An *in silico* Approach for the Identification of Clinically Significant SNPs in a Drug Metabolizing Gene CYP2C9 and Development of Tetra-primer ARMS PCR Based Genotyping Method for CYP2C9*3 (rs1057910)

By

Rifaquat Ahmed

Student ID: 16136027

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of Bachelor of Science in Biotechnology

Department of Mathematics and Natural Sciences

BRAC University

August 2020

© 2020, BRAC University

All Rights Reserved.

Declaration

It is hereby declared that

1.The thesis submitted is my original work while completing a degree at Brac University.

2.The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3.The thesis does not contain material which has been accepted or submitted for any other degree or diploma at a university or other institution.

4.I have acknowledged all of the main sources of help.

Student's Full Name & Signature:

Rifaquat Ahmed

ID 16136027

Approval

The thesis/project titled **"An *in silico* Approach for the Identification of Clinically Significant SNPs in a Drug Metabolizing Gene CYP2C9 and Development of Tetra-primer ARMS PCR Based Genotyping Method for CYP2C9*3 (rs1057910)"** submitted by

1. Rifaquat Ahmed (ID 16136027)

of Spring, 2016 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Biotechnology on "date to be given" of August 2020.

Examining Committee:

Supervisor:
(Member)

M. Mahboob Hossain, PhD
Professor, Department of Mathematics and Natural Sciences
BRAC University

Co-Supervisor:
(Member)

Md. Mahmudul Hasan Akash
Lecturer, Department of Genetic Engineering and Biotechnology
Dhaka University

Program Coordinator:
(Member)

Iftekhar Bin Naser, PhD
Assistant Professor, Department of Mathematics and Natural Sciences
BRAC University

Departmental Head:
(Chair)

A F M Yusuf Haider, PhD
Professor and Chairperson, Department of Mathematics and Natural Sciences
BRAC University

Dedicated to

my parents and my teachers

Acknowledgment

First of all, I would like to express my deepest gratitude to the Almighty to have allowed and enabled me to perform this thesis work. I would also like to thank my parents for their unconditional devotion and encouragement. They have been a constant source of inspiration and without this, the thesis would not have been possible.

I convey my sincere gratitude to my supervisor Dr. Md. Mahboob Hossain,Professor, Department of Mathematics and Natural Sciences, BRAC University for his guidance about this internship program.

I would also like to offer my most sincere gratitude to Md. Mahmudul Hasan Akash, Lecturer, Department of Genetic Engineering and Biotechnology, the University of Dhakafor his careful and precious guidance during the thesis work. He was sincere from the beginning till the end. He provided me insight and expertise in every step of this work. The support he provided me till the end of my thesis even after leaving the institution cannot be forgottenand I shall always be grateful for it.

I would also like to extend my thanks to Ms.Tabassum Rahman Sunfi, Lecturer, Md. Nazrul Islam, Lab officer, and AsmaBinte Afzal, Lab officer for their wise words and guidance throughout my thesis work.

Lastly, my heartfelt gratitude goes to my friends and well-wishers, SamihaAshreen, ZareenNawar, Sujoy Sarker, Nusrat Jalal Annesha, Mj Kenneth, and others for always encouraging me all the time which meant a lot.

In every phase of the thesis, the careful supervision, guidance,and encouragement of these respected people allowed me to successfully finish my thesis and also complete this thesis report. I consider this an extremely precious experience to which helped me gain knowledge and helped me polish my skills which I will use in the best possible way in the upcoming future ahead.

Rifaquat Ahmed

ID: 16136027

Abstract

The role of SNP is an important factor related to drug metabolism. The presence or absence of SNP can significantly alter the effect of the same amount of drug inside the body leading to various unwanted side-effects which can be harmful at times. The drug warfarin is one such drug that has a different effect based on the SNPs present inside the CYP2C9 gene metabolizing the drug. Various SNPs can be present at different positions in the CYP2C9 gene among which some of them can be in the active site which has a significant impact on drug metabolism. However, not all SNPs can have an effect on drug metabolism. As a result, bioinformatics tools were used in the study to identify the clinically significant SNPs from the CYP2C9 gene. Homology based modeling was used to model 3D proteins containing candidate SNPs which were docked with the ligand to be compared with positive control. The results showed data that could be used for comparison to know significant SNPs in the CYP2C9 gene. The CYP2C9*3 was found to be a clinically significant SNP that has a prevalence in the Bangladeshi population. Then, for the determination of the genotypes in the collected samples, a method was required for genotyping. There are various methods for genotyping of a SNP among which the tetra primer ARMS-PCR method was selected due to the process being inexpensive, time-efficient, and easy to perform in any diagnostic labs across the country. The primers were designed using a web-based primer designing tool and the PCR process was optimized for a successful PCR reaction. The optimization was successful which showed bands in the gel electrophoresis. However, due to the sample number being low, even though there were the bands for the outer primer and the band for wild type allele, no bands for mutant alleles were obtained. Thus, the prevalence of the mutant allele could not be confirmed. Further experimentation on more samples is required to obtain enough data to know the prevalence of mutant alleles in the population using the optimized PCR method.

Table of Contents:

List of Tables:

List of Figures:

Abbreviations

SNP	- Single Nucleotide Polymorphism
CAD	- Coronary Artery Disease
CVD	- Cardiovascular Diseases
PAD	- Peripheral Artery Disease
CD	- Communicable Disease
NCD	- Non-communicable Disease
CYP	- Cytochrome P450
ARMS	- Amplification Refractory Mutation System
PCR	- Polymerase Chain Reaction
DNA	- Deoxyribonucleic acid
NGS	- Next Generation Sequencing
RFLP	- Restriction Fragment Length Polymorphism
SWF	- S-warfarin
BLAST	- Basic Local Alignment Search Tool
PDB	- Protein Data Bank
NCBI	- National Center for Biotechnology Information
Spdbv	- Swiss PDB viewer

Chapter 1
Introduction

1.1 Pharmacogenomics and Pharmacogenetics

Pharmacogenomics and pharmacogenetics are similar words with slight differences. Both terms represent the genetic factors related to the drug response in a treated patient (Lu *et al*., 2015). While pharmacogenomics deal with the drug response in regards to the whole genome of an organism, pharmacogenetics studies the response of drugs related to a single gene only. The application of genomic technologies can be used in pharmacogenomics for a particular patient where the genes of the patient and the capability of the gene are taken into account for the application of treatment for a particular disease which can determine the effectiveness of that treatment along with the reasons for the failure of any past treatment (Ohadoma et al., 2019). While pharmacogenomics can lead to personalized medicine, there are scientific and technological shortcomings that need to be overcome for its full establishment. However, it is possible to optimize the treatment of a patient as much as possible by utilizing the results obtained from pharmacogenetic researches.

