Republic of the Philippines

Department of Science and Technology

Philippine Science High School - Eastern Visayas Campus

Pawing, Palo, Leyte 6501

PrimerGenTool- A Bioinformatics Platform to Train Site-Directed Mutagenesis Primer Designing

In partial fulfillment of the 4th Quarter requirements for S&T Research 3

Jefferson Louis G. Amores

Lyzter John S. Deuda

Rudy F. Pilande II

April 25, 2021

Chapter I

THE PROBLEM AND ITS SETTING

I. Background of the Study

Site-Directed Mutagenesis (SDM) is a valuable tool in the field of genetics and molecular biology as it allowed molecular researchers investigate the functional effects of specific sequence changes, therefore, enabling them to understand genetic and molecular mechanisms present in biological systems. SDM is the process of making mutations at a specific region in a genetic sequence using primers.

A primer is a short set of single-stranded DNA nucleotide sequence used in the polymerase chain reaction (PCR) technique and for a multitude of other experimental processes (Loftus, n.d.). The designing of optimal primers is therefore a key skill used alongside SDM. However, the task of designing primers is rather difficult and time-consuming with so many variables to consider. Because of issues such as these in biology, specialized tools for molecular data management, analysis, and generation of specific outputs based on these data have grown in numbers.

Bioinformatics is the application of statistics and information theory to genetic data. It employs computer science, statistics, and molecular biology to automatically and efficiently store, interpret, and utilize massive quantities of molecular data, making it an ideal tool for numerous molecular and genetic research such as SDM.

With the increasing appreciation of SDM due to newfound discoveries of its various applications, the need for experts in SDM and developing bioinformatics tools that aid users in performing SDM, particularly in generating mutagenic PCR primers, have exponentially increased. This need was further highlighted with the complexity of primer designing and with the majority of the few remaining commercial primer design

tools being outdated, functionally limited, overcomplicated for most users, and purely for experimentation purposes only (Guo et al., 2020).

Therefore, the researchers of this study aimed to create a convenient and easy-to-use open-sourced bioinformatics tool that teach and allow SDM primer designing to train more experts in SDM and to obtain maximum efficiency when conducting SDM-based studies, respectively.

II. Statement of the Problem

a. Main Objective

The main objective of the study is to develop a text and point-and-click-based bioinformatics platform prototype in JAVA that allows users to generate customized primers for site-directed mutagenesis and learn about site-directed mutagenesis primer design with an easy-to-use interface.

b. Specific Objectives

1. To develop a software prototype using JAVA that accepts inputs of either DNA or protein sequences.
2. To integrate a function that allows users to adjust the settings or conditions of their inputted sequences such as: 5' GC content, 3' GC content, melting temperature range, annealing temperature range, and GC termination for both forward and reverse primers into the prototype.
3. To integrate built-in dictionaries, guides, tutorials, and references regarding Site-directed Mutagenesis and Primer Design to the prototype.
4. To integrate a function that outputs generated mutagenic primers.

5. To evaluate the software output accuracy by comparing prototype outputs with *PrimerX* outputs.

III. Significance of the Topic

Site-directed mutagenesis is a fundamental tool in molecular biology. SDM is an artificial technique that uses oligonucleotide primers to mutate or alter a nucleotide sequence of a gene at a specified location, in contrast to general mutagenesis that employs mutagenic agents or compounds to randomly alter nucleotide sequences (El Gewely, Fenton, Buvang, & Xu, 2005).

Although SDM is most widely used to probe the structure and biological activity of nucleic acids and proteins, it can also be utilized in changing amino acid composition, destroying transcription factor binding sites, creating fusion proteins, and introducing or removing restriction enzyme recognition sites to aid cloning. More recently, however, such precise alterations are also being developed for in vivo gene or genome modifications. These techniques are revolutionizing our understanding of the genetic and molecular mechanisms in biological systems, which could lead to the development of new enzymes, therapeutics as well as improved agricultural applications.

Furthermore, even if an ample amount of research was invested in developing similar programs, most of these tools are now out of service, with the ones that are still functional being outdated, accessible only to those who have spent a significant amount of time studying primer design, and are for experimentations only. In line with this, the researchers created a SDM primer design tool that allowed and guided users in generating mutagenic primers specifically for SDM and conducting targeted mutations in a specific nucleotide sequence location with

ease, unlike other pre-existing tools that are limited to designing traditional primers for mutations at undefined sites.

Moreover, simulations and tutorial software have assisted teachers in educating students new lessons. They have also provided a platform through which students could learn a lesson at their own pace via a virtual experience. This educational software can lead to a deeper appreciation of Bioinformatics tools and its possible applications in education.

Overall, all aforementioned reasons have made an educational bioinformatics tool that teaches and allows the creation of primers and perform SDM an indispensable tool in biology, genetics, education, and the field of bioinformatics itself. Researchers and experts in the field of molecular biology, genetic engineering, and bioinformatics as well as instructors, teachers, and students in biology would therefore greatly benefit from the study.

IV. Scope and Limitations

The study was a one-year project dedicated to the creation of an educational bioinformatic tool that allows and teaches Site-Directed Mutagenesis primer designing. The research was conducted by the researchers from their respective homes due to the quarantine protocols and the COVID-19 pandemic. The researchers utilized JAVA for developing the software and its client-side improvements. The tool allowed users to import DNA or protein sequences, edit the parameters of their primer such as 5” GC content, 3” GC content, melting temperature range, annealing temperature range, and GC termination for both forward and reverse primers, and exported these mutagenic primers as plain text.

Furthermore, the tool would contain a built-in dictionary, wiki, documentation, and tutorial.

However, the exported primers are not ensured to work in actual follow-up experiments as the scoring and evaluating of generated primers for efficiency when used in actual experiments was beyond the scope of the study. Additionally, the accuracy of the outputs was based on comparisons with the outputs of a similar pre-existing primer design tool, PrimerX, which also claimed that their outputs where not ensured to be efficient when used in actual experiments.

Moreover, the quality of the primers depended on the capability or knowledge of the user on primer design and Site-Directed Mutagenesis as the software only have the QuikChange™ primer design protocol built-in. The user will have to manually set the primer parameters if they seek to use Site-Directed Mutagenesis protocols other than QuikChange™ Protocol. Furthermore, the Primer Design function had a limit of 8,000 base, while the Primer Characterization function had a limit of 90 bases.

Lastly, the software will only be available in Windows computers with at least a Java Runtime Environment 1.7.0.

V. Definition of Terms

3' flanking region (bp) - This refers to the length of the portion of a primer that is to the right of the mutated region, when the sequence is read 5' to 3'.

5' flanking region (bp) - This refers to the length of the portion of a primer that is to the left of the mutated region, when the sequence is read 5' to 3'.

Annealing temperature range- the temperature where the DNA primers can attach to the template DNA.

Bioinformatics- Bioinformatics is a subdiscipline of biology and computer science concerned with the acquisition, storage, analysis, and dissemination of biological data, usually molecular biology data.

COVID-19- is a highly infectious disease caused by a newly discovered coronavirus that has plunge the world into a pandemic in 2020.

Dimerize- when two primers complement and hybridize with each other.

DNA- deoxyribonucleic acid is the hereditary material in humans and almost all other organisms.

Enzymes- are biological molecules (typically proteins) that significantly speed up the rate of virtually all of the chemical reactions that take place within cells.

FASTA- in bioinformatics and biochemistry, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid sequences, in which nucleotides or amino acids are represented using single-letter codes.

GC content- In molecular biology and genetics, GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine (G) or cytosine (C)

HTML- HyperText Markup Language is the most basic building block of the Web. It defines the meaning and structure of web content.

Hybridize- a phenomenon in which single-stranded deoxyribonucleic acid or ribonucleic acid molecules anneal to complementary DNA or RNA

JAVA- Java is a class-based, object-oriented programming language that is designed to have as few implementation dependencies as possible.

Locus- is the specific physical location of a gene or other DNA sequence on a chromosome, like a genetic street address.

Mutations- is a change in a DNA sequence.

Nucleotide sequences- is a succession of bases signified by a series of a set of five different letters that indicate the order of nucleotides forming alleles within a DNA (using GACT) or RNA (GACU) molecule

Oligonucleotides- another name for primers

Primers- is a short nucleic acid sequence that provides a starting point for DNA synthesis.

Primer length (bp) - This refers to the total primer length, including the desired mutation.

Site- Directed Mutagenesis (SDM)- is a molecular biology method that is used to make specific and intentional changes to the DNA sequence of a gene and any gene products.

Windows- is a group of several proprietary graphical operating system families, all of which are developed and marketed by Microsoft.

Chapter II

REVIEW OF RELATED LITERATURE

Site-Directed Mutagenesis

Site-directed mutagenesis is the process of making mutations at known locations in a genetic sequence. SDM has become a crucial tool in the field of genetics and molecular biology. The method of making site-specific alterations of nucleotide sequences aids molecular researchers in investigating the functional effects of specific sequence changes therefore allowing them to understand genetic and molecular mechanisms present in biological systems such as the formation of proteins, heredity, and gene expression (Tee & Wong, 2013; Yang, Li, Du, Liu, 2017). A study by Tee & Wong (2013) tackled the role of mutagenesis methods in the rapidly developing field of genetics and evolutionary studies. Tee and his team identified the wide variety of applications that mutagenesis, specially SDM, could offer by surveying multiple mutagenesis methods and comparing the technical requirements and mutational spectra that they offer. With research effort now being invested in bioeconomy, SDM will continue to play a key role in biotechnology and genetics; SDM will remain as an indispensible tool in many laboratories as the importance of creating high quality mutant libraries is further accentuated by its extended applications in varied research fields.

