University of Central Florida
STARS

Honors Undergraduate Theses

UCF Theses and Dissertations

2020

Development of a Computer Algorithm for Generation of Primers for Nucleic Acid Sequence Based Amplification (NASBA)

Rohit Karnati
University of Central Florida

Recommended Citation

Karnati, Rohit, "Development of a Computer Algorithm for Generation of Primers for Nucleic Acid Sequence Based Amplification (NASBA)" (2020). *Honors Undergraduate Theses*. 723.
https://stars.library.ucf.edu/honorstheses/723

DEVELOPMENT OF A COMPUTER ALGORITIHM FOR GENERATION OF PRIMERS FOR NUCLEIC ACID SEQUENCE BASED AMPLIFICATION(NASBA)

BY

ROHIT KARNATI

A thesis submitted in partial fulfillment of the requirements
for the Honors in the Major Program in Biochemistry
in the College of Sciences
and in the Burnett Honors College
at the University of Central Florida
Orlando, Florida

Spring Term, 2020

Thesis Chair: Yulia Gerasimova, Ph.D.

Abstract

Nucleic acid sequence based amplification (NASBA) is a primer based isothermal method of RNA/DNA amplification. Currently, primer design for NASBA has been restricted to hand creating sequences of oligonucleotides that must follow a set of rules to be compatible for the amplification process. This process of hand-creating primers is prone to error and time intensive. The detection of mutants, post amplification, also offers a benefit in point of care scenarios and the design of hybridization probes for sequences in the region of amplification is also an erroneous and time intensive process. By creating a program to design primers and hybridization probes based on the set of rules provided for a sequence of user input DNA or RNA, one can avoid costly errors in primers design and save time. Utilizing Python (a high-level object-oriented programming language), along with a series of bioinformatic libraries such as Biopython and UNAfold one can definitively choose the best primer sequences for a given sample of DNA.

Acknowledgements

Foremost, would like to express my deepest gratitude to my thesis advisor, Dr.Yulia Gerasimova, for all her assistance and commitment to helping me through the entirety of the thesis project. I'd like to thank my committee chairs, Dr.Shaojie Zhang and Dr.Jonathan Caranto, for providing their knowledge and expertise to help make the project go along much more smoothly. I would also like to acknowledge all the time and help Ryan Connelly has provided me in addition to also providing the foundation for the project. I'd like to thank Mark Reed for all his help within the lab and also preparing some reagents for the experiments in this project. I'd also like to thank the rest of my lab mates for providing me their support and assistance in dealing with any roadblocks that arose.

I'd finally like thank Dr.Michael Zucker for graciously allowing me to use UNAfold at no cost.

Table of Contents

List of Figures

List of Tables

Introduction

1.1 Nucleic Acid Amplification Methods

Nucleic acid amplification has been an area of interest for biochemical and molecular biology research. Nucleic acid amplification relies on the activity of DNA and/or RNA polymerases – the enzymes that are responsible for DNA and RNA synthesis in living cells. Prokaryotic DNA replication is driven by DNA polymerase I (pol I). The Klenow fragment is a large portion pol I which lacks 5’ to 3’ exonuclease activity which catalyzes the stepwise addition of deoxyribonucleoside-5’-triphosphates (dNTPs) to the 3’-OH terminus of a nucleic acid primer based on the nucleotide sequence of a DNA template (Figure1).

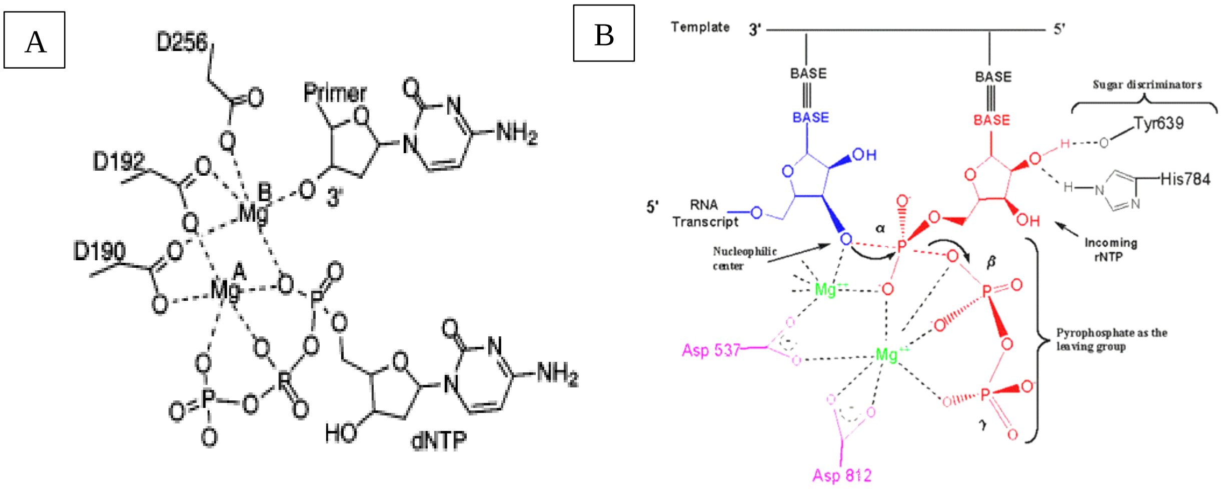


Figure 1: Panel A depicts the enzyme catalyzed reaction of nucleic acid polymerization in human DNA polymerase. The enzyme incorporates a divalent cation, such as Mg^{2+}, to form a coordinate ionic bridge between acidic amino acid residues (in this case aspartic acid), the 3’-OH of the growing strand, and the phosphates of the dNTP[1]. Panel B depicts a similar mechanism for T7 RNA polymerase[2].

It has been extensively used in molecular biology to synthesize a DNA strand complementary to the DNA template. Polymerase Chain Reaction (PCR) is a novel approach at utilizing DNA polymerase to create a double stranded DNA (dsDNA) amplicon form a single dsDNA analyte [1]. An intrinsic 3'→5' exonuclease activity of the Klenow fragment ensured high accuracy of the DNA fragment copying, but the thermal instability of the enzyme required the addition of a fresh portion of the enzyme for every new cycle. Discovery of thermostable DNA polymerases such as Taq polymerase isolated from the thermophilic bacterium *Thermus aquaticus* [1] allowed a means for PCR to be created.

The mechanism of PCR (Figure 2) consists of the following stages: denaturing the double-stranded template sequence, annealing the primers to the complementary fragments of the template by lowering the temperature, and extending with a thermal resistant enzyme [3]. All stages are performed in a thermal cycler, which ensures fast temperature changes and precise temperature control. PCR-based methods have found wide applications in sequencing, genotyping, gene cloning, and characterization of gene-related illness. However, for diagnostic applications, especially at the point-of-care (POC), reliance on the thermal cycler limits the affordability of PCR-based methods. The range of thermal stages, which causes the need for a thermal cycler in PCR, can be replaced with a relatively small number of thermal stages by using other amplification methods that can allow denaturation of double stranded nucleic acids in the same temperature range that is required fort the synthesis of a newly synthesized amplicon. This possibility for single-temperature, or isothermal, nucleic acid amplification caused several new amplification methods to be developed.