1.2 Single Nucleotide Polymorphism (SNP) and its significance

Single Nucleotide Polymorphism (SNP) is the single nucleotide change in a gene sequence that is present in more than one percent of the population. The term allele represents the variation of nucleotide in the gene where the SNP is considered as the minor allele which musthave a presence of 1% in at least one population e.g. CYP2C9*3 which is the change of A>C nucleotide at the position 1075 in the gene (Erichsen & Chanock, 2004). There is a significant role of SNP in influencing the effectiveness of drugs, the side-effects of drugs and also predicting the risk of certain diseases. SNP can also influence the phenotypic expression of certain genes which can be as minor as changing the hair color e.g. rs1805009 in the MC1R gene in chromosome 16 causes the hair color to be red (Lichtenwalter *et al*., 2019). There are other major roles of SNP such as, toxicity from chemotherapy due to the presence of a variant of the XRCC1 gene containing a SNP which changes the amino acid Arginine into Glutamine at the 399 position of the protein (D'Antonio *et al.,* 2014). Early screening of such significant SNP can provide a directive towards the successful treatment of a patient.

1.3 Cardiovascular Diseases and their importance

The cardiovascular system inside the human body mainly consists of the heart and the blood vessels that connect the heart to the rest of the body. This system is essential for the transport of essential nutrients and oxygen to the whole body. There are a variety of abnormalities that

may arise in the cardiovascular system. Cardiovascular disease (CVD) mainly mentions 4 types of abnormalities within the cardiovascular system (Lopez *et al*., 2020).

a) Coronary Artery Disease (CAD): It includes angina, myocardial infarction (MI), and heart failure. CAD is sometimes called coronary Heart Disease (CHD) as well.

b) Cardiovascular Disease (CVD): Brain stroke and transient ischemic attacks (TIA) are included in this.

c) Peripheral artery disease (PAD): Arterial disease involving limbs where too little blood flow to the limbs causing cramping pain or claudication.

d) Aortic atherosclerosis: Thoracic and abdominal aneurysms are part of this.

The CVDs are prevalent non-communicable diseases (NCD) around the globe. The disease is very widespread and an estimated 17.8 million people died due to CVD in 2017 which was one-third of all deaths around the world (W.C.R.C.W. Group, 2019). In Bangladesh the condition is similar. A study report shows that the deaths due to the communicable diseases (CD) declined from 52% to 11% whereas the deaths due to NCD increased to 68% from 8% between 1986 to 2006 (Karar *et al.*, 2009). According to the DGHS report of 2013, the CVDs had the top mortality rate in our country among all NCDs (Bangladesh Government Health bulletin, 2013). A study report on 12,338 participants with an age ≥35 shows that the prevalence of CVD was 4.5% and the risk factors in association with it are also rising significantly (Khanam *et al.*, 2019). The DGHS report of 2018, the mortality from CVDs are 30% of all NCDs (Bangladesh Government Health bulletin, 2018). The studies and reports clearly show the significance of CVDs from the perspective of Bangladesh.

1.4 Warfarin mechanism, metabolism, and its importance

1.4.1 Warfarin

Warfarin is one of the commonly used anticoagulants which prevents the formation of blood clots. It is a synthetic drug and has been used widely as its onset and duration of action can be predicted along with its above-average bioavailability (Hirsh, 1991). It is the racemic mixture of R- and S-enantiomers (figure 1.1). Both enantiomers have separate metabolism pathways. However, S-enantiomer is more potent which is three to five times more than R-enantiomer which makes the metabolism pathway of the former enantiomer more significant than the

later (Patel *et al.*, 2020). Warfarin is administered orally to prevent different complications related to CVDs.

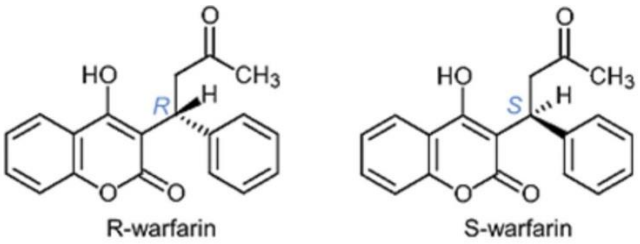


Figure 1.1: Structure of R- and S-enantiomers of warfarin which stays in a racemic mixture in the drug

1.4.2 Mechanism of action of warfarin

Warfarin is a vitamin K antagonist. Warfarin induces the anticoagulation effect by blocking the interconversion between vitamin K (reduced vitamin KH_2) and 2,3 epoxides (vitamin K epoxides) (figure1.2). Vitamin K dependent epoxide reductase (VKORC1) is the enzyme in human which is responsible for the regeneration of reduced vitamin KH_2 from its' epoxide. Warfarin inhibits the VKORC1 which hampers the vitamin KH_2 regeneration process. As a result, vitamin K epoxide is increased in liver and plasma whereas the reduced vitamin KH_2 amount is decreased. The enzyme γ-glutamate carboxylate uses vitamin KH_2 as a cofactor for γ-carboxylation of several vitamin K dependent clotting factors (prothrombin or factor II, VII, IX, and X) and proteins (protein C and protein S) which is involved in the blood clotting mechanism. Due to the deficiency of reduced vitamin KH_2,the blood clotting function is hampered. The γ-carboxylation causes the vitamin K dependent proteins to go through a conformational change where these proteins contain 10 to 13 Gla residues. However, due tothe function of anticoagulant, partially carboxylated or decarboxylated proteins are secreted which contain fewer Gla residues. Prothrombin molecules with fewer than six Gla residues retain 2% whereas molecules with 9 Gla residues retain 70% of its original activity [7].