SDM was first introduced in a 1974 study of Charles Weismmann about Qbeta phage study. Weissmann and his team inserted nucleotide analogs into specific locations of the Qbeta virion genome and discovered how the virus could rapidly mutate and become resistant against drugs. The researchers were able to study more comprehensively the structural and functional properties of proteins which would lead

to the discovery of reverse genetics or the process that is used to help understand the function of a gene by analysing the phenotypic effects caused by genetically engineering specific nucleic acid sequences within the gene (Weissmann, 2012). Mutagenesis have been developed over the years with it having numerous variants such as Random Mutagenesis, Transposon-based random mutagenesis, Altered target sequence length in random mutagenesis, and Focused Mutagenesis or Site-Directed Mutagenesis (Tee & Wong, 2013). The numerous variations of Mutagenesis have allowed for the method to be a flexible and effective tool capable of being used in numerous fields of study.

A study by Wu et al. on a rapid and efficient mutagenesis protocol found out that a site-directed mutagenesis protocol that involves ligating two amplified DNA fragments into vectors can potentially facilitate high-output genetic engineering and structure-function analyses consequentially greatly benefitting the field of molecular biological research (Wu et al., 2013). Additionally, a method for multi-site-directed mutagenesis based on homologous recombination have been developed by Liang and his team in 2012. SDM was used for the simultaneous introduction of up to three mutations in a plasmid DNA via homologous recombination. The approach was found to be compatible with a variety of mutations including degenerate codons. In contrast to other studies, this method utilizes the same set of reagents for both single- and multi-site mutagenesis assays, minimizes the required protocol steps, and produces high mutagenesis efficiencies (Liang, Peng, Li, Peterson, & Katzen, 2012). Moreover, targeted mutagenesis has been used to mitigate codon bias which resulted in the capability to change rare to more common codons by co-expressing genes that encode the tRNA. This allowed the microbial production and molecular engineering of industrial enzymes (Yang, Li, Du, Liu, 2017). SDM can therefore be utilized in

phenotypic improvement of single protein-encoding genes, viral evolution, metagenomics, functional study of proteins, and, overall, genomic engineering therefore allowing the ability to design primers at any desired manipulation of a genetic sequence (Karnik, Karnik, & Grefen, 2013).

Among the numerous Site-Directed Mutagenesis kits, Stratagene's QuikChange™ is one of the most useful, easiest-to-use, and preferred among users. QuikChange™ utilizes a high-fidelity DNA polymerase such as KOD hot start DNA polymerase, Pfu DNA polymerase, or Phusion® high-fidelity DNA polymerase to amplify the whole plasmid with complementary primer pairs, carrying the desired mutation in the form of mismatches to the original plasmid. This allows for minimal unwanted mutations (Zeng, F., Zhang, S., Hao, Z., 2018). This prompted the researchers to select the QuikChange™ as the primary basis for an ideal Site-Directed Mutagenesis protocols.

Furthermore, Polymerase Chain Reaction (PCR), also called "molecular photocopying," PCR is a fast and inexpensive technique used to copy large quantities of small segments of DNA. Significant amounts of a sample of DNA are necessary for molecular and genetic analyses therefore, studies of isolated pieces of DNA are nearly impossible without PCR amplification (NHGRI, 2019). This therefore made the researchers focus on PCR-based SDM instead of other SDM techniques.

Overall, SDM has been proven as an invaluable tool with numerous applications with several approaches to this technique having been published. However, these protocols are labor intensive or difficult and the success of the oligonucleotide-based mutagenesis is often dependent on the different factors to be optimized for the primer design (Agilent Technologies, 2015).

Primer Design

A primer is a short, single-stranded DNA sequence used in the polymerase chain reaction (PCR) technique and for a multitude of other experimental processes such as detection, cloning, sequencing, and many more. A pair of primers is used to hybridize with the sample DNA to define the region of the DNA that will be amplified for further analysis making it a crucial tool in SDM (Alfandary, 2015; El Gewely, Fenton, Buvang, & Xu, 2005; Loftus, n.d.). Primers that are not specific enough for a particular DNA template result in mispriming or fail in amplifying the target locus therefore protocols for optimal DNA Primer Design must be strictly observed. Thus, primer length limitations, temperature boundaries or melting point (primer and product), determining conserved regions, avoiding primer-dimers, and minding GC content must be properly determined (Biocompare: The Buyer's Guide for Life Scientists, 2013; El Gewely et al., 2005). Therefore, primer design can fundamentally determine the results of SDM studies (Chuang, Cheng, Yang, 2012).

Primer design will depend on the type of sequence change being desired. Although most of the basic PCR primer design rules hold, others may have to be adjusted as the template sequence places some constraints on the primer sequence.

Primers are short stretches of oligonucleotides that are synthesized in various lengths that amplify the target regions of a nucleotide sequence. The shorter the primers, the more efficiently they can anneal to a target DNA. Primers that are longer than 25 work better when troubleshooting closely related species, for instance as it allows for higher specificity and room for adding restriction enzyme sites to the primer end when cloning. For ideal amplification, the primer length must be 12 to 25 bases long. The greater the expected alterations in sequence, the longer the complementary region of

the oligonucleotide primer should be. Deletions or insertions require primers with complementary sequences of 12–15 base pairs or more on either side of the target (Biocompare: The Buyer's Guide for Life Scientists, 2013; El Gewely et al., 2005).

The Melting temperature (Tm) difference between primers must be limited to 2–3 degrees Celsius and the primer with the lower Tm will dictate the annealing temperature used in the reaction (El Gewely et al., 2005). The ideal melting point should be around 50 degrees or between 53 and 65 degrees Celsius as better results are obtained from primers with Tm greater than 45 degrees Celsius (Hall-Wheeler, n.d.). In some cases, a higher melting point may be required particularly in the QuikChangeTM method. An accurate estimate of Tm for primers between 18 and 24 base pairs in length can be calculated using the equation Tm 5 2 (AT) + 4 (GC) (El Gewely et al., 2005). The optimal annealing temperature (Ta) is determined empirically based on the Tm. The Ta is typically lower than the primers' Tm by approximately 5 degrees C to 10 degrees C. As a rule of thumb, the Tm of the primers can be estimated by adding 2 degrees C for each A or T and 4 degrees C for each G or C. (Biocompare: The Buyer's Guide for Life Scientists, 2013).

Whenever possible, the GC content of the primers should be between 35 and 65% with 50-55% being ideal. If the primers G-C content is less than 50%, the length of the primer may need to be increased to maintain the proper Tm. Moreover, mutagenic primers with polyC or polyG sequences can promote nonspecific annealing while long stretches of polyT or polyA can break from the template. Providing a clamp at the 30 - terminal with two or three G or C nucleotides promotes an increase in the chances of a successful reaction. the presence of G and C bases at the 3′ end of the primer—the GC clamp—helps promote correct binding at the 3′ end because of the stronger hydrogen bonding of G and C bases. Sequences containing more than three repeats of sequences

of G or C in sequence are usually avoided in the first five bases from the 3′ end of the primer because of the higher probability of primer-dimer formation or the primers amplifying themselves however, due to the experimental design of the QuikChange™, the primers would eventually dimerize. Therefore, the 3' ends where the primer should have no more than 3–4 bases of homology in order to have the least dimerization (Agilent Technologies, 2015; Biocompare: The Buyer's Guide for Life Scientists, 2013; El Gewely et al., 2005; Hall-Wheeler, n.d.)

A primer design with the ideal settings is crucial as the primer design fundamentally affects the results. The settings of the parameters must be adjusted accordingly based on the experiment at hand. Due to the constraints that site-directed mutagenesis imposes on the location of the primers, meeting all these parameters is not always possible. Numerous researches are available to help guide researchers in designing their own primers. However, the most optimal method to follow when designing primers is to make use of the many available primer design tools (Chuang, Cheng, Yang, 2012; Hall-Wheeler, n.d).

Bioinformatics Tools

Bioinformatics is an interdisciplinary study that involves computer science, molecular biology, genetics, statistics, and more. The objective of Bioinformatics is to develop theoretical and computational models and technologies for solving problems based on large quantities of data, usually molecular data. Bioinformatics is basically the use of information technologies in collecting, storing, and produce data conveniently, developing technological tools to analyze biological data, and using these tools interpret the results from a biological perspective (Cummings et al., 2010; Kovarik et al., 2013).

The discovery of the importance and various applications that site-directed mutagenesis offers have increased the need to develop bioinformatics tools that aid users in carrying out SMD particularly in generating mutagenic or modified PCR primers. Bioinformatic tools, specifically primer design software tools, are crucial tools in SDM as such tools can aid in the selection and formation of the optimal mutagenic primers which is imperative as primer design fundamentally determines the results of SDM studies (Chuang, Cheng, Yang, 2012).

Primer design tools can assist in PCR primer design for new and experienced users alike. These tools reduce the cost and time involved in experimentation by lowering the chances of failed experimentations through simulations. However, no two primer programs will ever have the same exact generated primers, even when basic parameters are equivalently set (El Gewely et al., 2005).

There used to be numerous primer design software available on the internet. These were URPD, Mutation Maker, CODEHOP, GeneWalker, NetPrimer, Primer3, The Primer Generator, Web Primer, MutScreener, and PrimerX. However, only a handful of these tools remain accessible to the public.

Primer3 is an online bioinformatics tool for PCR primer generation from nucleotide sequences. It was the first ever PCR primer design software introduced in the 1990s and was the basis of almost all other primer design tools.

yoUR Primer Design(URPD) is a web-based specific product primer design tool, that combines the NCBI Reference Sequences (RefSeq), UCSC In-Silico PCR, memetic algorithm (MA) and genetic algorithm (GA) primer design methods to obtain specific primer sets (Chuang, et al., 2012).