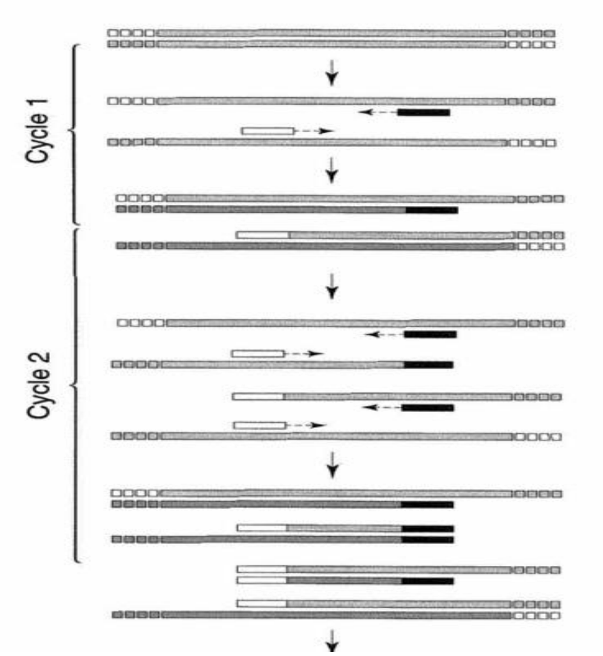


Figure 2: Schematics of Polymerase Chain Reaction (PCR) cycling shown through 2 cycles. It should be noted that classic PCR deals with amplification of double stranded DNA but other variations exist such as reverse transcriptase PCR (RT-PCR) that can amplify RNA. The amount of amplicon can be estimated by using the formula 2^n where n is the number of cycles [1].

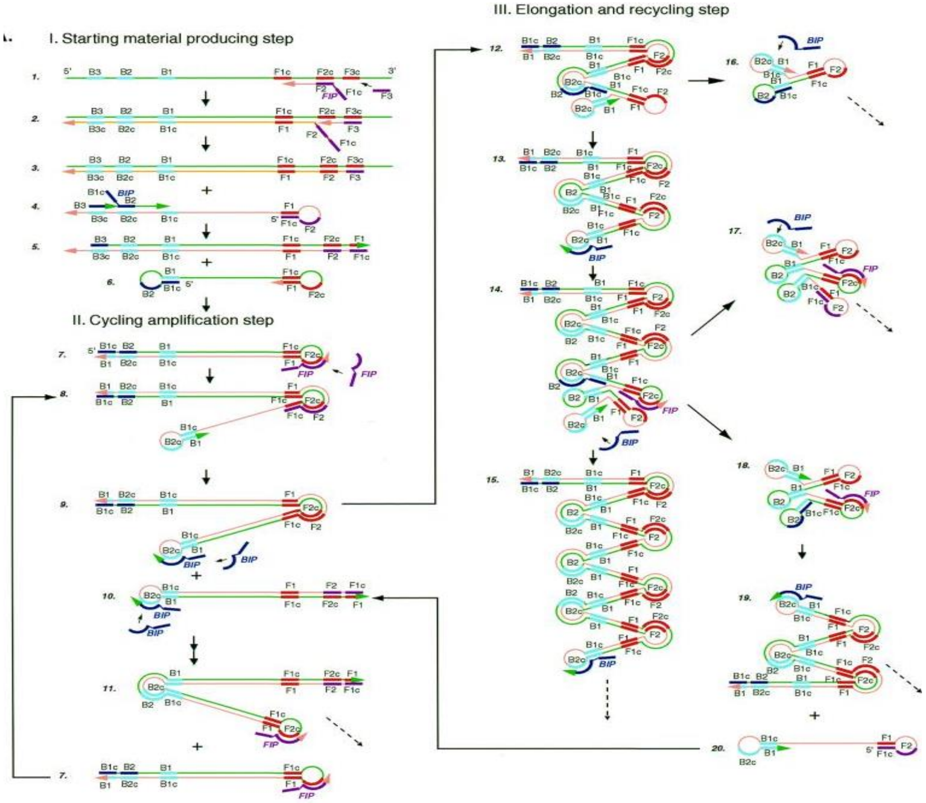

I. Starting material producing step
II. Cycling amplification step
III. Elongation and recycling step
FIP
BIP

Figure 3: Schematics for Loop mediated isothermal AMPlification (LAMP). LAMP is distinct from other forms of amplification in that it uses 4 different primers to create an amplicon with multiple stem loops [3].

There are a number of DNA and RNA amplification strategies that can be performed at a constant temperature. Different techniques offer their own strengths and setbacks, but the overall goal for all the techniques is to increase the amount of a nucleic acid fragment of interest efficiently without the need of precise temperature control and cycling**.** The most widely used methods for isothermal amplification include Loop mediated isothermal AMPlification (LAMP) and Strand Displacement Amplification (SDA). LAMP is conducted at 65 °C and utilizes four (or six) different primers to create side by side inverted repeats of the target sequence (Figure 3) [2]. The product would require another round of denaturation for a fluorescently labeled probe to bind and allow for quantification of the amplification product.

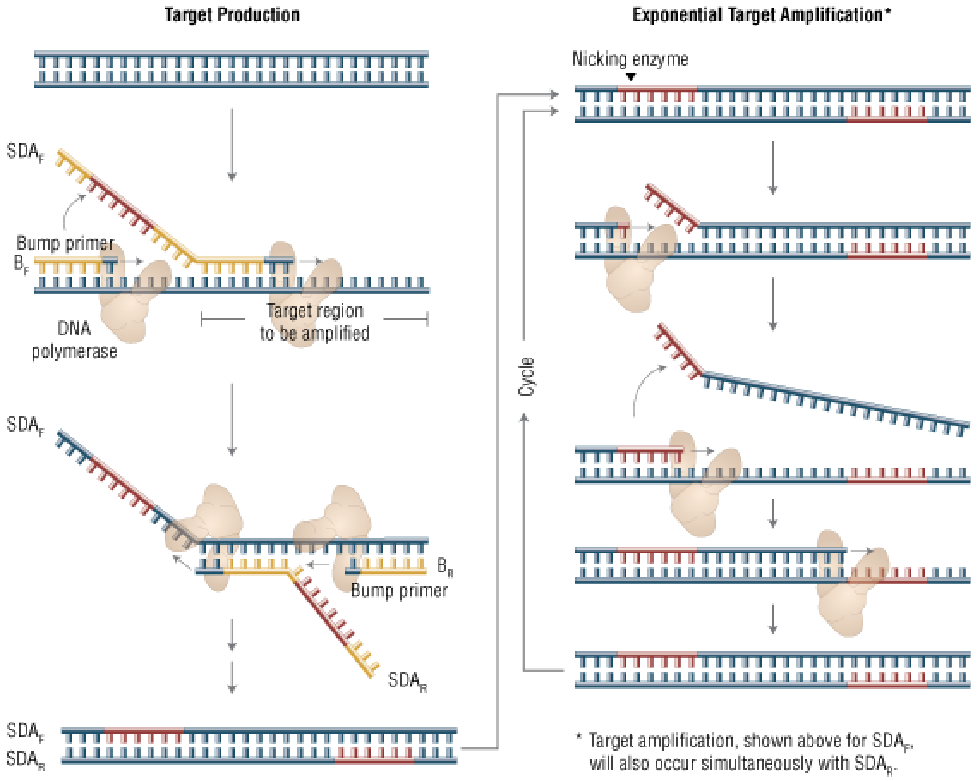

Target Production
Exponential Target Amplification*
Nicking enzyme
SDA$_F$
Bump primer
B$_F$
DNA polymerase
Target region to be amplified
SDA$_F$
B$_R$
Bump primer
SDA$_R$
SDA$_F$
SDA$_R$
Cycle
* Target amplification, shown above for SDA$_F$, will also occur simultaneously with SDA$_R$.