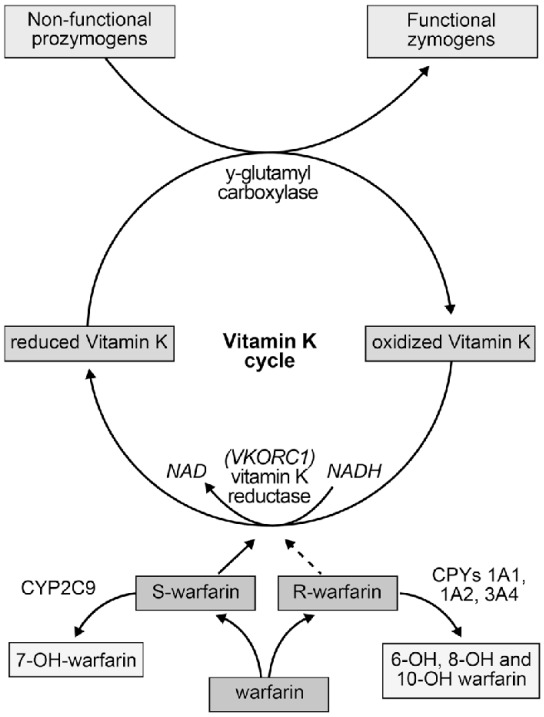


Figure 1.2: Warfarin mechanism of action involves the VKORC1. This enzyme is inhibited by warfarin. The production of reduced vitamin K is thus stopped which ultimately leads to the anticoagulation effect.

1.4.3 Importance of warfarin

Warfarin is used for over 60 years and it has been accepted for the treatment of different types of complexities regarding the CVDs (Dumont *et al.*, 2013). The drug is used for the prevention and treatment of venous thromboembolism (VTE) with deep vein thrombosis, atrial fibrillation, pulmonary embolism, and mechanical heart valves (Baillargeon *et al.*, 2012). Due to the easy availability and diverse usage in different complications, warfarin is used as an anticoagulant for decades. However, the drug also contains some fatal side effects and needs to be in constant monitoring as under- or over-coagulation might occur which can potentially be life-threatening. Warfarin has been linked to a variety of drugs causing major bleeding in the patient (Baillargeon *et al.*, 2012). The internal hemorrhage can be very severe when unnoticed in due time. The common scenario of warfarin use accompanied by an antibiotic can result in over anticoagulation (Juurlink *et al.*,2007). Additionally, there are other side effects like coumarin induced skin necrosis due to the imbalance caused by the rapid reduction of protein C activity (Esmon *et al.*,1987). There are other adverse effects of the drug which includehypersensitivity, rash, alopecia, diarrhea, unexplained drop in haematocrit, purple toes, jaundice, hepatic dysfunction, nausea, vomiting, and pancreatitis

(Mc *et al.*, 2012). The various adverse side-effects of the drug are related to the dosage which depends on the metabolism of the drug by specific proteins. Hence, the study on the proteins can help to determine the dosage requirement of the drug to avoid fatal life-threatening side-effects while obtaining the benefits of the drug to its fullest.

1.4.4 Warfarin Metabolism

Warfarin being a racemic mixture has 2 different enantiomers which are R- and S-enantiomers. The different enantiomers of warfarin are metabolized by different proteins of cytochrome p450 (CYP) protein family. Warfarin acts as a substrate to yield hydroxylated metabolites by the various CYP proteins. The R-enantiomer is metabolized by CYP1A1, CYP1A2, CYP3A4, CYP2C18, CYP2C19 of the CYP protein family to produce 4-, 6-, 8-, 10-hydroxylated metabolites whereas S-enantiomer is mainly metabolized by CYP2C9 protein to yield 6- and 7-hydroxylated metabolite (Kaminsky *et al.*, 1997). Due to the ratio of S-enantiomer being greater in the drug, it has much more significance in the metabolism pathway, drug-drug interaction and drug response.