Mutation Maker is an open source mutagenic oligo design software for large-scale protein engineering experiments. Mutation Maker is not only specifically tailored to multi-site random and directed mutagenesis protocols, but also pioneers bespoke mutagenic oligo design for de novo gene synthesis workflows (Hiraga et al., 2020). Similar to Primer3, MutationMaker is built for more advance, high throughput protein engineering processes.

NetPrimer combines algorithms with a web-based interface, allowing the user to analyze primers. NetPrimer analyzes inputted primers and outputs their Melting Temperature (Tm), Primer Secondary Structures, and other primer properties similar to a function of PrimerGenTool. However, PrimerGenTool can design mutagenic primers itself in addition to primer characterization. PrimerGenTool is therefore a more flexible and diverse product when it comes to performing Site-Directed Mutagenesis experiments.

Lastly, PrimerX is a web-based program written to automate the design of mutagenic PCR primers for site-directed mutagenesis.

All the aforementioned software were the only commercially available platforms left. Although almost all these available software share functions similar to what PrimerGenTool offers, some have limited functionality, while others have significantly complex functions than what the researchers will be developing.

However, the ability to perform highly-advanced functions risk the price of being too complicated to use, thus making it difficult for infrequent and inexperienced users or researchers that only seek for a fast and convenient primer design tool that can be accessed anytime.

The researchers used PrimerX as the basis for their prototype's output accuracy as PrimerX was developed specifically for SDM primer design. Although, PrimerX is not a published study or the developers have no published documentation regarding PrimerX, there are numerous published studies and online research forums that reviewed primer design software tools that recommended PrimerX (Guo et al., 2020). Furthermore, PrimerX was the software that was utilized for numerous patented methodologies or protocols for primer design, protein engineering, and for performing research on molecular mechanisms (Widjojoatmodjo et al., 2020).

Bioinformatics Tools Used in Education

Over the last two decades, bioinformatics has been an integral part of life science. Many research areas now need bioinformatics skills; however, bioinformatics education is not well incorporated at the undergraduate level due to a shortage of bioinformatics faculty in life sciences education. The supply of trained researchers is normally insufficient to meet the demand. One study by de-Diego et al. aims to create a self-contained educational platform that hosts a variety of resources and databases for bioinformatics research and allows for supervised training. All necessary applications and databases are installed locally on the system, minimizing the system's reliance on the internet. Furthermore, the eBioKit's architecture has proven to be an excellent blend of portability and efficiency. Databases, applications, and tutorials are the three forms of content in the eBioKit. Human and nonhuman model organisms are also supported in research. To support any bioinformatics discipline, tools for sequencing homology search, protein structure prediction, next-generation sequencing (NGS) data analysis, functional annotation, and genome-wide association studies (GWASs) are included, among others. There are two simple access modes for the eBioKit. Using a web browser is the most popular method. Students can also link via Secure Shell (ssh) on a terminal

using a command-line interface. The eBioKit provides a number of tools that make certain common tasks in service and user management easier. These administration tools, which can be run individually as command-line programs, are compiled in a Java framework called "eBioKit Control Panel," which has a user-friendly interface on both the desktop and the command-line. As a result, the eBioKit has become an important part of training and study at a number of universities and organizations, including the Pan African Bioinformatics Network (H3ABioNet) as part of the Human Heredity and Health in Africa (H3Africa) initiative, the Southern Africa Network for Biosciences (SAnBio) initiative, the Biosciences Eastern and Central Africa (BecA) center, and the Biosciences Eastern and Central Africa (BecA) center, and the International Glossina Genome Initiative.

A study by Kovarik and his team in 2013 investigated the effects of their bioinformatics teacher professional development model and bioinformatics curricula on cognitive traits, specifically awareness, engagement, self-efficacy, and relevance in high school teachers and STEM students. Their program included the best practices in adult education and diverse resources to empower teachers to integrate STEM career information into their classrooms. Kovarik and his team conducted Pre–post surveys which showed significant growth among teachers in their preparation to teach the curricula and to infuse career awareness into their classes. The 289 Introductory Unit students showed significant gains in awareness, relevance, and self-efficacy, while the 41 Advanced Unit students showed gains in all target cognitive areas. The results of their study were further explored in the context of recommendations for other bioinformatics tools or programs that aim to increase student interest in STEM careers (Kovarik et al., 2013).

Furthermore, a study by Reisdorph, and his team in 2013 aimed to identify the best methods in teaching advanced technologies and concepts such as genomics, proteomics, and bioinformatics. Reisdorph hypothesized that a comprehensive hands-on training course in genomics, proteomics and informatics in a coherent, experimentally-based framework can help address the issue of the lack of proper training for experts in advanced biological fields such as Genomics and Proteomics. Reisdorph developed course content with theoretical and practical or "hands-on" experience. The course included comprehensive lectures and laboratories in mass spectrometry and genomics, extensive hands-on experience with instrumentation and software, video demonstrations, workshops, online sessions, and keynote speakers. The educational course received positive reviews from participants and had a substantial impact on grant writing and review, manuscript submissions and publications (Reisdorph, 2013). The researchers of the current study incorporated Reisdorph's formula for teaching biological concepts into their software.

The results of these studies led to the researchers of the current paper, to program a bioinformatics tool that can help in teaching biological concepts, specifically Site-Directed Mutagenesis, in classrooms.

Due to time constraints, the researchers of the current paper would not be evaluating the effectiveness of their bioinformatics tool through testing respondent cognitive traits, awareness, engagement, self-efficacy, and relevance. The researchers would employ an eLearning Tool Evaluation rubric from Lauren M. Anstey & Gavan P.L. Watson from the Centre for Teaching and Learning, Western University and guidelines from Gregg B. Jackson, Associate Professor and Coordinator of the Education Policy Program, George Washington University which were (Alexander, B.,

2021). Transforming Digital Learning and Assessment: A Guide to Available and Emerging Practices and Building Institutional Consensus. Stylus Publishing, LLC.).

CHAPTER III

METHODOLOGY

This chapter covered the following parts: research design, the methods, and the statistical treatment to be used for the research.

RESEARCH DESIGN

The research used a developmental type of research design as the researchers aimed to create a user-friendly text-based bioinformatics platform in Java that allowed users to generate customized primers for site-directed mutagenesis experiments and design workflows with an easy-to-use interface and learn about primer design. This type of research design allowed the researchers answer the main problem and sub-problems.

PREPARATION

Preparation of Hardware and Software

Windows laptops that can run JAVA were utilized for the creation of the software program. The researchers made use of the Java programming language in creating the platform since Java is platform-independent, meaning programs written in Java can be ran on several different types of computers with the Java Runtime Environment or the JRE installed (Lowe & Burd, n.d.). The researchers created the program using NetBeans, specifically version 8.2, since it is the already available Integrated Development Environment or the IDE. The Java Development Kit or the JDK, which includes the JRE, is already installed in the researchers' laptops. The JDK also functions as an interpreter and a compiler.

Creating the Flowchart

A flowchart is necessary because it will serve as an aid for the researchers to visualize and understand a process more clearly. A flowchart is also useful for the debugging stage. Following is the flowchart of the program.

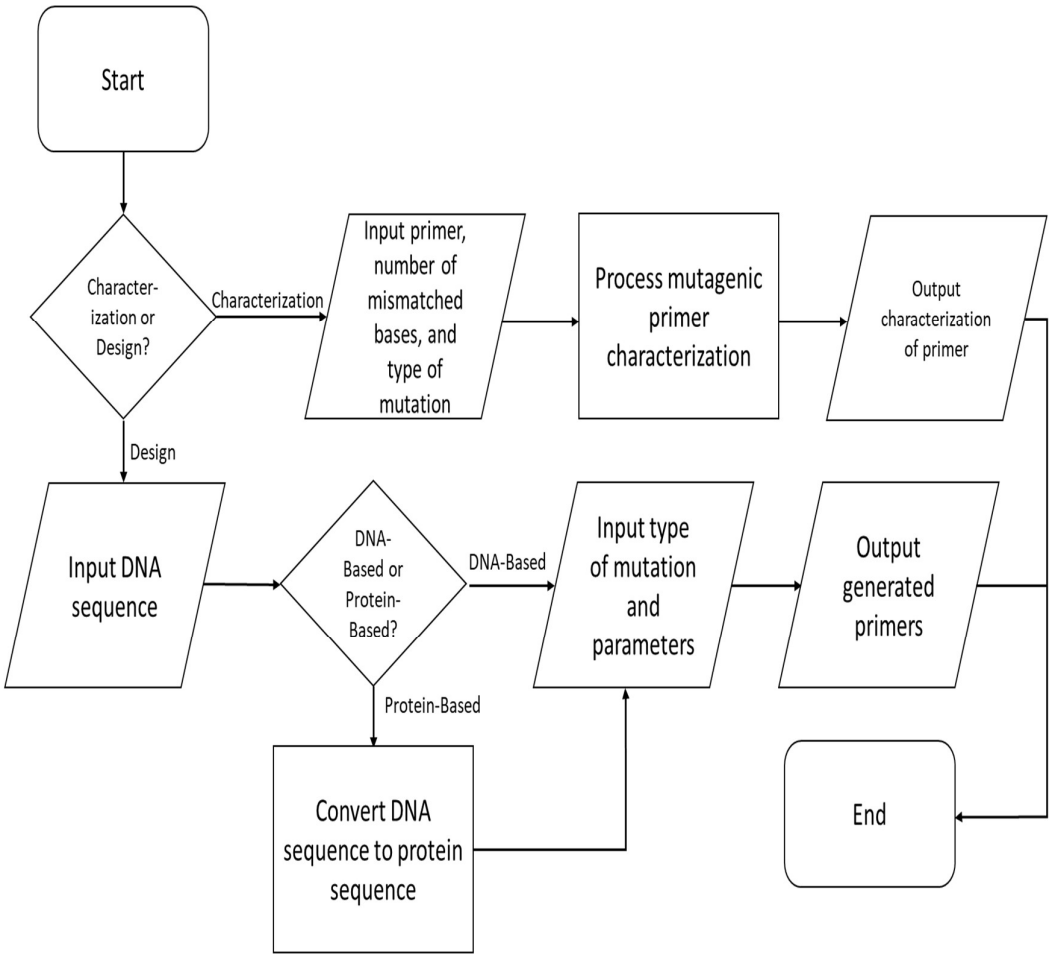

Figure 1. Rough Flowchart of the Program

DEVELOPMENT

From the flowchart above, what will follow are the wireframes and descriptions for each part of the program. Hovering on any of the buttons or elements in the program will show a pop up of its description to help the user to easily maneuver the application.