Figure 4: Schematic of strand displacement amplification (SDA). SDA can be identified by the extension of forwards and reverse primers to make two doubles stranded products that will then be nicked and displaced to form amplicon[4].

SDA is conducted at 37 °C and utilizes a nicking restriction endonuclease to nick a double-stranded DNA (dsDNA) amplicon and a DNA polymerase with strand-displacing activity, such as exo-Klenow or Bst DNA polymerase, to elongate the 3’-OH containing nicked fragment of one template strand and displace the other nicked fragment. (Figure 4). Repetition of polymerization-nicking-extension cycles allows for exponential amplification of the template fragment. Interrogation of SDA amplicons is limited due to generation of a double-stranded amplicon. Recent developments have been made in the detection of SDA amplicon such as DNA probes that successfully bind double stranded DNA [3], or the use of intercalating dyes that can be detected through gel electrophoresis.

Another similar method of amplification is Nicking Enzyme Amplification Reaction (NEAR) offers a simple and fast approach at isothermal amplification. NEAR can be characterized by its ability to provide nucleic acid amplification without the need of complicated primer design. It utilizes nicking enzyme and a phosphorylated template strand to create single stranded DNA amplicon in an isothermal fashion [2].

All of the above-mentioned techniques except NEAR (which produces a short single stranded DNA amplicon) produce double-stranded DNA (dsDNA) amplicons, which complicates the downstream analysis of the sequence of the amplicon using hybridization probes. To overcome this limitation, additional enzymes that specifically cleave unwanted strands from the dsDNA product, or asymmetric amplification conditions have been suggested. This solution adds complexity to the amplification reaction and/or compromises the efficiency of the reaction [2]. On the contrary, transcription-based methods generate a single-stranded RNA (ssRNA) amplicon, which can be conveniently interrogated with a complementary probe in the follow-up analysis. Examples of such isothermal techniques are Transcription-Mediated Amplification (TMA) and Nucleic Acid Sequence Based Amplification (NASBA) [5]. TMA is able to synthesize ssRNA amplicon by utlizing a dsDNA that is transcribes from an initial ssRNA template. This process is isothermal and shows amplification faster than that of PCR, since more than 100 copies of RNA amplicon can be made per cycle in TMA as opposed to 2 in PCR. The mechanism of the NASBA reaction is described in Figure 4. It requires the coordinated action of the three enzymes – Reverse Transcriptase (RT) from the Avian Myeloblastosis virus (AMV), ribonuclease H (RNase H), and T7 RNA polymerase. AMV RT catalyzes the synthesis of a complementary DNA (cDNA) sequence based on an RNA template and is naturally employed by retroviruses [3]. RNase H is an enzyme that catalyzes the cleavage of RNA in the RNA/DNA hybrid using a hydrolytic mechanism [3]. T7 RNA polymerase (T7 DdRp) recognizes the double stranded T7 promoter of the product and catalyzes the downstream synthesis of RNA from a template of dsDNA [6]. An initial template of single-stranded RNA, the positive sense strand, is annealed with the first primer (P1) that includes a T7 promoter region. The primer is then elongated by AMV RT at 41 °C using deoxynucleoside-5'-triphosphates (dNTPs). This extension product is isolated from the DNA/RNA hybrid by RNase H, which degrades the RNA portion of the hybrid. A second primer (P2) attaches to the cDNA formed, at which point the DNA polymerase activity of AMV RT allows the formation of a dsDNA. T7 RNA DdRp recognizes the double stranded promoter region and then forms a complementary antisense RNA strand from a template of dsDNA being read form 5' to 3'. The antisense RNA is bound by P2, causing AMV RT to create a cDNA copy of the RNA. The RNA/cDNA hybrid is cleaved by RNase H, producing a cDNA strand that is then bound by P1. AMV RT creates another dsDNA product which is used to create more antisense RNA from the T7 promoter region by T7 RNA polymerase. These steps can be cycled multiple times to produce more antisense RNA [3].

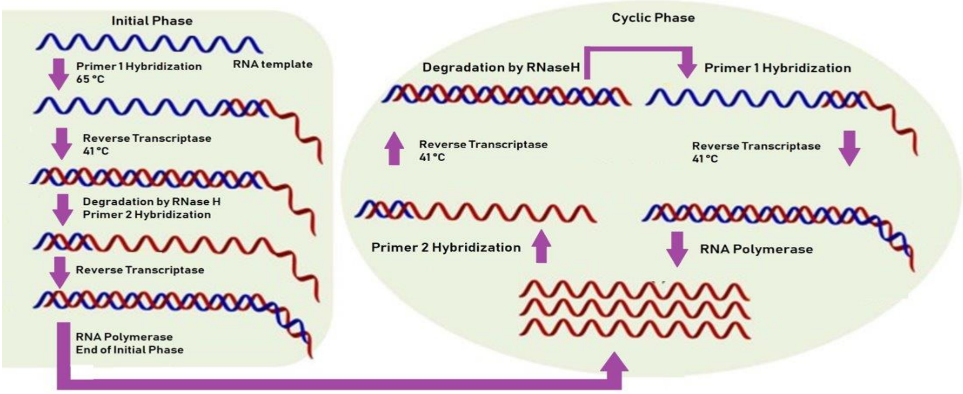


Figure 5: Schematics of the NASBA reaction. The initiation phase can be described as an attempt at creating a double stranded DNA from RNA using RT and RNase. The cyclic phase that follows can be described as making negative sense RNA from a template of double stranded DNA, which serves as a precursor for more double stranded DNA [6]

The NASBA reaction was originally proposed to amplify genome fragments of RNA viruses [6]. For sequence-specific analysis of NASBA products, a molecular beacon probe has been used, which also allows for real time quantification of the RNA amplicon in a sample [4]. NASBA is primarily used for RNA amplification but can be used for DNA amplification, though it is not as efficient in the latter due to the addition of two more denaturing steps [3]. The two denaturing steps account for denaturing dsDNA of the template and creating a positive sense RNA amplicon. Following the first denaturing step at 95°C, P1 will anneal to the DNA.AMV RT will form a double-stranded DNA as an extension product of P1 with a T7 promoter region. After another denaturing step, P2 (the reverse primer) will start the formation of another extension product that is complementary to the first extension product. The extension products form a double-stranded DNA product with a double-stranded T7 promoter region on the 5’ end. At this point, AMV RT has denatured due to the second heating and must be resupplied to the system along with T7 DdRp and RNase H. T7 DdRp recognizes the double-stranded promoter region of the double stranded DNA product and the cyclic phase of NASBA can commence as normal.

NASBA has shown great potential for POC scenarios due to it having a greater sensitivity than commonly used amplification methods in the detection of certain viruses and bacteria such as hepatitis C virus, West Nile virus and *M. pneumoniae*. It was reported that NASBA had limit of detection 1000 times than that of RT-PCR for the detection of West Nile Virus while being a less expensive option [2]. NASBA was also used to detect hepatitus C virus, for which NASBA exhibited a sensitivity 1000 times greater to the commonly used Quantiplex HCV-RNA assay [2]. NASBA has also been used in the detection of variations in the 16S rRNA fragment of *M. pneumoniae* types 1 and 2 which originally took 6 days with the commonly used charcoal differential assay (CCDA) and can now be done in 26 hours [5].