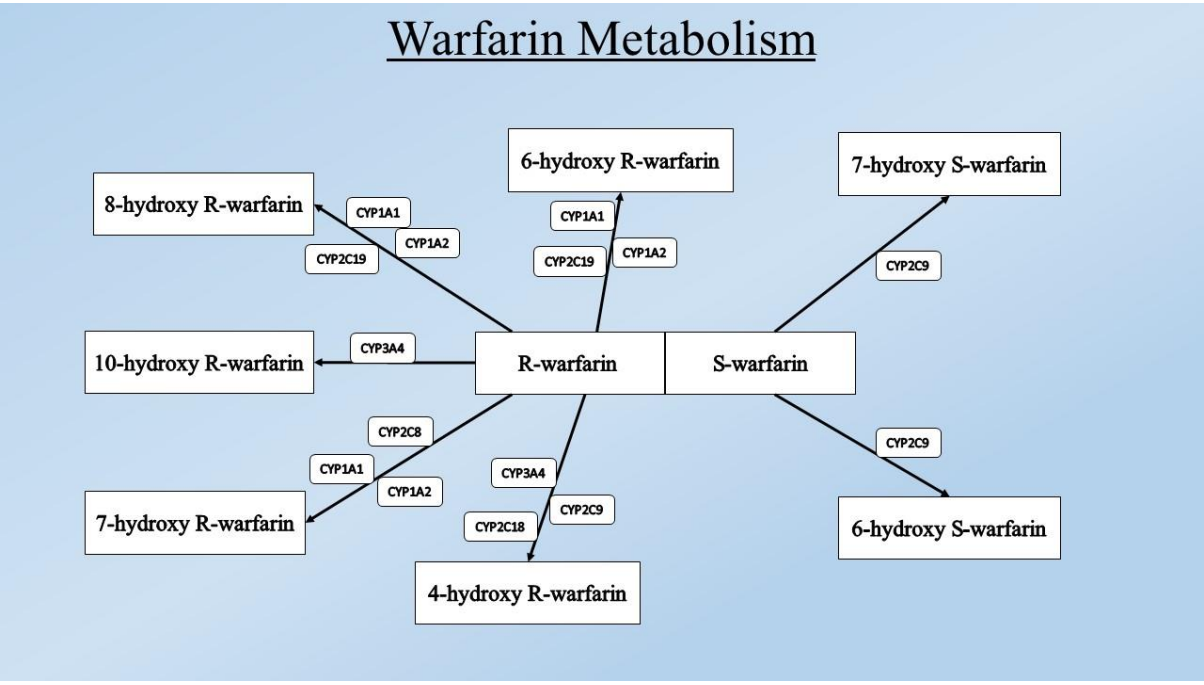


Figure 1.3: Warfarin is a racemic mixture consists of 2 types of enantiomer which are metabolized by different types of CYP protein. R-enantiomer is metabolized by CYP1A1, CYP1A2, CYP3A4, CYP2C18, CYP2C19 of CYP protein family to produce 4-, 6-, 8-, 10-hydroxylated warfarins whereas S-enantiomer is metabolized by CYP2C9 protein to yield 6- and 7-hydroxylated warfarin which is later excreted from the body through urine after conjugation.

1.5 Cytochrome P450 (CYPs)

1.5.1 CYPs and their importance

Cytochrome P450 proteins are membrane-bound hemeproteins. They are present in different parts of the body which include the kidney, liver, intestinal wall, and plasma where they play a vital role in metabolizing drugs, detoxifying xenobiotics (foreign drugs and chemicals), and homeostasis. Drug metabolism is achieved in 2 phases which are Phase I (oxidation, reduction, and hydrolysis) and Phase II (conjugation). Phase I involves the discontinuation of drug activity by the addition of a functional group or activation of the prodrug into active drug whereas in Phase II the drug is conjugated to an exogenous substance like sulfate or glycine which is finally excreted out of the body through urine (Issa *et al.*, 2017). The CYPs are the proteins that work in Phase I of drug metabolism. The primary function of CYPs is the oxidation of various xenobiotics into hydrophilic compounds which are later excreted from the body (McDonnell *et al.*, 2013). The prolonged presence of xenobiotics inside the body can create a toxic effect inside the body and by converting these substances, the CYP helps in keeping the body healthy. Moreover, CYP acts as a catalyst in steroid hormone synthesis which is vital in maintaining various key roles in the body (McDonnell *et al.*, 2013).

CYP proteins are classified through similar gene sequences where a family number, sub-family letter, and isoform number are assigned to each protein. For instance, CYP1A1 is the Cytochrome P450 protein having assigned family 1, sub-family A, and isoform 1. The various CYP protein can metabolize one drug. However, one CYP protein can metabolize multiple drugs as well. There is a possibility of drug-drug interaction when multiple drugs follow the same metabolizing pathway which makes the proteins significant for in-depth study.

1.5.2 Cytochrome P450 family 2, subfamily C, member 9 (CYP2C9)

CYP2C9 is a part of the Cytochrome P450 superfamily which is found in the endoplasmic reticulum of the cell. Among the CYP2C subfamily, CYP2C9 is present in the highest amount in the liver which amounts to approximately 20% of the total CYPs in the liver catalyzing approximately 15% of total drugs metabolized by CYPs making CYP2C9 the third most significant CYP (Daly *et al.*, 2017). CYP2C9 metabolizes multiple types of drugs which include anticoagulants, nonsteroidal anti-inflammatory drugs, sulfonylurea oral hypoglycemic drugs. The protein also has a significant role in metabolizing and clearance of drugs from other therapeutic classes including anticonvulsant phenytoin, the diuretic

torsemide, and the antihypertensive losartan (Daly *et al.*, 2017). There are several SNPs present in CYP2C9 which in many cases determine the drug response of a variety of drugs that is metabolized by this protein. For example, rs2153628 and rs1799853 in CYP2C9 have been found to have a relation with indomethacin response in the treatment of patent ductus arteriosus (PDA) whereas rs1057910 influences the drug response of warfarin (Smith *et al.*, 2017) (Rosdi *et al.*, 2016). By applying pharmacogenetics, the drug response of those particular SNPs could be discovered and by screening a particular SNP for a particular disease, the treatment can be modified for an individual to obtain the best result from his treatment.

1.5.3 Relationship between warfarin and CYP2C9 SNPs

Several SNPs in CYP2C9 have been reported to have an influence in the altered drug response of warfarin. The particular variants of the protein which are CYP2C9*2 containing the SNP rs1799853 and CYP2C9*3 containing the SNP rs1057910 have been reported to have decreased activity compared to the wild type variant CYP2C9*1 (Hosseinkhani *et al.*, 2018). As a result, the variant proteins cannot metabolize warfarin as effectively as the wild type variant resulting in altered drug response. There are also some other variants of the protein such as CYP2C9*5, CYP2C9*6, and CYP2C9*11containing the SNPs rs28371686, rs28371686, and rs18371685 which have been found to be putative poor metabolizer alleles that can alter the drug response of warfarin (Limdi & Veenstra, 2008). This pharmacogenetic relation between warfarin and CYP2C9 variants can determine the efficacy of the drug through maintaining proper dose requirements while avoiding potentially harmful side-effects.

1.6 Determination of clinically significant SNP in CYP2C9

1.6.1 Determination through literature review

The study of CYP2C9 variants has been performed for a long time. As a result, many research articles can be obtained related to the different variants of the CYP2C9 protein. There are multiple clinically significant SNPs that have been observed to alter the drug response of warfarin through cohort studies. The 2 major variants of the CYP2C9 which are found commonly are CYP2C9*2 (rs1799853) and CYP2C9*3 (rs1057910) that have reduced catalytic activity found in the patients of European, African and Asian population (Limdi&Veenstra, 2008). Among these SNPs, the CYP2C9*3 variant has a frequency of 2-4% whereas CYP2C9*2 is normally not obtained in the Asian population but both variants

are more prevalent in the European American population (Limdi & Veenstra, 2008). The other variants such as CYP2C9*5 (rs28371686), CYP2C9*6 (rs28371686) and CYP2C9*11 (rs18371685) has reported to have 0.9-1.8%, 0.1% and 1.5% frequency respectively in African American population (Limdi & Veenstra, 2008). Another study has reported that CYP2C9*1/3 and CYP2C9*3/3 have a frequency of 4.6% and 3.4% with no presence of the CYP2C9*2 in the Bangladeshi population (Chowdhury *et al.*, 2017). The significance of CYP2C9*3 (rs1057910) can be determined to be a clinically significant SNP from the perspective of the Bangladeshi population.