Main Menu add iba na tabs for learning

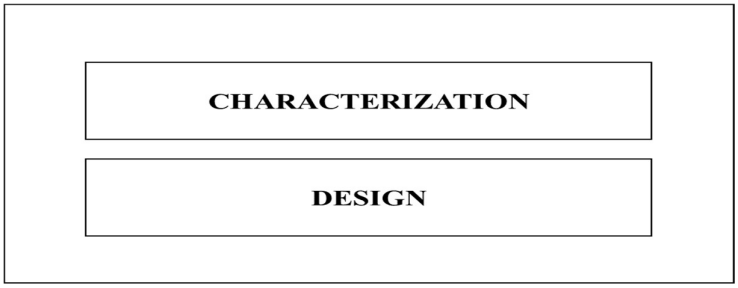

Figure 2. Wireframe of the Program: Main Menu

The program will prompt a menu choosing between two buttons which are Characterization and Design.

Characterization's function is to characterize a user-inputted primer based on user-inputted number of mismatched bases and type of mutation – substitution, insertion, or deletion.

Design's function is to generate primers based on user-inputted DNA sequence, user-inputted DNA sequence with mutations, and the parameters for the primers the user wish to generate.

Characterization

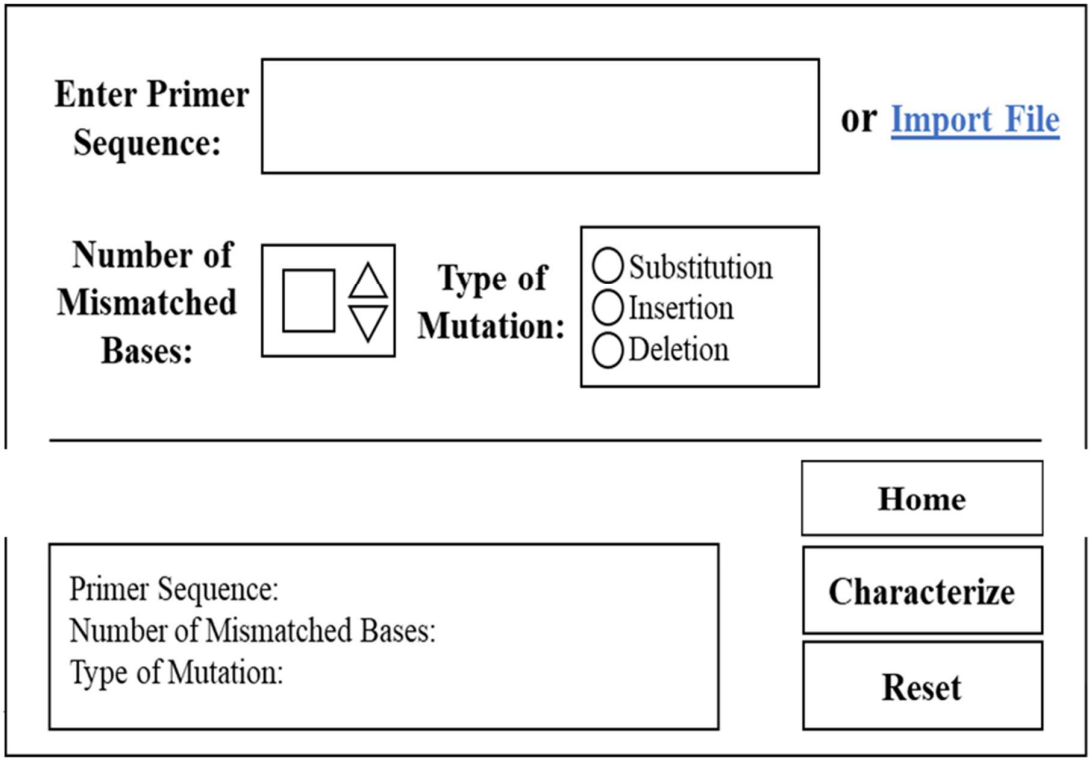

Figure 3. Wireframe of the Program: Characterization (Input)

Should the user choose Characterization, the program will prompt for the primer sequence, number of mismatched bases, and type of mutation. The user can choose to manually type or paste the primer sequence, or an "Import File" option is available if the user wishes to import a file for the input. File format for the input is limited to FASTA. Any character from the inputted primer sequence that is not G, T, A, or C will be trimmed. Below the input section is the review section, where the user can recheck the inputted values. Following are the "Characterize" and "Reset" buttons. Clicking on "Characterize" will proceed to the next page, while clicking on "Reset" will clear the inputted values.

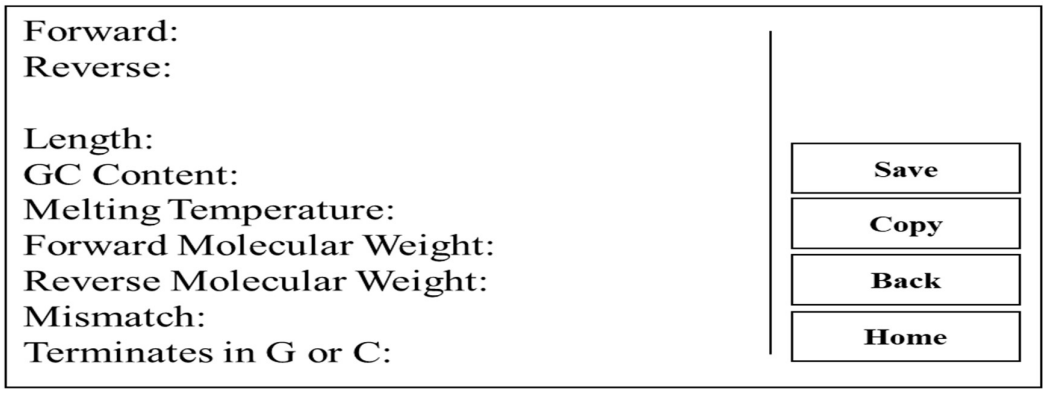

Figure 4. Wireframe of the Program: Characterization (Output)

Next will display the output – the characterization of the primer. The output will contain the forward sequence of the primer, the reverse sequence of the primer, and the characterization. The characterization will contain the length of the primer sequence; the GC content (in percentage); the melting temperature (in degrees Celsius), in which, depending on the inputted type of mutation, will follow the following formulas: for substitutions, $T_m = 81.5 + 0.41(\%GC) – 675/N – (\%mismatch)$, and for deletions and insertions, $T_m = 81.5 + 0.41(\%GC) – 675/N$ wherein T_m is the melting temperature of the primer, %GC is the GC content, and %mismatch is the ratio of the number of mismatched bases and the length of the primer sequence; the molecular weights of the forward and reverse sequences (both in kilograms per mole), in which, will follow the following formula: MW = (313.21 x nA) + (304.20 x nT) + (329.21 x nG) + (289.19 x nC) + 18:02 - 80:00 wherein MW is the molecular weight, and nA, nT, nG, and nC are the number of A, T, C, and G base pairs, respectively; mismatched bases (in percentage), which is the ratio of the number of mismatched bases; and the length of the primer sequence; and check if the sequence terminates in G or C. Following are the “Save”, “Copy”, “Back”, and “Home” buttons. Clicking on “Save” will prompt the user for the filename and save the output in the “.txt” file. Clicking on “Copy” will copy the output to the clipboard. Clicking on “Back” will go back to the previous page. Clicking on “Home” will go back to the main menu.