Given NASBA's success in POC situations, it should be noted that proper primer design is vital for a successful run through of NASBA. A set of empirical rules reported in the literature [3] is currently used for manual NASBA primer design. The cumbersome process of primer design requires a large number of primers to be made and tested to decide the best sequences. The manual method for primer design is time intensive and can be prone to error. There have been other methods of amplification with similar setbacks in primer generation that have programs made to simplify the process. For example, primer design software is available for the PCR (e.g. primer-BLAST) and LAMP reactions (e.g. PrimerExplorer), which were shown to have similar obstacles in a manual approach. Although the programs mentioned prior are similar in their approach at primer design and selection, the specific structure of NASBA primers, such as the T7 promoter region, is local to this amplification method and requires a new program. Most of these programs are also not open source and offer no functionality for the detection of mutants. This can be done within the region of amplification by detecting misalignments between the wild type and mutant genomes and designing hybridization probes for those misalignments. Given that there are a consistent set of rules, primers and probes for mutants can be made via a program for a user-defined region of a genome. The purpose of this project was to create an efficient program capable of designing NASBA primers and probes for mutants within the region of amplification by utilizing a number of well-established bioinformatics libraries and applications such as Biopython and UNAFold.

1.2 Language and Setup

Python's extensive biotechnology-based libraries, along with its ability to communicate between multiple applications written in different languages, makes it the language of choice for implementation of a procedural algorithm such as the one used in NASBA primer development. Python is also a scripting language that can make code easily understood between multiple people, making it a viable option for a program that could be repurposed for similar projects. Use of python and the compilation of Python code requires the use of a development environment Integrated DeveLopment Environment (IDLE). Use of any extraneous tools or libraries in Python requires them to be included in the same directory as the program using them.

1.3 Biopython

Biopython is an extensive Python based bioinformatics package that works to incorporate many useful tools in a single repository. One such useful tool is BLAST (Basic local alignment search tool), an algorithm for finding the statistical significance in the similarity between a query and a database of choice. Figure 6 shows in greater detail how BLAST works to create an alignment. A score is assigned by the number of matches and mismatches for a set of three nucleotides or amino acids, called a query word, through a local alignment. The maximum score for the combinations of local alignments is used to identify the correct alignment between the queries and subject. BLAST will be useful in creating an alignment of multiple genomes to create a multiple alignment object (a data type found that orders a number of queries against a subject). Figure 7 shows a visualization of a multiple alignment object, which allows one to compare a wild type genome or protein against multiple mutant genomes to amino acid sequences. Biopython also includes the Entrez module which is able to access the NCBI database to retrieve a genome or amino acid sequence of choice with a tag associated with each one, called an accession number.

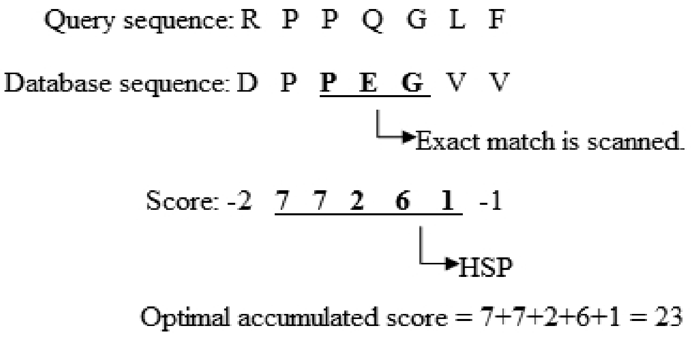
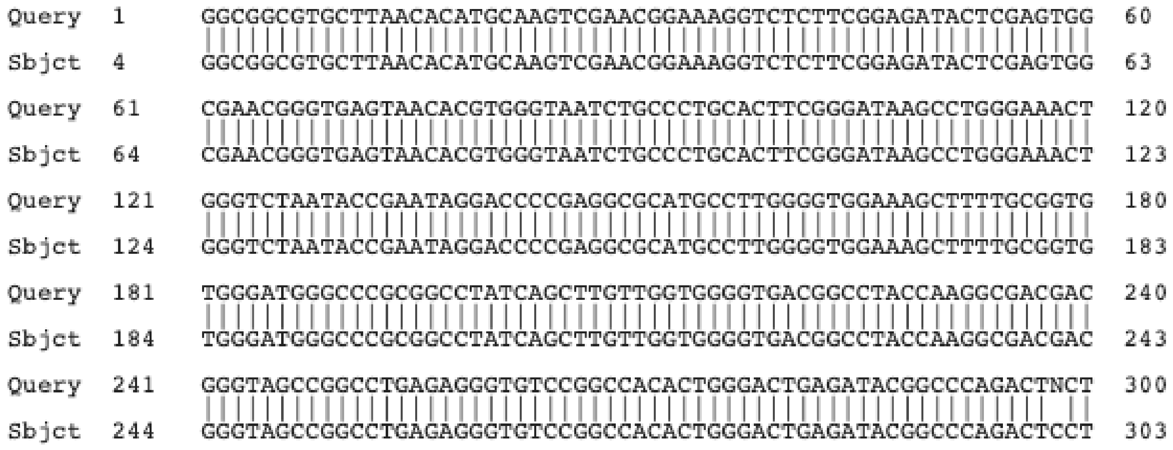

Query sequence: R P P Q G L F
Database sequence: D P P E G V V
Exact match is scanned.
Score: -2 7 7 2 6 1 -1
HSP
Optimal accumulated score = 7+7+2+6+1 = 23

Figure 6: The BLAST search algorithm. This process is used to score different possible alignments until the one with the best score is considered the optimal alignment. This method takes deletions into account as well [7]

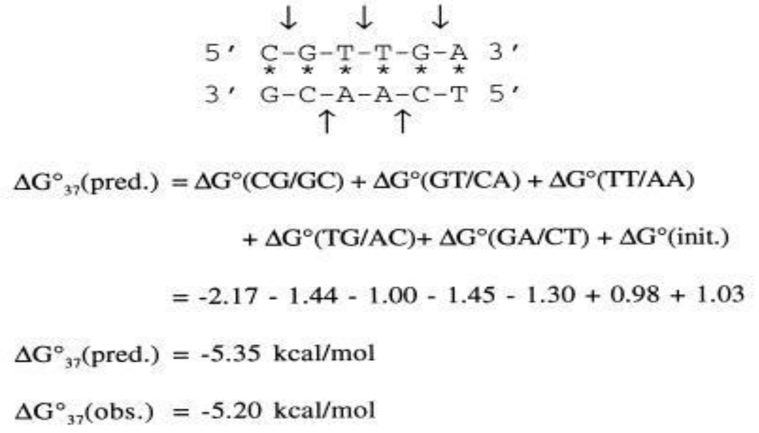


```
Query  1    GGCGGCGTGCTTAACACATGCAAGTCGAACGGAAAGGTCTCTTCGGAGATACTCGAGTGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4    GGCGGCGTGCTTAACACATGCAAGTCGAACGGAAAGGTCTCTTCGGAGATACTCGAGTGG  63

Query  61   CGAACGGGTGAGTAACACGTGGGTAATCTGCCCTGCACTTCGGGATAAGCCTGGGAAACT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  64   CGAACGGGTGAGTAACACGTGGGTAATCTGCCCTGCACTTCGGGATAAGCCTGGGAAACT  123

Query  121  GGGTCTAATACCGAATAGGACCCCGAGGCGCATGCCTTGGGGTGGAAAGCTTTTGCGGTG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  124  GGGTCTAATACCGAATAGGACCCCGAGGCGCATGCCTTGGGGTGGAAAGCTTTTGCGGTG  183

Query  181  TGGGATGGCCCGCGGCCTATCAGCTTGTTGGTGGGGTGACGGCCTACCAAGGCGACGAC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  184  TGGGATGGCCCGCGGCCTATCAGCTTGTTGGTGGGGTGACGGCCTACCAAGGCGACGAC  243

Query  241  GGGTAGCCGGCCTGAGAGGGTGTCCGGCCACACTGGGACTGAGATACGGCCCAGACTNCT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  244  GGGTAGCCGGCCTGAGAGGGTGTCCGGCCACACTGGGACTGAGATACGGCCCAGACTCCT  303
```