1.6.*2 In-silico* determination of clinically significant SNPs

In-silico analysis is becoming much popular in molecular biology due to the advancement in technology in the last few decades. There are various bioinformatic tools such as PyRx, Swiss PBD viewer (spdbv), chimera, etc. Each of these tools has its own algorithm and function. Some can design protein from sequences, some can create protein to drug interaction, and so on. These tools can be used to determine the possible altered response of a drug for different variants of a protein by mimicking the drug to protein interactions in the human body. All the important SNPs present in a significant amount of the population can be considered to be analyzed through bioinformatics tools. The different proteins containing a specific SNP can be designed through homology modelling and the drug to protein interaction can be observed. The comparison can be done to observe the difference in protein-drug interaction in the wild type protein varies from the variants. By comparing the two interactions, the most significant SNP can be determined through in-silico analysis.

1.6.3 CYP2C9*3 (rs1057910)

CYP2C9*3 has been reported to be a common CYP2C9 SNP in the Asian population and it was also observed that the SNP plays a significant role in dose requirement for a patient who requires warfarin treatment (Limdi & Veenstra, 2008). From the perspective of the Bangladeshi population, the SNP having a 4.6% frequency in heterozygous condition and frequency of 3.4% in homozygous condition can determine the variation in dose requirement during warfarin treatment (Chowdhury *et al.*, 2017). This variant of CYP2C9 has a change in base adenine to cytosine in the 1075^{th} position of the gene causing the change of amino acid isoleucine to leucine. The change in amino acid can directly affect the active site of the protein where the drug warfarin bind has a substrate. The change in the active site can reduce the binding affinity of the drug in the active site causing changes in drug metabolism. As a

consequence, drug activity alters in a patient. In which case the dosage of the drug needs to be modified for better results through treatment.

1.7 Detection of CYP2C*3 (rs1057910) using tetra-primer ARMS PCR

1.7.1 Principles behind tetra-primer ARMS PCR

ARMS stand for Amplification Refractory Mutation System. This is a type of PCR reaction which is specifically designed for genotyping. This technique is highly sensitive and specific for genotype determination. ARMS PCR uses four different primers for the specific determination of a base in a gene. Among the four primers, two primers are called outer primers which bind at a significant distance from the nucleotide that needs to be determined. The other two primers are called inner primers between which one is designed for a wild type allele binding in one strand of the DNA whereas the other is designed for mutant allele binding in the other strand of the DNA. The last nucleotide of the inner primers must be placed on the nucleotide in the DNA that needs to be genotyped. There is another factor that increases the specificity of the process which is the deliberate mismatch of a nucleotide a few bases away from the 3' end nucleotide. As a result, if the allele is wild type, the mutant primer cannot bind properly due to having 2 different mismatches whereas wild type primer can bind easily and give results in a PCR reaction. In the case of the heterozygous genotype, both primers can bind on the opposite DNA strands and give 2 separate bands in gel electrophoresis which can be used to determine the genotype. The process can be easily understood in Figure 1.4 which illustrates the whole process. This process is very advantageous in determining a single SNP in a gene however detection of another closely related SNP which might be inherited due to linkage disequilibrium in a multiple reaction is not possible as the outer primers are spread out from an SNP. The process uses a single PCR tube. Hence, the primer designing becomes complex because there is a possibility of PCR reaction failing due to too much variation is PCR conditions for different primers. As a consequence, suitable primer designing becomes the biggest challenge in the ARMS PCR technique.

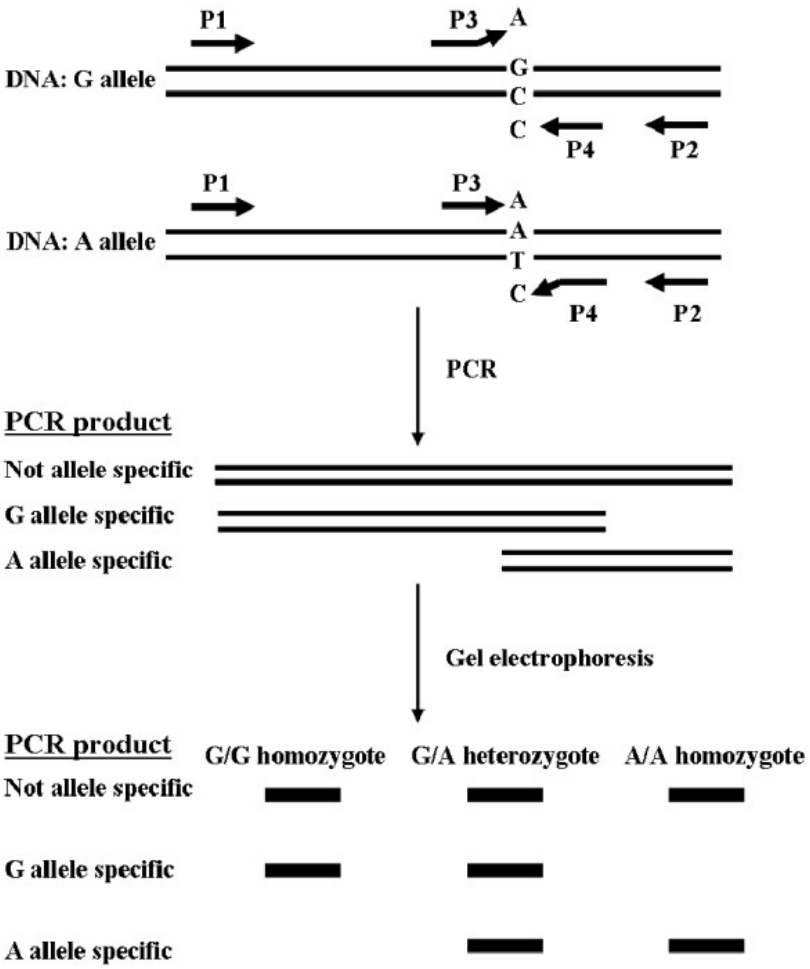


Figure 1.4: The figure illustrates the ARMS PCR method used for genotype detection where 4 different primers are used to determine the genotype of an allele.