Design

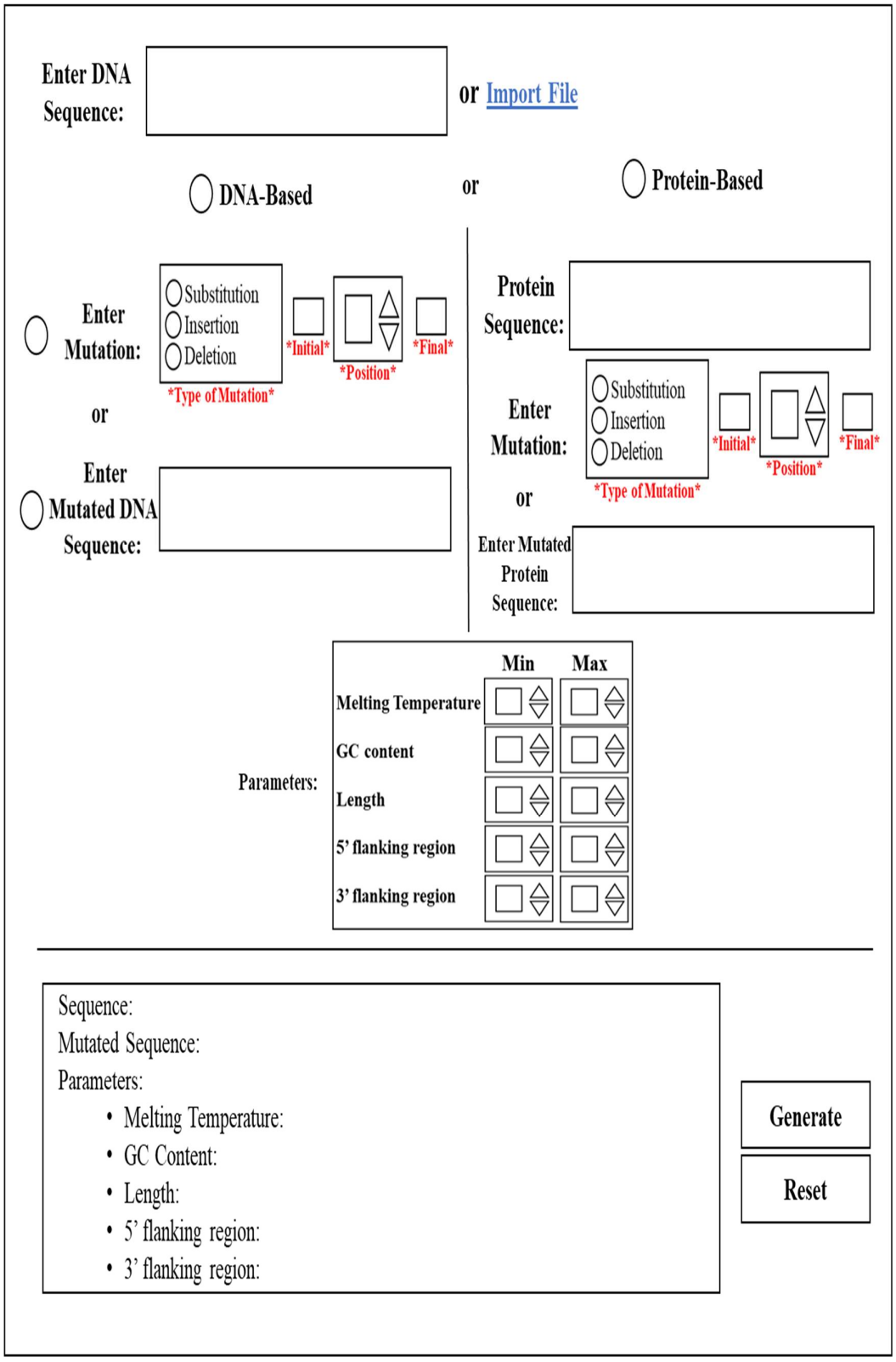

Figure 5. Wireframe of the Program - Design (Input)

In the main menu of the program, the user was asked to choose between Characterization and Design. Should the user choose Design, the program will prompt the user to input a DNA sequence. The user can choose to manually type or paste the primer sequence, or an “Import File” option is available if the user wishes to import a file for the input. File format for the input is limited to FASTA. Any character from the inputted sequence that is not G, T, A, or C will be trimmed.

Next, the user will be prompted to choose between “DNA-Based” or “Protein-Based”. Choosing “Protein-Based” will translate the user-inputted DNA sequence into a protein sequence. Next will be desired mutation of the user. For the DNA-based mutation, it will be by nucleotide bases, but for the Protein-based mutation, it will be by codons. The user has the option to enter the mutation (what type of mutation, what base or codon, what position, and for substitution, what base or codon to substitute). Following these will be the parameters for the primers the user wishes to generate. There will be the default parameters, which will be based on *QuikChange™ Site-Directed Mutagenesis Kit by Stratagene®*, wherein the range for the melting point is 75°C - 85°C, the range for the GC content is 40% - 60%, the range for the length is 25-45 base pairs, and the range for the 5’ and 3’ flanking regions is 11-21 base pairs. The user has the option to edit these parameters. This includes the minimum and maximum melting temperature (in degrees Celsius), the minimum and maximum GC content (in percentage), the minimum and maximum length of the base pairs, the minimum and maximum 5’ flanking region, and the minimum and maximum 3’ flanking region.

Below the input section is the review section, where the user can recheck the inputted values. Following are the “Generate” and “Reset” buttons. Clicking on “Generate” will proceed to the next page, while clicking on “Reset” will clear the inputted values.

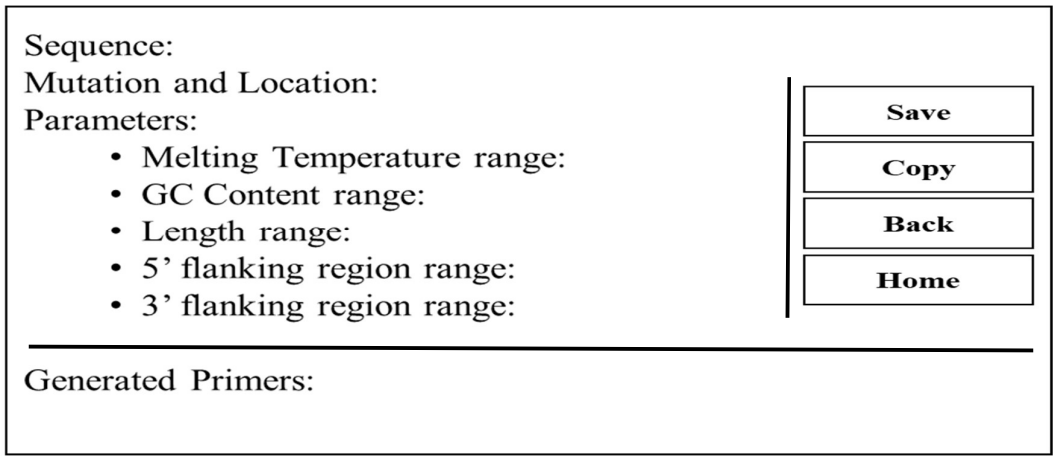

Figure 6. Wireframe of the Program - Design (Output)

Based on the user's inputted values and information, the program will generate primers that satisfy the given parameters of the user-inputted sequence. Generation of these primers is discussed in chapters one and two of the research paper. The program will then output the results, with options "Save", "Copy", "Back", and "Home". Clicking on "Home" will go back to the main menu. Clicking on "Copy" will copy the output to the clipboard. Clicking on "Back" will go back to the previous page. Clicking on "Home" will go back to the main menu.

The researchers also included sections for a built-in dictionary, wiki, documentation, and a tutorial.

EVALUATION OF SYSTEM

The researchers will evaluate and see to it that the program is fully-functional, displays no errors, and outputs the correct information. Evaluation is important to demonstrate the success or progress of the research (Evaluation: What is it and why do it?, n.d).

Beta-Testing

The researchers tested the prototype of the program to see if there are any bugs and errors present. The program was tested for its functionality and efficiency according

to essential parameters and guidelines. Specifically, the researchers checked if the program produced the correct output:

- Site-directed mutagenesis
- Protein sequence-based design
- DNA sequence-based design
- Instructions/Help function
- Selection of mutagenic primer types
- Primer sequence
- Primer properties

Moreover, the researchers tested the accuracy of their software's outputs by comparing the PrimerGenTools outputs with PrimerX outputs. In order to avoid bias in data input, the specific inputs that were given to the program to test its output were provided by 10 random Philippine Science High School – Eastern Visayas Campus Grade 12 students.

The results will be subjected for conclusions and discussions.

CHAPTER 4

RESULTS AND DISCUSSIONS

Results for Primer Characterization can be seen in Tables 1 to 10. Out of all the 10 test cases, there is no PrimerX and PrimerGenTool outputs that are 100% similar. Overall, the outputs had differences of values with at most two decimal places.

Data Analysis and Research Findings

PRIMER CHARACTERIZATION

INPUT					PrimerX	PrimerGenTool	Results
	Primer Sequence	Number of Mismatched Bases	Type of Mutation				
1	GTACTCAGTCAGAGTCGA	4	Insertion	Forward	GTACTCAGTCAGAGTCGA	GTACTCAGTCAGAGTCGA	Yes
				Reverse	TCGACTCTGACTGAGTAC	TCGACTCTGACTGAGTAC	Yes
				Length (base pairs)	18	18	Yes
				GC Content (%)	50	50	Yes
				Melting Temperature (°C)	53.8	53.79	-0.01
				Forward Molecular Weight (Da)	5523.68	5523.68	Yes
				Reverse Molecular Weight (Da)	5474.65	5474.65	Yes
				Mismatch (%)	22.22	22.22	Yes
				Ends in G or C (Yes/No)	No	No	Yes

Table 1. Test Case 1: GTACTCAGTCAGAGTCGA

					PrimerX	PrimerGenTool	
2	GTACCCTAGATGCTATTCGGATATCGAT CGATTGCAGCATC	41	Substitution	Forward	GTACCCTAGATGCTATTCGGATATCGATCGATTGCAGCATC	GTACCCTAGATGCTATTCGGATATCGATCGATTGCAGCATC	Yes
				Reverse	GATGCTGCAATCGATCGATATCCGAATAGCATCTAGGGTAC	GATGCTGCAATCGATCGATATCCGAATAGCATCTAGGGTAC	Yes
				Length (base pairs)	41	41	Yes
				GC Content (%)	46.34	46.34	Yes
				Melting Temperature (°C)	59.9	59.65	-0.25
				Forward Molecular Weight (Da)	12575.31	12575.3	-0.01
				Reverse Molecular Weight (Da)	12633.35	12633.3	-0.05
				Mismatch (%)	24.39	24.4	0.01

Table 2. Test Case 2: GTACCCTAGATGCTATTCGGATATCGATCGATTGCAGCATC

					PrimerX	PrimerGenTool
3	CATCTGTTGATGGA	0	Insertion	Number of mismatched bases is less than 1.		Number of mismatched bases is less than 1.

