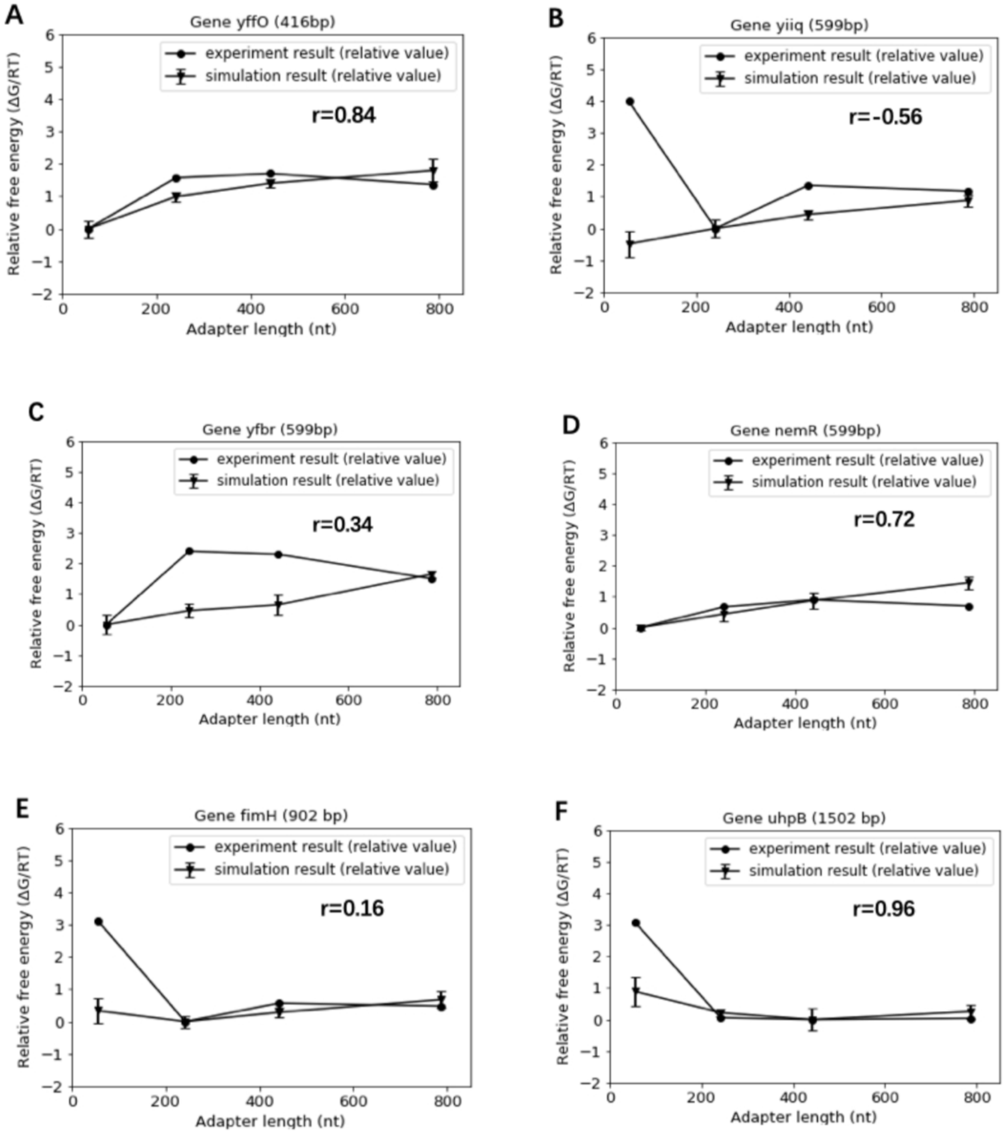

Figure 4.3.2 Comparison between experiment and simulation results. A to F represent six groups for data respectively for six E. coli genes. Pearson correlation coefficients are labeled as r value in the figures.