1.7.2 Detection of CYP2C9*3 (rs1057910)

There are multiple techniques for the detection of CYP2C9*3 such as sequencing through the Sanger method or Next Generation Sequencing (NGS), restriction fragment length polymorphism (RFLP), Molecular beacon, SNP microarray, TaqMan assay. However, these techniques are costly and require much expertise. ARMS PCR technique having high specificity and sensitivity is a very suitable method for the detection of CYP2C9*3 and one of the reasons is that it requires a minimal amount of equipment and any lab with a PCR set-up can perform it easily. The technique starts from the designing of suitable primers which can be used effectively in PCR reaction. Then after setting up and optimizing the PCR reaction with specific primers for CYP2C9*3, the SNP can be detected efficiently.

1.8 Objectives of the study

The study was designed and performed based on the following objectives-

- To determine the most significant SNPs in the CYP2C9 gene *in silico.*
- To design primer for the tetra-primer ARMS PCR method.
- To optimize the tetra-primer ARMS PCR method and validation by Sanger Sequencing.
- To determine the frequency of the alleles of the study SNP in the Bangladeshi population.

Chapter 2
Materials and Methods

2.1 Study Design

The whole experiment is shown in the flowchart in Figure 2.1, where the experiment was performed in 2 parts which are not shown in the figure. The first part included the determination of a significant SNP in the CYP2C9 gene and designing allele-specific primers for that specific SNP using various bioinformatic tools. This portion of the experiment is also known as *in silico* experiment. The second part is the in vitro experiment where samples were collected from volunteers with their consent from which DNA was extracted using an established protocol. Following the extraction of DNA, the designed primers were used for PCR reaction in vitro where the PCR conditions were optimized for a successful PCR reaction with reproducibility and gel electrophoresis was performed to visualize the result of the PCR reaction and also genotyping of the DNA samples were done from the gel electrophoresis result.

Flowchart showing the procedure of the whole study

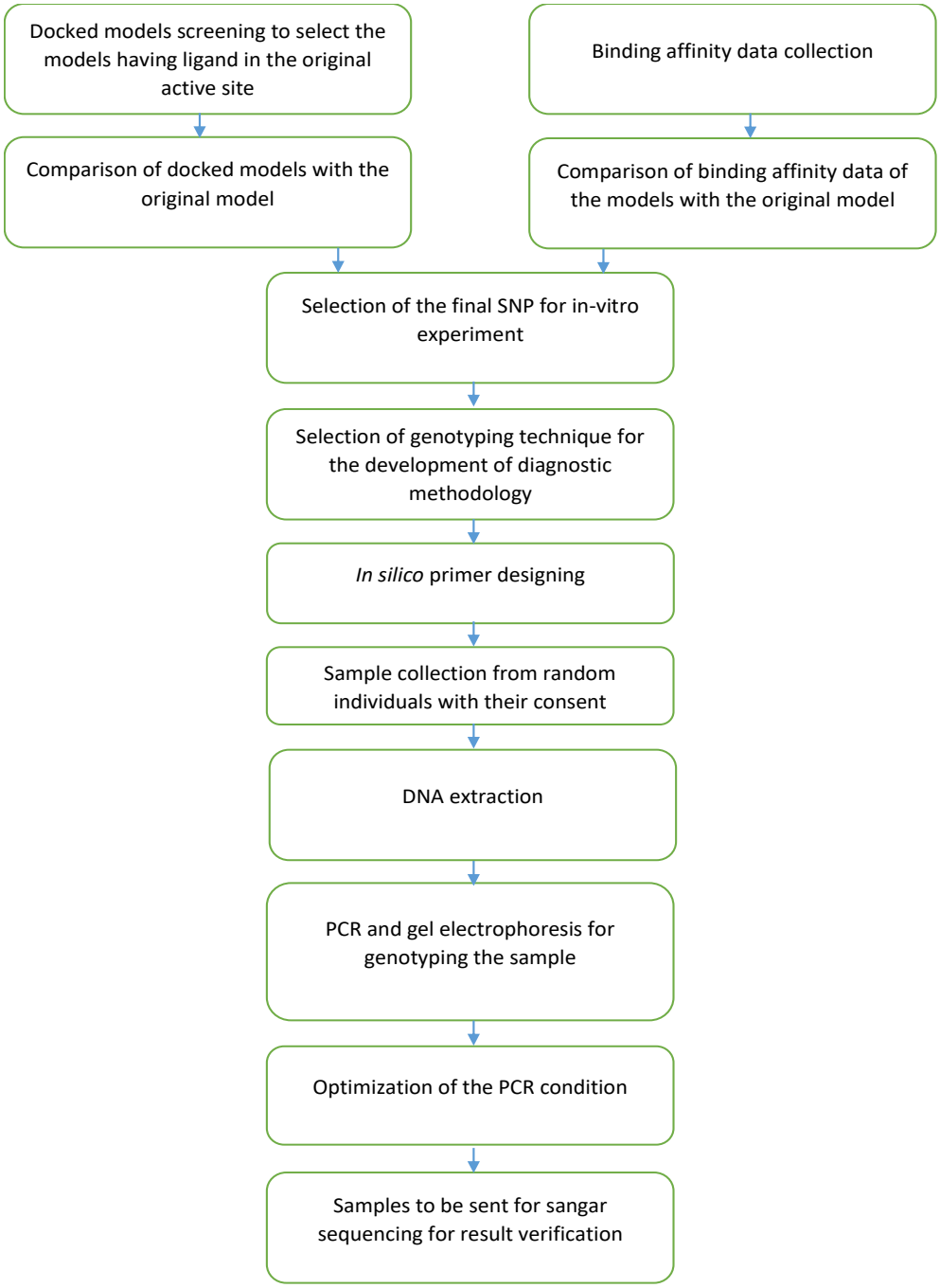


Figure 2.1: The flowchart shows the whole methodology designed to perform the whole study.