Figure 7: A multiple alignment object for a mutant of Mycobacterium abscessus. The program will have the same approach in ordering mutant genomes (aminoglycoside resistant Mycobacterium abscessus) against a wild type genome (the subject)

1.4 UNAfold

UNAfold is a Linux-based, command line-driven package of programs that determine nucleic acid secondary structure and hybridization through a series of thermodynamic calculations. UNAfold's thermodynamic calculations for free energy and assumptions for entropy are outlined by "nearest neighbor Watson-Crick base pair" [7] method shown in Figure 8.



$$\Delta G^\circ_{37}(\text{pred.}) = \Delta G^\circ(\text{CG/GC}) + \Delta G^\circ(\text{GT/CA}) + \Delta G^\circ(\text{TT/AA}) + \Delta G^\circ(\text{TG/AC}) + \Delta G^\circ(\text{GA/CT}) + \Delta G^\circ(\text{init.})$$

$$= -2.17 - 1.44 - 1.00 - 1.45 - 1.30 + 0.98 + 1.03$$

$$\Delta G^\circ_{37}(\text{pred.}) = -5.35 \text{ kcal/mol}$$

$$\Delta G^\circ_{37}(\text{obs.}) = -5.20 \text{ kcal/mol}$$

Figure 8: Calculations for a nucleic acid sequence using the nearest neighbor Watson-Crick Base pair method. All nearest neighbor free energies are literature values that can be referenced from the sources provided in this paper [5].

Materials and Methods

2.1 Methods for Program Development

As previously mentioned, since the program is written in Python, the program will require a development environment for compilation such as Integrated DeveLopment Environment (IDLE) or through an extension in IDLE. All packages required, except for UNAFold, are in the folder or directory that the Python script will be executed in for ease of use. It is crucial that the script is run on either a Linux or Mac terminal due to the system commands used in the program.

The multi sequence alignment that was discussed before in Figure 6 can be made for any retrieved wild type genome and mutant genome. An alignment can be useful in pinpointing mutations after amplifying a region of the wild type genome or finding extraneous primer binding sites within the genome. Biopython also includes a query system to retrieve multiple genomes through its Entrez module. By utilizing the Entrez module and BLAST from Biopython, the program retrieves a wild type genome along with its corresponding mutant genomes, if not already retrieved locally, and create a multiple alignment object. Primers can be designed for a region of the wild type genome and mutations found in the same region of the multiple alignment object can have sensors designed for post amplification detection.

The program will only need to utilize the UNAfold scripts for thermodynamic calculations of single and double-stranded nucleic acids in order to determine the most energetically favorable primer and probe complexes. All the free energies of nucleic acid strands are computed for 41 °C and the system they are in is assumed to be in equilibrium. The hybrid2 script has an output of the minimum free energy (mfe) of the target, primer, and target/primer complex. The hybrid2 script can also output the mol fraction of target, primer, and target/primer complex at different temperatures. The latter will be more useful for the program due to the DNA NASBA primers binding at 41 °C. The Perl script, hybrid2.pl, can be called with the system command:

```
hybrid2.pl --tmin  --tmax  --NA DNA --sodium\  magnesium
--exclude B --exclude BB --\ A0 5e-8 --B0 1e-7
target.seq.
```

2.2 Materials and Instruments

Monarch® Total RNA Miniprep Kit was purchased from New England BioLabs® Inc. (Ipswich, MA). NASBA Liquid Kit Complete was purchased from Life Sciences Advanced Technologies (St. Petersburg, FL). Oligonucleotides (primers for NASBA) purchased from IDT, Inc. (Coralville, IA) and used without purification. RNase/DNase free water was obtained from ThermoFischer Scientific (Waltham, MA) and used in all the experiments. Luria-Bertani (LB) medium was from Acros Organics Inc.(Waltham, MA), lysozyme was from Ambion Inc.(Austin,TX).

Concentrations of oligonucleotide stocks and total bacterial RNA were determined based on the absorbance measurements performed using a NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer from ThermoFisher Scientific (Waltham, MA). NASBA reactions were done using a C100 Touch Thermocycler purchased from Bio-Rad

(Hercules, CA). The agarose gel was imaged using a BIO-RAD Gel Documentation system from Bio-Rad Laboratories Inc. (Hercules, California).

2.3 Total RNA preparation

Total RNA was isolated from *E. coli* to serve as a template for the NASBA reaction. Initially, *E.coli* was inoculated within 15mL of LB (Luria-Bertani) liquid medium and grown until the optical density of the sample (measured at a wavelength of 600 nm) is 1 (measured by using a NanoDrop^{TM} One - UV-Vis Spectrophotometer). The cell were pelleted by centrifugation at 16000×g, 4ºC for 10 minutes. The cells were resuspended in RNase free water to a volume of 500 µL. The sample was then vortexed until homogenous and incubated with 1 mg/mL lysozyme for 5 min at room temperature (20 ºC). 1 mL of RNA Lysis buffer from the Monarch Total RNA Miniprep Kit was added to the solution and vigorously vortexed 2-3 times, 10 seconds at a time. The solution was then centrifuged at 16000×g, 4ºC for 2 minutes. 800 µL of the supernatant was transferred from the microcentrifuge tubes to a genomic DNA removal column, and the flow through was collected. The flow through was centrifuged at 16,000×g, 4ºC for 30 seconds and diluted with an equivalent volume of ethanol before placing 800 µL of the solution into an RNA purification column. The column was spun at 16,000×g, 4ºC for 30 seconds before the flow through was discarded from the collection column. 500 µL of RNA priming buffer was added to the column and spun at 16,000×g, 4ºC for 30 seconds in a microcentrifuge before discarding the flow through from the collection column. 500 µL of RNA wash buffer was added to the column and spun at 16,000×g, 4ºC for 30 seconds in a microcentrifuge before discarding the flow through from the collection column. The previous step was then repeated to ensure residual ethanol was mostly washed out. The column was then placed in a microcentrifuge tube, and RNA was eluted by using 50 µL of RNase free water.

2.4 NASBA reaction

Samples (12 µL total after enzyme addition) were prepared with master mixes that included the NASBA NT mix (a 6×mixture of NTP's and dNTP's), and the NASBA reaction buffer(a 3X mixture of Tris-HCl, pH 8.5 at 25ºC, $MgCl_2$ KCl, DTT, Dimethyl Sulfoxide). Once the master mix was aliquoted into each sample, the forward and reverse primers were added such that they both have final concentrations of 1.5 µM. The samples were made using RNase-free water. The samples for RNA amplification had 1 µL of total *E.coli* RNA added to them. The no-target control (NTC) samples contained 1 µL of water instead of RNA. At this point, all sample were loaded into the C100 Touch Thermocycler at 65 ºC for 2 minutes to allow for RNA denaturation to enable primer annealing while the samples were cooled down to 41ºC for 10 min. The samples were then removed, and 3 µL NASBA enzyme cocktail (AMV RT, RNase H, T7 RNA polymerase, BSA, and high MW sugar matrix) was added to 9 µL of every sample. At this point, all samples were placed back into the C100 Touch Thermocycler and incubated at 41 ºC for 120 minutes.