4.4 Discussion

LASSO probes can enable long-read DNA sequencing and multiplexed cloning efforts if they are highly robust and efficient. This study evaluated the adapter region of a LASSO probe, both experimentally and computationally, to determine the effects of AL on capture efficiency. For a given target size, it was determined that there was a distribution of efficiency for LASSO probes where intermediate adapter sizes led to the best capture results. These empirical results motivated an in-silico effort to understand LASSO probe binding and help account for these results. A coarse-grained model, oxDNA, partially reproduced the discrepancies in efficiency between different LASSOs and MIPs and provided a proof-of-concept that computational simulations could be a useful quality control check to design LASSOs in the future for optimal results given a pre-specified length of DNA target. There remain several open areas to build upon this first-generation computational model that can improve predictions.

The simplicity of the model should be firstly taken into consideration if we start to analyze the experiment-simulation difference from the aspect of simulation. The free energy of target-probe interaction is presumed to be the main factor that causes the efficiency discrepancy. The post-interaction hybridization of bases was ignored. This term is considered to have an equal contribution to the total target-capture free energies of all the probes since the MIPs and LASSOs for one gene share the same extension and ligation arms. Nevertheless, the internal stress which varies as the probe length changes could have an influence on the unzippering energy.(57) Considering this, the term may also influence the capturing performance of the probes due to the different constraints during hybridization.

Moreover, the target was treated as a DNA fragment that could move without constraints from surrounding genes. In reality, the target is likely located in large genome fragments due to the imprecision of sonicating DNA before a capture experiment is conducted. This factor may have an impact on the behavior of the target due to constraints associated with the long chains of the genome fragments. This might also be a clue behind some of the differences between experiment and simulation. Finally, the experimental errors can also be a result of artifacts from PCR amplification of a captured target library and sequencing itself.

Another factor that may take account of the difference between **Figure 2.1-1** and the simulation results is the intervention of non-target capture. The simulations were unable to include non-specific captures. Therefore, the model did better in predicting the most favorable probe for a specific target but it was not capable of considering a probe with the least side effects of non-target capture.

In summary, this chapter introduced a computational model of LASSO binding that roughly predicts experimental capture results. It is envisioned that, with further improvements of this computational model, such knowledge can help customize adapter sequences for future large library preparations and obviate the need for empirical determination of an optimal LASSO adapter size.

CHAPTER 5 CONCLUSION

As it has been shown by the previous results, LASSO probes provide high performance in capturing and cloning long DNA targets and thus become a promising candidate to address the technical problem in long DNA synthesis. A critical characteristic of LASSO is the design of its single strand LA. More interestingly, change of the AL leads to differences in capturing efficiency, which was clearly observed in experiments. Therefore, to explain the implicit mechanism of this difference will contribute to selecting an appropriate adapter for a specific target and further optimizing the design of LASSO. Two hypotheses respectively regarding the intervention of secondary structures and the free energy of probe-target interaction were proposed as potential explanations for this phenomenon. Secondary structures are predicted by Mfold sever and demonstrated largely reduced motifs at high temperature under experimental condition. On the other hand, free energies of probe-target interactions were predicted by simulation based on a coarse-grained model. The quantified free energies from the simulation were compared with the relative free energies which were transformed from RPKM values from sequencing results of library experiments that were used to evaluate the capturing and cloning efficiency. Some consistencies could be observed from this experiment-simulation comparison. This result implies that the probe-target interaction free energy is the main reason that causes the difference of capturing efficiency, which potentially becomes a critical criteria of pre-experiment design of LASSO. Furthermore, different preferences to different ALs, which occurs when the target length change, have been detected in simulation and some of the experiments. MIPs are more favorable when the target is relatively short (~400bp and ~600bp). On the other hand, LASSOs provide better performances when the target grows to around 1000 and 1500bp long.