2.1.1 *In silico* analysis and primer designing

This part of the experiment was performed to obtain the most significant SNP in the CYP2C9 gene through the use of widely used bioinformatic tools. Furthermore, the sequence of the CYP2C9 gene was also obtained in this part of the experiment which was used to design allele-specific primers for the specific SNP to perform the tetra-primer ARMS PCR. The

suitability of the primers was also checked using various means in this portion of the experiment

2.1.1.1 Selection of SNPs for bioinformatic analysis

The significant SNPs in a protein are mostly found on the nucleotide sequence of the amino acids in the active site which directly interact with the ligand. However, sometimes an SNP changing the amino acid which does not directly interact with the ligand can also be significant and finding those SNPs require in vitro experiments. The SNPs causing a significant change in the amino acid was found out. This was done in the following way:

From a web browser, the URL: https://www.ncbi.nlm.nih.gov was used to visit the NCBI database. The word "*Homo sapiens* CYP2C9" was written in the search box and "Gene" was selected as the database from beside the search box and the search was pressed. The CYP2C9 (ID: 1559) was selected from the search result. "See SNP Geneview Report" was selected from the "Genotypes" part of the "Variation" section. Then, the URL: https://www.uniprot.org was used to visit the UniProt with a web browser. In the search option, "Homo sapiens CYP2C9" was typed, and "UniProtKB" was selected from the dropdown box beside the search box. The "search" option was pressed. From the search result, the entry name "CP2C9_HUMAN" (entry no: P11712) was selected. In the "Structure" section, the PDBe is pressed which is situated in the "Links" section of the 1OG5 model was which contained the CYP2C9 structure bound with the warfarin ligand. In the new window, the "Ligands and Environments" was selected in the "1og5 overview" section in the "Quick links" part. The "Environment details" option of the SWF 502(A) was selected in the new window. Now from the "EBI resources (SWF)" section, the option "Binding site details" was selected. The new window showed the amino acids in the active site of the protein which directly interacts with the ligand s-warfarin along with the positions of the amino acids. Afterward, these amino acid positions were noted down and the SNPs in these amino acids were found out from the Geneview report which was opened earlier from the NCBI database. The SNPs causing significant amino acid changes like changing a basic amino acid to an aromatic amino acid in the active site which can have a significant impact on the protein-ligand interaction were shortlisted for further bioinformatic analysis.

2.1.1.2 Obtaining the nucleotide sequence and amino acid sequence of CYP2C9

The nucleotide sequence of the CYP2C9 gene was retrieved from the NCBI database whereas the amino acid sequence and 3D structure of the protein of this gene were obtained from the UniProt database. The steps which were performed are mentioned below:

From a web browser, the URL: https://www.ncbi.nlm.nih.gov was used to visit the NCBI database. The word "*Homo sapiens* CYP2C9" was written in the search box and "Gene" was selected as the database from beside the search box and the search was pressed. The CYP2C9 (ID: 1559) was selected from the search result. Various information regarding the gene was shown in the new window from which "Go to nucleotide: GenBank" was pressed which is situated under the "Genomic regions, transcripts, and products" section. The FASTA format of the CYP2C9 nucleotide sequence was downloaded from this window of the website from the "Send to:" option. This retrieved sequence was later modified to use it for the primer designing.

To retrieve the amino acid sequence of the human CYP2C9 gene, the URL: https://www.uniprot.org was visited with a web browser. In the search option, "*Homo sapiens* CYP2C9" was typed, and "UniProtKB" was selected from the dropdown box beside the search box. The "search" option was pressed. From the search result, the entry name "CP2C9_HUMAN" (entry no: P11712) was selected. From the "Sequences (2+)" section, the "FASTA" of the Isoform 1 was pressed which revealed a new window containing the FASTA sequence of the amino acid of CYP2C9. The sequence was copied in a new notepad file in .txt format which was used for different purposes later on.

2.1.1.3 Finding homologous protein sequence using NCBI BLAST

Homologous proteins are required for the homology-based modelling of protein 3D structure. The NCBI BLAST (Basic Local Alignment Search Tool) is an excellent tool to find homologous proteins (Altschul *et al.*, 1990). The URL: https://blast.ncbi.nlm.nih.gov was visited using a web browser from which the "Protein BLAST" was selected. The previously retrieved FASTA amino acid sequence of the CYP2C9 protein was uploaded. The database was selected as "Protein Data Bank(PDB)" from the database dropdown box while the rest of the options were kept in default. The "BLAST" option was pressed. After a few moments, the result of the BLAST was shown in a table. The result was sorted by the "Per. Identification" and after sorting 5 different models were selected. The IDs of those proteins were shortlisted

which was used later for 3D protein model retrieval to be used in the homologous protein modelling.

2.1.1.4 Retrieving 3D models of proteins

The RCSB PDB data bank was used to retrieve the 3D models of the proteins (Berman *et al.*, 2000). From a web browser, the URL: https://www.rcsb.org was visited. In the search box, the ID of the specific protein was typed and searched. In the new window, the "Download Files" option was pressed and from the dropdown box, the "PDB Format" option was selected. By repeating this process, the protein models of the IDs 1OG5, 1PQ2, 1R9O, 5XXI, and 4GQS in PDB formats were obtained.