Table 3. Test Case 3: CATCTGTTGATGGA

					PrimerX	PrimerGenTool		
4	GTAC TACG CTAG GAAC	3	Insertion	Forward	GTACTACGCTAGGAAC	GTACTACGCTAGGAAC		Yes
				Reverse	GTTCCTAGCGTAGTAC	GTTCCTAGCGTAGTAC		Yes
				Length (base pairs)	16	16		Yes
				GC Content (%)	50	50		Yes
				Melting Temperature (°C)	50.1	50.08		-0.02
				Forward Molecular Weight (Da)	4890.27	4890.27		Yes
				Reverse Molecular Weight (Da)	4872.25	4872.25		Yes
				Mismatch (%)	18.75	18.8		0.05
				Ends in G or C (Yes/No)	Yes	Yes		Yes

Table 4. Test Case 4: GTAC TACG CTAG GAAC

					PrimerX	PrimerGenTool		
5	GTAC GTAC	2	Substitution	Forward	GTACGTAC	GTACGTAC		Yes
				Reverse	GTACGTAC	GTACGTAC		Yes
				Length (base pairs)	8	8		Yes
				GC Content (%)	50	50		Yes
				Melting Temperature (°C)	-7.4	-7.375		0.025
				Forward Molecular Weight (Da)	2409.64	2409.64		Yes
				Reverse Molecular Weight (Da)	2409.64	2409.64		Yes
				Mismatch (%)	25	25		Yes
				Ends in G or C (Yes/No)	Yes	Yes		Yes

Table 5. Test Case 5: GTAC GTAC

					PrimerX	PrimerGenTool		
6	GCGTGTAC	5	Substitution	Forward	GCGTGTAC	GCGTGTAC		Yes
				Reverse	GTACACGC	GTACACGC		Yes
				Length (base pairs)	8	8		Yes
				GC Content (%)	62.5	62.5		Yes
				Melting Temperature (°C)	-40	-39.75		0.25
				Forward Molecular Weight (Da)	2425.64	2425.64		Yes
				Reverse Molecular Weight (Da)	2394.63	2394.63		Yes
				Mismatch (%)	62.5	62.5		Yes
				Ends in G or C (Yes/No)	Yes	Yes		Yes

Table 6: Test Case 6: GCGTGTAC

					PrimerX	PrimerGenTool		
7	GTCGATCGATCGATGTCA	5	Deletion	Forward	GTCGATCGATCGATGTCA	GTCGATCGATCGATGTCA		Yes
				Reverse	TGACATCGATCGATCGAC	TGACATCGATCGATCGAC		Yes
				Length (base pairs)	18	18		Yes
				GC Content (%)	50	50		Yes
				Melting Temperature (°C)	72.7	72.65		-0.05
				Forward Molecular Weight (Da)	5514.67	5514.67		Yes
				Reverse Molecular Weight (Da)	5483.66	5483.66		Yes
				Mismatch (%)	27.78	27.8		0.02
				Ends in G or C (Yes/No)	No	No		Yes

Table 7: Test Case 7: GTCGATCGATCGATGTCA

						PrimerX	PrimerGenTool				
8	AGCTAGGATCGATATGCATCTATCAGCTACGATCGATCAGCTACATCAGTA	3	Insertion		Forward	AGCTAGGATCGATATGCATCTATCAGCTACGATCGATCAGCTACATCAGTA	AGCTAGGATCGATATGCATCTATCAGCTACGATCGATCAGCTACATCAGTA		Yes	
					Reverse	TACTGATGTAGCTGATCGATCGTAGCTGATAGATGCATATCGATCCTAGCT	TACTGATGTAGCTGATCGATCGTAGCTGATAGATGCATATCGATCCTAGCT		Yes	
					Length (base pairs)	51	51		Yes	
					GC Content (%)	43.14	43.14		Yes	
					Melting Temperature (°C)	85.1	86.69		1.59	
					Forward Molecular Weight (Da)	15666.36	15666.4		0.04	
					Reverse Molecular Weight (Da)	15719.37	15719.4		0.03	
					Mismatch (%)	5.88	5.88		Yes	
					Ends in G or C (Yes/No)	No	No		Yes	

Table 8: Test Case 8: AGCTAGGATCGATATGCATCTATCAGCTACGATCGATCAGCTACATCAGTA

						PrimerX	PrimerGenTool		
					Forward	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA		Yes
9	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	13	Deletion		Reverse	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT		Yes
					Length (base pairs)	55	55		Yes
					GC Content (%)	0	0		Yes
					Melting Temperature (°C)	71.6	71.57		-0.03
					Forward Molecular Weight (Da)	17164.57	17164.6		0.03
					Reverse Molecular Weight (Da)	16669.02	16669		-0.02
					Mismatch (%)	23.64	23.6		-0.04
					Ends in G or C (Yes/No)	No	No		Yes

Table 9: Test Case 9: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

					PrimerX	PrimerGenTool	
10	TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG	8	Deletion	Forward	TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG	TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG	Yes
				Reverse	CACACACACACACACACACACACACACACACA	CACACACACACACACACACACACACACACACA	Yes
				Length (base pairs)	40	40	Yes
				GC Content (%)	50	50	Yes
				Melting Temperature (°C)	87.9	87.94	0.04
				Forward Molecular Weight (Da)	12606.22	12606.2	-0.02
				Reverse Molecular Weight (Da)	11986.02	11986	-0.02
				Mismatch (%)	20	20	Yes
				Ends in G or C (Yes/No)	No	No	Yes

Table 10: Test Case 10: TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG

Tables 1 to 10 showcases the inputs of the 10 PSHS-EVC respondents, and their corresponding outputs in PrimerX and PrimerGenTool. The last column of the table displays whether the outputs are similar. The last column also displays the difference between the outputs.