2.5 Gel electrophoresis

The agarose gel was cast with Gel Red as the staining dye in order to be able to visualize each band under UV light. The samples containing 1 μL NASBA products or NTC, 6 µL of 2X RNA gel loading dye (ThermoFisher Scientific) and 8 µL of RNase-free water were loaded into the agarose gel (2%) to be run at 100 V for 40 min.

A JPEG image of the gel from the Gel Documentation system was converted to a grayscale image and analyzed for pixel intensity. The pixel intensities for each band of the NASBA samples and their respective no target controls were taken relative to that of the ladder by using PhotoShop by Adobe.

Results and Discussion

An algorithm is a step by step solution at solving a problem and flow charts can be used to visually depict the algorithm. Figure 9 is the flow chart that was followed when designing the program.

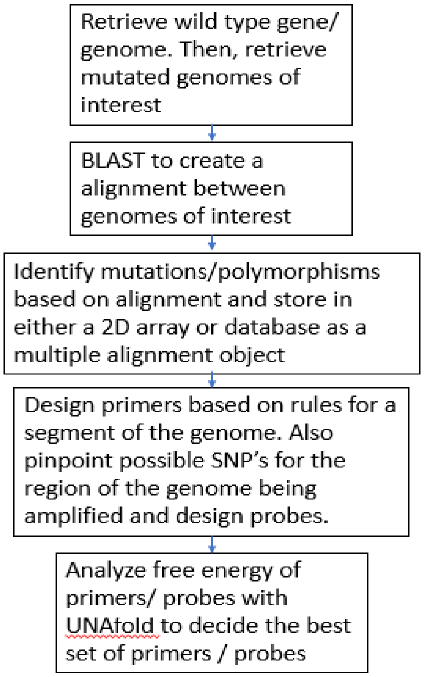
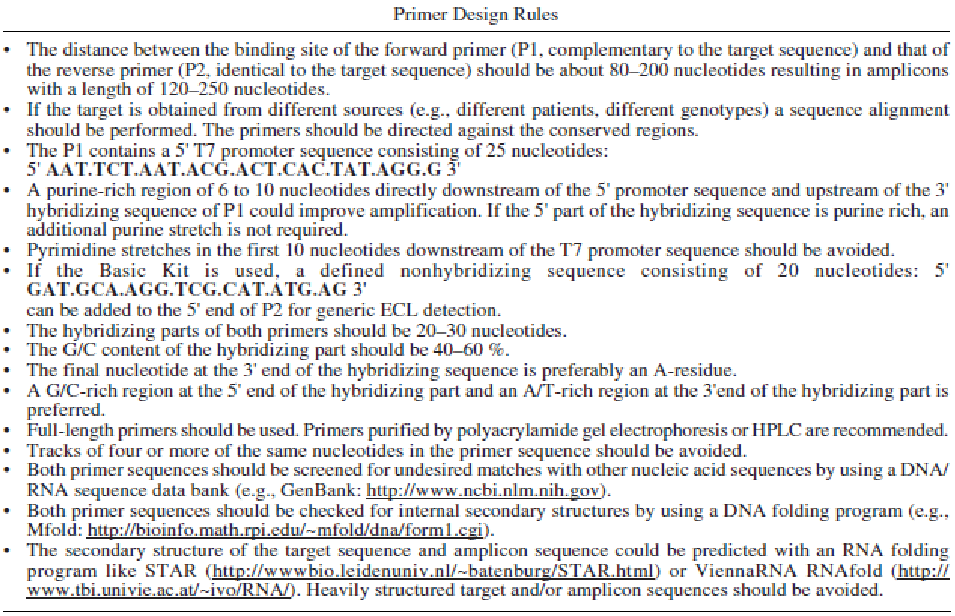

Retrieve wild type gene/ genome. Then, retrieve mutated genomes of interest
BLAST to create a alignment between genomes of interest
Identify mutations/polymorphisms based on alignment and store in either a 2D array or database as a multiple alignment object
Design primers based on rules for a segment of the genome. Also pinpoint possible SNP's for the region of the genome being amplified and design probes.
Analyze free energy of primers/ probes with UNAfold to decide the best set of primers / probes

Figure 9: Flowchart for a NASBA primer design program. The organization of the program is straight forward and does not require any loops.

As described in the flowchart, the project was focused on designing primers to amplify a fragment of an RNA template from a sequence that is retrieved from a database. A second portion of the program, also included in the flowchart, is focused on designing sensors to detect mutations such as single nucleotide polymorphisms (SNP's) or deletions. These mutations can be verified by comparing various mutant genomes against the wild type genome using the BLAST comparison algorithm. BLAST works to align two given sequences by matching three letter similarities between the two and noting any dissimilarities. After aligning the genome and noting the discrepancies, a 2D array was constructed to keep track of every sequence and mutation as a multiple alignment object. Primers can be designed after the user specifies the region of interest in the genome. The rules for primer design can be referenced in Figure 10. The target sequence and primers were tested for the secondary structure with the most positive free energy and other areas of binding within the genome to determine the most energetically favorable primer/analyte complex.



Primer Design Rules

- The distance between the binding site of the forward primer (P1, complementary to the target sequence) and that of the reverse primer (P2, identical to the target sequence) should be about 80–200 nucleotides resulting in amplicons with a length of 120–250 nucleotides.
- If the target is obtained from different sources (e.g., different patients, different genotypes) a sequence alignment should be performed. The primers should be directed against the conserved regions.
- The P1 contains a 5' T7 promoter sequence consisting of 25 nucleotides:
  5' **AAT.TCT.AAT.ACG.ACT.CAC.TAT.AGG.G** 3'
- A purine-rich region of 6 to 10 nucleotides directly downstream of the 5' promoter sequence and upstream of the 3' hybridizing sequence of P1 could improve amplification. If the 5' part of the hybridizing sequence is purine rich, an additional purine stretch is not required.
- Pyrimidine stretches in the first 10 nucleotides downstream of the T7 promoter sequence should be avoided.
- If the Basic Kit is used, a defined nonhybridizing sequence consisting of 20 nucleotides: 5' **GAT.GCA.AGG.TCG.CAT.ATG.AG** 3'
  can be added to the 5' end of P2 for generic ECL detection.
- The hybridizing parts of both primers should be 20–30 nucleotides.
- The G/C content of the hybridizing part should be 40–60 %.
- The final nucleotide at the 3' end of the hybridizing sequence is preferably an A-residue.
- A G/C-rich region at the 5' end of the hybridizing part and an A/T-rich region at the 3'end of the hybridizing part is preferred.
- Full-length primers should be used. Primers purified by polyacrylamide gel electrophoresis or HPLC are recommended.
- Tracks of four or more of the same nucleotides in the primer sequence should be avoided.
- Both primer sequences should be screened for undesired matches with other nucleic acid sequences by using a DNA/RNA sequence data bank (e.g., GenBank: http://www.ncbi.nlm.nih.gov).
- Both primer sequences should be checked for internal secondary structures by using a DNA folding program (e.g., Mfold: http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi).
- The secondary structure of the target sequence and amplicon sequence could be predicted with an RNA folding program like STAR (http://wwwbio.leidenuniv.nl/~batenburg/STAR.html) or ViennaRNA RNAfold (http://www.tbi.univie.ac.at/~ivo/RNA/). Heavily structured target and/or amplicon sequences should be avoided.

Figure 10: The rules for NASBA primer design. Each rule accounts for key aspects of NASBA such as the T7 promoter region of the primer [5].