By having further understanding and control of LASSO adapter sizes, there can be better assurance of the deployment of LASSO probes in a wide range of applications where versatility of these reagents will be necessary. Ultimately, this simulation engine can be built into a front-end user interface for the public to create LASSOs on their own with customized binding arms and ALs/sequences to enable broad adoption of the technique by the research community.

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7.

2. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107(1):1-8.

3. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.

4. Schatz MC, Phillippy AM. The rise of a digital immune system. Gigascience. 2012;1(1):4.

5. Beaucage SL, Caruthers MH. Deoxynucleoside Phosphoramidites - a New Class of Key Intermediates for Deoxypolynucleotide Synthesis. Tetrahedron Lett. 1981;22(20):1859-62.

6. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-Directed, Spatially Addressable Parallel Chemical Synthesis. Science. 1991;251(4995):767-73.

7. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. Nature Methods. 2014;11(5):499-507.

8. Nilsson M, Dahl F, Larsson C, Gullberg M, Stenberg. Analyzing genes using closing and replicating circles. Trends Biotechnol. 2006;24(2):83-8.

9. Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci U S A. 2007;104(22):9387-92.

10. Landegren U, Schallmeiner E, Nilsson M, Fredriksson S, Baner J, Gullberg M, et al. Molecular tools for a molecular medicine: analyzing genes, transcripts and proteins using padlock and proximity probes. J Mol Recognit. 2004;17(3):194-7.

11. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418-26.

12. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78-94.

13. Claverie JM. Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet. 1997;6(10):1735-44.

14. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23(6):673-9.

15. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods. 2009;6(5):315-6.

16. Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R. A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc Natl Acad Sci U S A. 2008;105(27):9296-301.

17. Shen P, Wang W, Krishnakumar S, Palm C, Chi AK, Enns GM, et al. High-quality DNA sequence capture of 524 disease candidate genes. Proc Natl Acad Sci U S A. 2011;108(16):6549-54.

18. Tosi L, Sridhara V, Yang Y, Guan D, Shpilker P, Segata N, et al. Long-adapter single-strand oligonucleotide probes for the massively multiplexed cloning of kilobase genome regions. Nat Biomed Eng. 2017;1.

19. Ouldridge TE, Louis AA, Doye JP. DNA nanotweezers studied with a coarse-grained model of DNA. Phys Rev Lett. 2010;104(17):178101.

20. Kim SH, Quigley GJ, Suddath FL, McPherson A, Sneden D, Kim JJ, et al. Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. Science. 1973;179(4070):285-8.

21. Pabo CO, Sauer RT. Protein-DNA recognition. Annu Rev Biochem. 1984;53:293-321.

22. Tinoco I, Jr., Uhlenbeck OC, Levine MD. Estimation of secondary structure in ribonucleic acids. Nature. 1971;230(5293):362-7.

23. Larsen N, Zwieb C. SRP-RNA sequence alignment and secondary structure. Nucleic Acids Res. 1991;19(2):209-15.

24. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406-15.

25. Zuker M. On finding all suboptimal foldings of an RNA molecule. Science. 1989;244(4900):48-52.

26. Hospital A, Goni JR, Orozco M, Gelpi JL. Molecular dynamics simulations: advances and applications. Adv Appl Bioinform Chem. 2015;8:37-47.

27. Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications: Elsevier; 2001.

28. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Equation of state calculations by fast computing machines. J. Chem. Phys. 1953;21(6):1087-92.

29. Kirkwood John G. Statistical mechanics of fluid mixtures. J. Chem. Phys.1935;3(5):300-13.

30. Zwanzig RW. High-temperature equation of state by a perturbation method. I. Nonpolar gases. J. Chem. Phys. 1954;22(8):1420-6.

31. Torrie GM, Valleau JP Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J. Chem. 1977;23(2):187-99.

32. Liu W, Saint DA. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. Anal Biochem. 2002;302(1):52-9.

33. da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A. SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. Int J Plant Genomics. 2008;2008:412696.

34. Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics. 2014;30(18):2670-2.

35. Diep D, Plongthongkum N, Gore A, Fung HL, Shoemaker R, Zhang K. Library-free methylation sequencing with bisulfite padlock probes. Nat Methods. 2012;9(3):270-2.

36. SantaLucia J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Natl. Acad. Sci. U.S.A. 1998;95(4):1460-5.

37. SantaLucia Jr J, Hicks DJ. The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struc. 2004;33:415-40.

38. Everaers R, Kumar S, Simm C. Unified description of poly-and oligonucleotide DNA melting: Nearest-neighbor, Poland-Sheraga, and lattice models. Phys. Rev. E. Stat. Nanlin Soft Matter Phys. 2007;75(4):041918.

39. Becker NB, Everaers RJ. DNA nanomechanics: How proteins deform the double helix. J. Chem. Phys. 2009;130(13):04B602.

40. Savelyev A, Papoian GA. Chemically accurate coarse graining of double-stranded DNA.Nat. Aca Sci. 2010;107(47):20340-5.

41. Ouldridge TE. Coarse-grained modelling of DNA and DNA self-assembly: Springer Science & Business Media; 2012. https://ora.ox.ac.uk/objects/uuid:b2415bb2-7975-4f59-b5e2-8c022b4a3719

42. Hill TL. An introduction to statistical thermodynamics: Courier Corporation; 1986. https://books.google.com/books?hl=en&lr=&id=QttFDwAAQBAJ&oi=fnd&pg=PP1&dq=Hill+TL.+An+introduction+to+statistical+thermodynamics&ots=_O7CWv7FHS&sig=629cWT0e7WcOQICUiGcbXigbIyk - v=onepage&q=Hill TL. An introduction to statistical thermodynamics&f=false

43. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleid Acid Res. 2006;34(2):564-74.

44. Ouldridge TE, Sulc P, Romano F, Doye JP, Louis AA. DNA hybridization kinetics: zippering, internal displacement and sequence dependence. Nucleic Acids Res. 2013;41(19):8886-95.

45. Sulc P, Romano F, Ouldridge TE, Rovigatti L, Doye JP, Louis AA. Sequence-dependent thermodynamics of a coarse-grained DNA model. J Chem Phys. 2012;137(13):135101.

46. Romano F, Hudson A, Doye JP, Ouldridge TE, Louis AA. The effect of topology on the structure and free energy landscape of DNA kissing complexes. J Chem Phys. 2012;136(21):215102.

47. Harrison RM, Romano F, Ouldridge TE, Louis AA, Doye JPJapa. Coarse-grained modelling of strong DNA bending II: cyclization. 2015. https://arxiv.org/abs/1506.09008

48. Wang Q, Pettitt BM. Sequence Affects the Cyclization of DNA Minicircles. J Phys Chem Lett. 2016;7(6):1042-6.

49. Ouldridge TE, Louis AA, Doye JP. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. J. Chem. Phys. 2011;134(8):02B627.

50. Whitelam S, Geissler PL. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. J. Chem. Phys. 2007;127(15):154101.

51. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PAJJocc. The weighted histogram analysis method for free‐energy calculations on biomolecules. I. The method. 1992;13(8):1011-21.

52. Berg BA, Neuhaus T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. Phys Rev Lett. 1992;68(1):9.

53. Grigoryev SA, Arya G, Correll S, Woodcock CL, Schlick T. Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. Proc Natl Acad Sci U S A. 2009;106(32):13317-22.

54. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8.

55. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

56. Lee Rodgers J, Nicewander WA. Thirteen ways to look at the correlation coefficient. Am Stat. 1988;42(1):59-66.

57. Tee SR, Wang ZJ. How well can DNA rupture DNA? Shearing and unzipping forces inside DNA nanostructures. ACS Omega. 2018;3(1):292-301.