Functionality

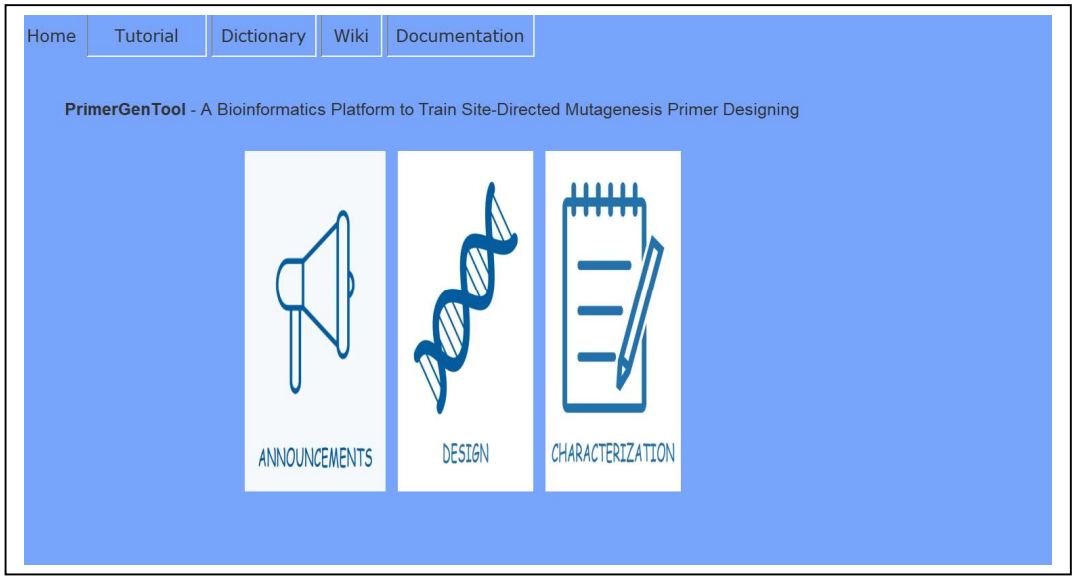

Figure 7. Main Menu

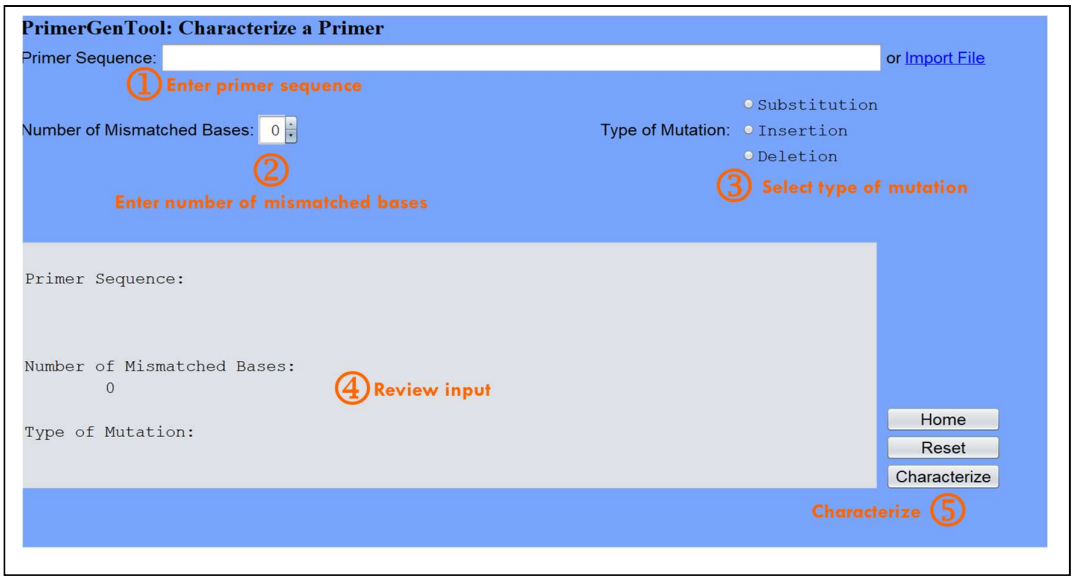

Figure 8. Tutorial

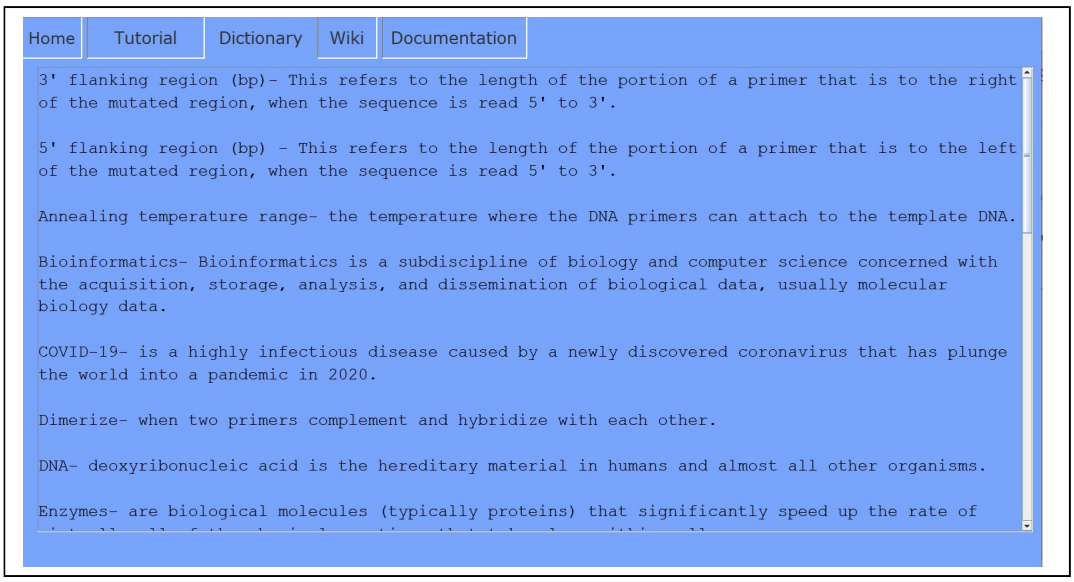

Figure 9. Dictionary

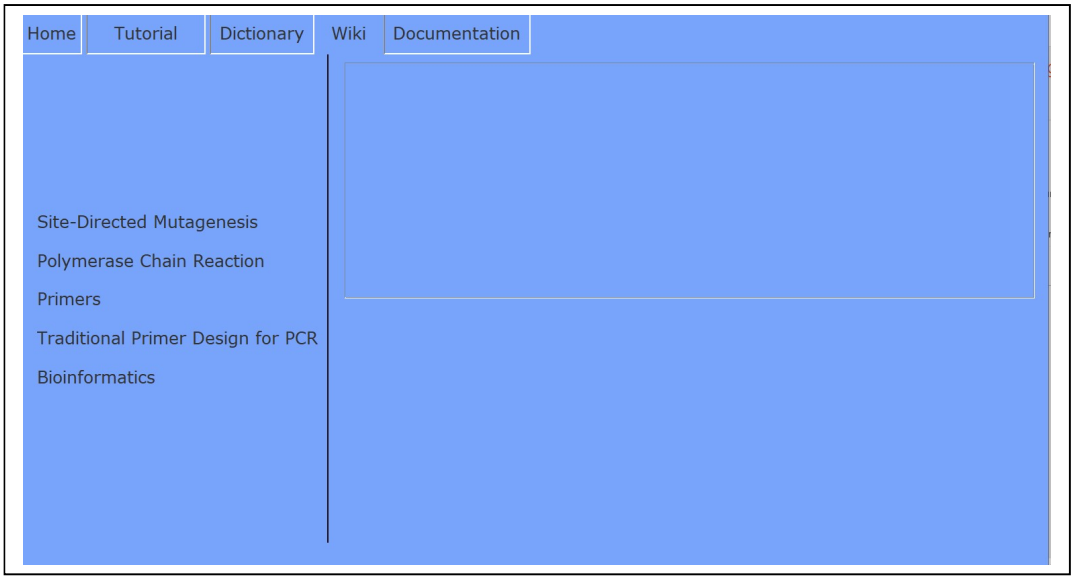

Figure 10. Wiki

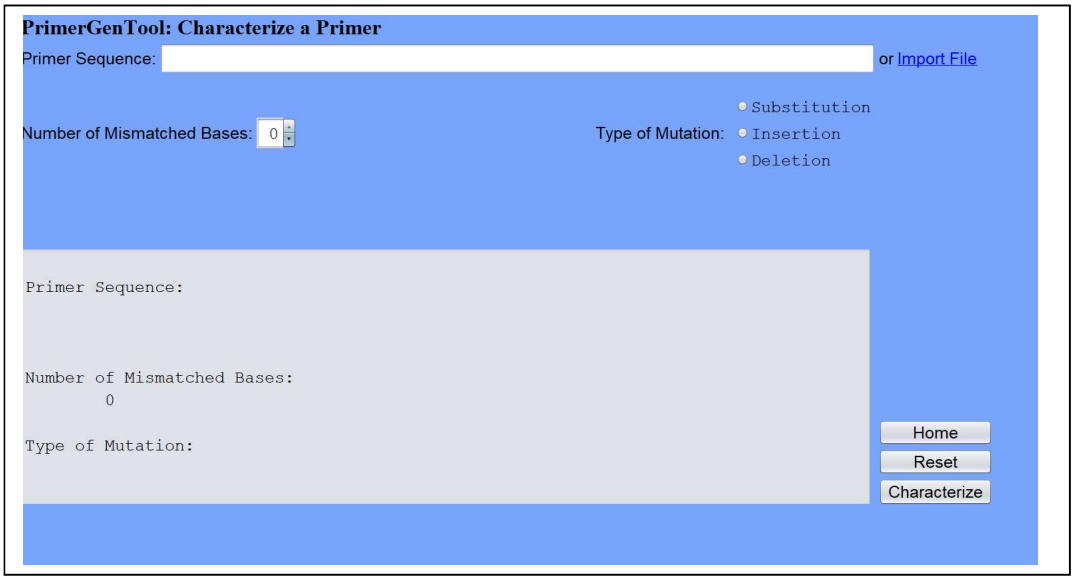

Figure 11. Characterization: Input

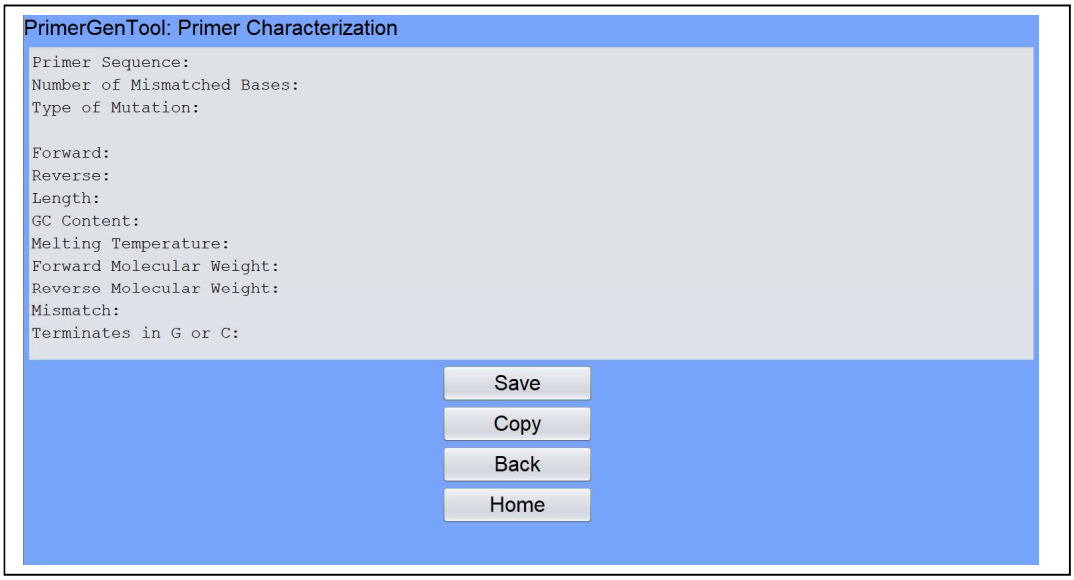

Figure 12: Characterization: Output

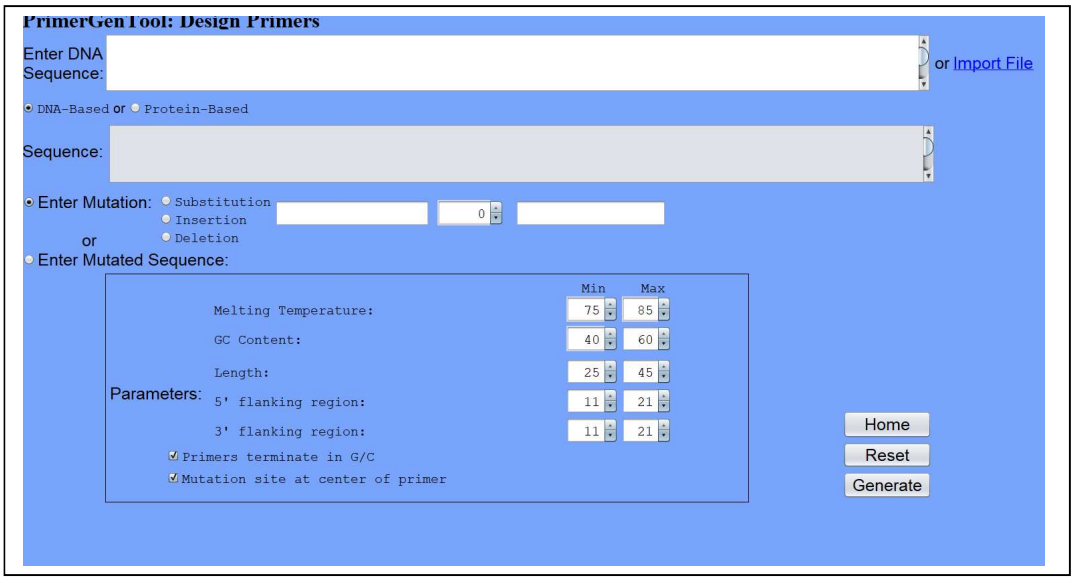

Figure 13: Design: Input

Figure 14: Design Output

Figures 7 to 14 showcases the finished prototype of the study.

Discussion of Findings

Results suggest that PrimerGenTool was able to meet the software objectives.

For the primer characterization, it was impossible to achieve a complete 100% similarity. The reasons for which were not tackled in detail in the study however, the researchers found studies that suggest that this may be due to the two software being compared (PrimerX and PrimerGenTool) were built using different programming languages (web development software and JAVA, respectively) and therefore have a slight difference in computing processes, especially in the estimation or rounding-off process. Furthermore, the researchers had different algorithms in tackling the problem. Moreover, the developers of PrimerX set a disclaimer stating that the outputs of PrimerX cannot be identified as accurate and efficient, when used in experiments, as evaluating that aspect was beyond their scopes and limitations.

CHAPTER 5

Summary, Conclusion and Recommendation

In this study, the researchers developed a text and point-and-click- based bioinformatics platform prototype in JAVA that allowed users to generate customized primers for site-directed mutagenesis and learn about site-directed mutagenesis primer design.

PrimerGenTool was able to meet the software and research objectives. The tool accepted inputs of either DNA or protein sequences, allowed users to adjust the settings or conditions of their inputted sequences, had a built-in dictionary, tutorial, and references regarding Site-directed Mutagenesis and Primer Design, outputted the generated mutagenic primers.

However, as expected, the tool was not able to meet 100% accuracy with regards to primer design and primer characterization when compared with the outputs of PrimerX. The reasons for which were not tackled in detail in the study however, the researchers found studies that suggest that this may be due to the two software being compared were built using different programming languages and algorithms.

In conclusion, the tool has achieved its goals. The bioinformatics functions of the tool include primer characterization and primer design, while the educational features include having a built-in dictionary, tutorial, documentations, and references regarding site-directed mutagenesis, primer design, bioinformatics, and other related topics. The application was also developed using JAVA, with a functional point-and-click and text-based User Interface.

For future studies, the researchers recommend increasing the number of test cases to be compared with the outputs of PrimerX for primer design and primer characterization and let Bioinformatics and Biotechnology experts evaluate the software when evaluating the efficiency and accuracy of the outputs. It is ideal if experts on the relevant fields, conduct

traditional site-directed mutagenesis and primer design, and compare the traditional outputs with the software's. Future studies may also utilize Python and connect their software to the internet to allow more complex functions such as expression systems easier access to third-party services to make processes smoother and easier.

Furthermore, future researchers should conduct an experiment such as implementing the software in a biology class for a significant amount of time and conduct pre-post surveys with the participants. Additionally, utilize research-backed ergonomics and psychology in designing the user interface. This is to ensure that the educational aspect of the software is evaluated properly with statistical-significance and that the software is an effective educational tool.

REFERENCES

Agilent Technologies. (2015). QuikChange II site-directed mutagenesis kit instruction manual.