The program initially retrieves the genomes of the wild type organism and those with mutated genes. These genomes can either be retrieved locally by the user or by using the BioPython Entrez module to access GenBank and retrieve the corresponding genomes. The user should specify the nucleotide numbers for the region of amplification, which should be known prior to using the program. If desired, the user can identify the mutation sites between the aligned target sequences of the wild type and mutant genomes. With the target sequence specified, the rules (Figure 10) can be applied by using straightforward algorithms for each rule of primer design. The program will find the reverse complement of the target sequence in the 5' to 3' position so the reverse primer can be constructed. The primers are checked for proper GC content/weighting as well as the absence of triplets. GC content/weighting is determined by adding up all G and C's in the sequence and dividing them by the number of nucleotides in the primer sequence. Triplets are identified by iteratively looking through each nucleotide in the primer sequences and findings repeats of the same nucleotide. Those that don't meet the above specifications are then removed from the list of possible primers. Reverse primers have a 25 nucleotide T7 promoter region at the beginning of the sequence attached to them and a terminal adenine residue.

After all possible primers are designed, they are energetically analyzed via UNAfold. Any system commands made by the program should be compatible with either UNAfold or Mfold as long as they are stored in the same directory as the script being run. The program will analyze the primers for the least negative Gibbs free energy of the primer's secondary structure, which will be used to find the mol fractions of the species (P1, P2, target, and all combinations of the three) in the system to see which primer set has the greatest P1/P2/target ratio when compared to the rest of the primer sets. BLAST is run on all of the primers against the wild type genome to determine if there are any other possible binding sites. The best primer was then shown along with all other possible primers and their corresponding free energies/ binding sites. The following is an example of the output for the program that would appear in the terminal:

```
TTGCTGACGAGTGGCGGACGGGTGA
free energy:-3.805 kCal
TTGCTGACGAGTGGCGGACGGGTGAGTA
free energy:-3.805 kCal
TTGCTGACGAGTGGCGGACGGGTGAGTAA
free energy: -3.805 kCal
AATTCTAATACGACTCACTAAGGGAGAAGGCTTGCGACGTTATGCGGTATTTA
free energy: -3.928 kCal
AATTCTAATACGACTCACTAAGGGAGAAGGCTTGCGACGTTATGCGGTATTTAGCTA
free energy: -5.929 kCal
forward primers:
['TTGCTGACGAGTGGCGGACGGGTGA', 'TTGCTGACGAGTGGCGGACGGGTGAGTA',
'TTGCTGACGAGTGGCGGACGGGTGAGTAA']
reverse primers:
['AATTCTAATACGACTCACTAAGGGAGAAGGCTTGCGACGTTATGCGGTATTTA',
'AATTCTAATACGACTCACTAAGGGAGAAGGCTTGCGACGTTATGCGGTATTTAGCTA']
```

The ordering of the array the primer lists starts with the “best” primer sequence leading up to the “worst” primer sequence.

Similarly, to how primers are designed, sensors for post-amplification detection can also be constructed and energetically tested to decide the best sequence. The sensor design that was tested for the program is a split DNA based sensor leading to the possibility of either a split aptamer or split deoxyribozyme based sensor. Our lab currently works with split G4 deoxyribozyme[11], cascade deoxyribozyme[12], split 10-23 deoxyribozyme[13], and split dapoxyl aptamer (SDA) derived sensors[10]. In terms of price, limit of detection, and simplicity, SDA is the best option out of the four. When compared to the other probes SDA has a straightforward design and a limit of detection of about 10 nanomolar for NASBA products on a synthetic analyte [7]. Since there will be a focus on post amplification detection of mutation sites, sensors limit of detection values will not need to be scrutinized in order determine which sensor to use due to a large amount of target/amplicon being available.

The lengths of the sensors are taken into account, since longer target-complementary fragments bind stronger to the target, thus ensuring better sensitivity. Shorter sequences allow for the sensor to be SNP-specific. There is a prompt for the user to specify sensor length and mutation of interest. The probes are to be energetically analyzed after they're constructed by using UNAfold to determine the mol fractions of the primer and primer sequence complexes.

Several primer sets for NASBA reaction of a fragment of *E. coli* 16S rRNA (nts 90-190 or nts 45 to 160) were generated using the developed software. Some sequences are listed in Table 1. Experiments were focused on verifying the accuracy of the program's primer design as well as the calculations for determining the free energies of the primer sets. By designing two sets of primers with varying free energies for two different regions of *E.coli* 16S rRNA, one could verify the functionality of the program depending on the presence/ amount of amplification. With regards to free energy, it was predicted that a more positive free energy should correlate to a greater amount of amplification. However, all primer sets developed from the program should yield some amount of amplification.

Table 1: Primers Generated In-Silico For Regions Of *E.coli* 16s rRNA

Amplified region	Forward primer	Reverse primer[a]
nt 90-190	**FP1:** TTGCTGACGAGTGGCGGACGGGTGA free energy:-3.805 kCal **FP2:** TTGCTGACGAGTGGCGGACGGGTGAGTA free energy:-3.805 kCal **FP3:** TTGCTGACGAGTGGCGGACGGGTGAGTAA free energy: -3.805 kCal	**RP1:** AATTCTAATACGACTCACTAAGGGAGAAGGCTTGCGACGTTATGCGGTATTTA free energy: -3.928 kCal **RP2:** AATTCTAATACGACTCACTAAGGGAGAAGGCTTGCGACGTTATGCGGTATTTAGCTA free energy: -5.929 kCal
nt 45-160	**FP4:** GCCTAACACATGCAAGTCGAA free energy:-0.094 kCal **FP5:** GCCTAACACATGCAAGTCGAACGGTA free energy:-0.797 kCal **FP6:** GCCTAACACATGCAAGTCGAACGGTAA free energy:-1.149 kCal **FP7:** GCCTAACACATGCAAGTCGAACGGTAACA free energy:-1.149 kCal	**RP3:** AATTCTAATACGACTCACTAAGGGAGAAGGTTCCAGTAGTTATCCCTCCCA free energy: -6.223 kCal **RP4:** AATTCTAATACGACTCACTAAGGGAGAAGGTTCCAGTAGTTATCCCTCCCATCA free energy: -6.223 kCal **RP5:** AATTCTAATACGACTCACTAAGGGAGAAGGTTCCAGTAGTTATCCCTCCCATCAGGCA free energy: -6.675 kCal

[a] Nucleotides in the reverse primer sequences containing the T7 RNA polymerase promotor sequence are underlined.

Table 2: Combinations of Primers Used for NASBA Trials

	Forward Primer/Reverse Primer Set	*Amplicon Region*
RK1	FP2 and RP1	90 to 190
RK2	FP2 and RP2	
RK3	FP4 and RP3	45 to 160
RK4	FP5 and RP3	

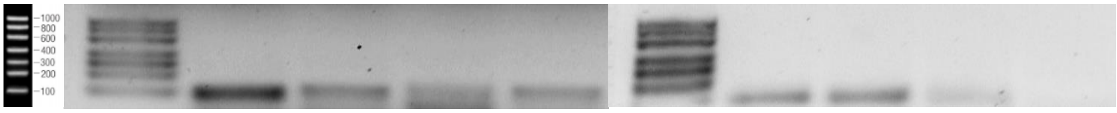


Figure 11: An agarose gel in which the NASBA samples as well as their NTC's were ran in. A low range riboruler was used as the ladder in order to indicate the number of bases within each band.