Alexander, B. (2021). Transforming Digital Learning and Assessment: A Guide to Available and Emerging Practices and Building Institutional Consensus. Stylus Publishing, LLC.

Alfandary, B. R. (2015). Primer Design: Tips for an Efficient Process. Genome Compiler Corporation. Retrieved from http://www.genomecompiler.com/tips-for-efficient-primer-design/ on November 15, 2020.

Biocompare: The Buyer's Guide for Life Scientists. (2013). Retrieved from https://www.biocompare.com/Bench-Tips/133581-Primers-by-Design-Tips-for-Optimal-DNA-Primer-Design

Chuang, L.-Y., Cheng, Y.-H., & Yang, C.-H. (2012). URPD: a specific product primer design tool. BMC Research Notes, 5(1), 306. https://doi.org/10.1186/1756-0500-5-306

Cummings, M. P., & Temple, G. G. (2010). Broader incorporation of bioinformatics in education: opportunities and challenges. Briefings in bioinformatics, 11(6), 537-543.

El Gewely, R., Fenton, C., Buvang, E., & Xu, H. (2005). Mutagenesis: Site-specific. Wiley and Sons. doi: 10.1038/npg.els.0003842

Evaluation: What is it and why do it? (n.d.). Retrieved from http://meera.snre.umich.edu/evaluation-what-it-and-why-do-it

Guo, J., Starr, D., & Guo, H. (2020). Classification and review of free PCR primer design

software. *Bioinformatics*.

Hall-Wheeler, C. (n.d.). Primer Design Considerations | Genomics | University of Nevada, Las Vegas. Retrieved November 18, 2020, from https://www.unlv.edu/genomics/equipment-services/primer-design-considerations

Hiraga, K., Mejzlik, P., Marcisin, M., Vostrosablin, N., Gromek, A., Arnold, J., … Bitton, D. A. (2020). Mutation Maker, An Open Source Oligo Design Software for Mutagenesis and De Novo Gene Synthesis Experiments. BioRxiv, 2020.06.26.171819. https://doi.org/10.1101/2020.06.26.171819

Karnik, A., Karnik, R., & Grefen, C. (2013). SDM-assist software to design site-directed mutagenesis primers introducing “silent” restriction sites. BMC Bioinformatics, 14 , 105. doi: 10.1186/1471-2105-14-105

Kovarik, D. N., Patterson, D. G., Cohen, C., Sanders, E. A., Peterson, K. A., Porter, S. G., & Chowning, J. T. (2013). Bioinformatics education in high school: implications for promoting science, technology, engineering, and mathematics careers. CBE—Life Sciences Education, 12(3), 441-459.

Liang, X., Peng, L., Li, K., Peterson, T., & Katzen, F. (2012). A method for multi-sitedirected mutagenesis based on homologous recombination. Analytical Biochemistry, 427 (1), 99 - 101. doi: https://doi.org/10.1016/j.ab.2012.05.002

Loftus, S. (n.d.). Primer. [Genome.Gov](http://genome.gov/). Retrieved November 18, 2020, from https://www.genome.gov/genetics-glossary/Primer

Lowe, D., & Burd, B. (n.d.). What Is Java, and Why Is It So Great? Retrieved from

https://www.dummies.com

NHGRI. (2019, March 9). Polymerase Chain Reaction (PCR) Fact Sheet. Genome.Gov. https://www.genome.gov/about-genomics/fact-sheets/Polymerase-Chain-Reaction-Fact-Sheet#:%7E:text=Sometimes%20called%20%22molecular%20photocopying%2C%22,copy%20%2D%20small%20segments%20of%20DNA.

Reisdorph, N., Stearman, R., Kechris, K., Phang, T. L., Reisdorph, R., Prenni, J., ... & Geraci, M. (2013). Hands-on workshops as an effective means of learning advanced technologies including genomics, proteomics and bioinformatics. Genomics, proteomics & bioinformatics, 11(6), 368-377.

Tee, K. L., & Wong, T. S. (2013). Polishing the craft of genetic diversity creation in directed evolution. Biotechnology Advances, 31 (8), 1707–1721. doi: 10.1016/j.biotechadv.2013.08.021

Weissmann, C. (2012). The end of the road. Prion, 6(2), 97–104. https://doi.org/10.4161/pri.19778

Widjojoatmodjo, M. N., Godeaux, O., Williams, K. L., Callendret, B. C. S., & Sadoff, J. C. (2020). U.S. Patent Application No. 16/642,082.).

Wu, D., Guo, X., Lu, J., Sun, X., Li, F., Chen, Y., & Xiao, D. (2013). A rapid and efficient one-step site-directed deletion, insertion, and substitution mutagenesis protocol. Analytical Biochemistry, 434 (2), 254 - 258. doi: https://doi.org/10.1016/j.ab.2012.11.028

Yang, H., Li, J., Du, G., & Liu, L. (2017). Microbial Production and Molecular Engineering of Industrial Enzymes. Biotechnology of Microbial Enzymes, 151-165. doi:10.1016/b978-0-12-803725-6.00006-6

Zeng, F., Zhang, S., Hao, Z. et al. Efficient strategy for introducing large and multiple changes in plasmid DNA. Sci Rep 8, 1714 (2018). https://doi.org/10.1038/s41598-018-20169-8