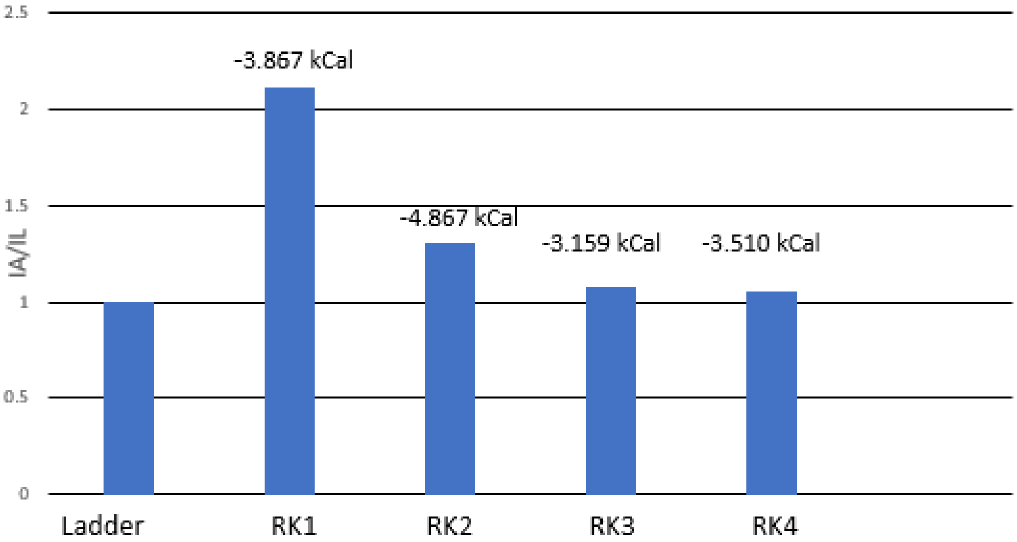


Figure 12: A comparison of the grey scale intensities of the bands of the agarose gel containing samples from table 10 obtained with NASBA. IA/IL is the pixel intensity of the amplicon divided by the pixel intensity of the ladder.

Table 1 indicates the regions of *E.coli* 16S rRNA amplification and every primer generated for those regions. Table 2 shows the different primers used for the primer sets in each NASBA trial. The amplicons of RK1 and RK2, as well as the amplicons of RK3 and RK4, should theoretically have had the same nucleic acid sequence apart from the primer regions. All comparisons of NASBA trials should be between those that have similar amplicon. Table 1 also indicates the free energy of every primer used in the experiment. We hypothesized that a greater free energy (less negative) of a primer's secondary structure, which corresponds to a less stable secondary structure, would allow the primers to bind more readily to the amplicon region to allow for greater amplification efficiency, which would correlate with the free energy value. All the tested primer sets allowed for some degree of amplification, as expected (Figure 11). As it was predicted, RK1 produced a greater amount of amplicon compared to RK2, since the average calculated free energy of the forward and reverse primer's secondary structures was greater (Figure 12). RK3 and RK4 had similar amounts of amplification due to the average calculated free energy ($(\text{Energy}_{FP} + \text{Energy}_{RP})/2$) of the forward and reverse primer's secondary structures being minutely different (Figure 12). The bands identified for the NTC1, NTC2, and NTC3 samples had a visibly larger Rf value than those found in the RK1-RK4 samples and can be noted as primer bands. The relative amount of amplification can be identified by the naked eye when viewing the agarose gel (Figure 11) and by viewing a comparison of the pixel intensities of each sample/corresponding NTC from a grey scale distribution of the gel's image (Figure 12). It can be assumed that there was amplicon of the desired sense

RNA in all samples and our initial hypothesis regarding primer free energy still stands true but further testing should be done considering the number of trials required to definitively state the program has no error.

Limitations

The run time complexity of BLAST is approximately O(kN), where N is the size of the database, and k is the size of the query word. This can pose to be a limitation if the size of the wild type database is too large. A possible solution for this would be to use BLAT which is able to search the query against the database.

Future Work

A graphical user interface (GUI) would make the program much easier to use and probe design for NASBA is still pending. Both of these extensions for the program would build on the established functionality of the program. Probe development for SNP’s is still underway as well. There are also plans to upload the program to GitHub in order to make the program open source.

Conclusion

We have developed a NASBA primer design program that has the capability to detect for possible mutation sites in the region of amplification. The program was tested by developing primers for two different regions of *E.coli* 16S rRNA and conducting NASBA using those primer sets. The primers successfully produced amplicon, and the amount of amplification correlated to the calculated free energy of the secondary of the primer sets. This program is able to tackle a large roadblock in using NASBA for POC scenarios and will hopefully make the process much more easily accessible.

References

1. Eom, S., Wang, J. & Steitz, T. Structure of Taq polymerase with DNA at the polymerase active site. Nature 382, 278–281 (1996). https://doi.org/10.1038/382278a0

2. Catalytic Mechanism. http://academic.brooklyn.cuny.edu/chem/zhuang/Nicolas/catalyti.htm (accessed Mar 11, 2020).

3.McPherson, M. J., & Møller, S. G. (2003). PCR. Oxford: BIOS.

4. Biolabs, N. E. Isothermal Amplification & Strand Displacement. https://www.neb.com/products/pcr-qpcr-and-amplification-technologies/isothermal-amplification-and-strand-displacement/isothermal-amplification-and-strand-displacement (accessed Mar 23, 2020).

5.Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N., & Hase, T. (2000). Loop-mediated isothermal amplification of DNA. *Nucleic acids research*, *28*(12), E63. https://doi.org/10.1093/nar/28.12.e63

6.Zhao, Y., Li, Q., Chen, F., & Fan, C. (2015). Isothermal Amplification of Nucleic Acids. Chemical Review. doi:10.1021

7. Deiman, B., Aarle, P. V., & Sillekens, P. (2002). Characteristics and Applications of Nucleic Acid Sequence-Based Amplification (NASBA). Molecular Biotechnology,20(2), 163-180. doi:10.1385/mb:20:2:163

8. Chang, C., Chen, C., Wei, S., Lu, H., Liang, Y., & Lin, C. (2012). Diagnostic Devices for Isothermal Nucleic Acid Amplification. Sensors,12(6), 8319-8337. doi:10.3390/s120608319

9. Xia, T., Santalucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., . . . Turner, D. H. (1998). Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson−Crick Base Pairs†. Biochemistry,37(42), 14719-14735. doi:10.1021/bi9809425

10. Kikuchi, N., Reed, A., Gerasimova, Y. V., & Kolpashchikov, D. M. (2019). Split Dapoxyl Aptamer for Sequence-Selective Analysis of Nucleic Acid Sequence Based Amplification Amplicons. Analytical Chemistry,91(4), 2667-2671. doi:10.1021/acs.analchem.8b03964

11. Reed, A. J., Connelly, R. P., Williams, A., Tran, M., Shim, B. S., Choe, H., & Gerasimova, Y. V. (2019). Label-Free Pathogen Detection by a Deoxyribozyme Cascade with Visual Signal Readout. Sensors and actuators. B, Chemical, 282, 945–951. https://doi.org/10.1016/j.snb.2018.11.147

12. Gerasimova, Y. V.; Cornett, E. M.; Edwards, E.; Su, X.; Rohde, K. H.; Kolpashchikov, D. M. Deoxyribozyme Cascade for Visual Detection of Bacterial RNA. ChemBioChem 2013, 14 (16), 2087–2090. doi: 10.1002/cbic.201300471

13. Tian, Y.; Mao, C. DNAzyme Amplification of Molecular Beacon Signal. Talanta 2005, 67 (3), 532–537. doi: 10.1016/j.talanta.2005.06.044