2.1.1.5 Homology-based modelling of protein 3D structure

To prepare a homologous protein containing a SNP, the first requirement is the FASTA amino acid sequence containing the specific amino acid which should be present instead of the wild type amino acid when the gene contains that specific SNP. So, the FASTA amino acid sequence of the CYP2C9 protein was opened in a word editing tool and for 7 SNPs that were selected earlier, 7 different word files were created each containing FASTA amino acid sequences with a single amino acid change corresponding to their SNPs. The bioinformatic tool "EasyModeller 4.0" was opened (Kuntal *et al.*, 2010). Then, by clicking the "Browse" option, the FASTA amino acid sequence containing a SNP was selected. Then, by pressing the "Add Template" option, the 5 PDB format files of the protein models retrieved were opened. Then all 5 templates were selected by clicking the checkbox beside each template. The new protein 3D model was created using these 5 models as reference. Next, the template sequences were aligned with each other in the "Align Template" section where the option "Align Template" was pressed, and then in the next dialogue box "Proceed" was selected. Then the query sequence was aligned with the previously aligned template sequences in the "Align Query" section where the option "Align Query with Templates" was selected and "Proceed" was selected in the next dialogue box. Finally, in the "Build Model" section, the option "Generate Model" was pressed. A new dialogue box appeared where the number of models to generate was selected as 3 and the other options were selected as "NO" and Ok was pressed. The 3D models of the proteins were obtained after the above process was completed.

2.1.1.6 Selection of the best protein model

The Easymodeller 4.0 designs three 3D models for each protein containing a specific SNP. The selection of the best protein model among the 3 was necessary for the following works. For this step, the "SAVES" server was used which was found in the URL: https://servicesn.mbi.ucla.edu/SAVES (Pontius *et al*., 1996). In the SAVES server, three criteria namely Verify 3D, ERRAT, and Procheck were selected and then each of the 3 models was uploaded separately. The results of the 3 models of a protein sequence containing a specific SNP were compared to select the best protein model for that specific SNP.

2.1.1.7 Energy minimization of the modeled proteins

Swiss PDB Viewer (spdbv) tool was used for the energy minimization of the modeled proteins (Guex *et al.*, 1997). The PDB file of the model for energy minimization was opened in spdbv. Then, the "All" option was pressed from the "Select" in the menu bar and after that, the "Energy Minimization" was pressed from the "Tools" in the menu bar. The new model is saved as a separate PDB file.

2.1.1.8 Retrieval of ligand 3D structure

In the UniProt database, the structure of 1OG5 was previously used to know the active sites of the CYP2C9 protein where the ligand s-warfarin can bind. This 1OG5 model visualized a protein-ligand interaction between CYP2C9 and s-warfarin. The IUPAC InChi key of that s-warfarin was obtained from UniProt. The website PubChem was visited from the URL: https://pubchem.ncbi.nlm.nih.gov (Kim *et al.*, 2019). The InChi key was copied in the search box and the search was pressed. The result will show the corresponding s-warfarin ligand for that InChi key which was pressed. From the new window, the 3D structure of the ligand was downloaded in SDF format by pressing the download option in the "3D Conformer" section to retrieve the 3D structure of the ligand.

2.1.1.9 Protein-ligand docking

AutoDockVina in PyRx software was used for this docking of protein with the ligand (Trot & Olson, 2010). The energy minimized 3D model of a designed protein was used to dock with the 3D model of the ligand that was retrieved previously. At first, the energy minimized model file was loaded from the "Load Molecule" option in the File menu. Next, the SDF format file of the ligand was loaded in PyRx from the "Import" option of the File menu where in the next dialogue box, the "Chemical Table File – SDF" option was selected and

Next was pressed. The next step was to minimize the energy of the ligand to make it stable. To do that, the "Minimize a Molecule" option was pressed and in the next dialogue box, the force field was selected as "mmff94", the optimization algorithm was selected as "Steepest Descent", the number of steps was set as 2000 and ok was pressed. Next, the right mouse button was clicked on the protein molecule, and the "Make Macromolecule" option was selected from the "AutoDock" option in the context menu. Then, the right mouse button was clicked on the energy minimized ligand, where the "Convert Selected to AutoDock Ligand (pdbqt)" option was selected from the context menu. The next step is to dock the ligand and the protein. For this, the converted pdbqt format of ligand and the protein model folder was selected from Ligand and Macromolecules respectively from the AutoDock section. The "Vina Wizard" option was selected next and the "Forward" option was pressed from the Select Molecules section. Then, the "Maximize" option was pressed from the Vina Search Space section to maximize the space in which the protein can dock, and forward was pressed to start the docking. Different types of data were found after the docking which included the binding affinity of all the positions in the protein where the ligand can bind. The binding affinity data obtained was then saved as CSV file and the docking data were saved by using the "Export" option in the file menu which was used later for the next steps of the experiment.

2.1.1.10 Preparing positive control for protein-ligand interaction analysis

The 1OG5 model of the protein was opened in BIOVIA discovery studio v4.5. From the sidebar where all the molecules present in the model are shown, all the molecules except the isomer A protein are deleted. Then the model is saved as a PDB formal file in the saves option in the File menu. This cleaned model was used in spdbv for energy minimization and then used as a macromolecule in PyRx to complete the previous procedures of docking. The binding affinity data was also kept for data comparison with other models that contained SNPs. Then the PyMOL v2.3 software was opened (DeLano, 2002). The model 1OG5 was opened in PyMOL. Then the model was shown as a cell from the show option. The cell was then adjusted with the mouse so that the cell is fully visible. Then from the sequence was revealed by pressing "Sequence" in the "Display" in the menu bar. From the sequence, SWF was then selected and shown as a stick. Next, the pdbqt format file of the docked clean model was opened which revealed the ligands coinciding with the 1OG5 ligand. The closest orientation of the ligand of the cleaned model with the ligand of the 1OG5 was noted. Then the UCSF Chimera v1.12 was opened (Pettersen *et al.*, 2004). The "ViewDock" option from

"Surface/Binding Analysis" in the Tools menu was selected. In the new window, the pdbqt format files of both the ligand and protein model obtained from PyRx data was opened. Then from the file menu "Save PDB" was selected. In the new window, the protein molecule and the ligand which was noted in the PyMOL was selected and saved. This new model was used as a positive control in later steps.

2.1.1.11 Docked model screening

The UCSF Chimera v1.12 was used for screening the docked models created from the PyRx (Pettersen *et al.*, 2004). The "ViewDock" option from "Surface/Binding Analysis" in the Tools menu was selected. In the new window, the pdbqt files of both the ligand and protein models which were obtained previously by exporting the PyRx data were selected and opened. The docked models with different positions of ligands were shown which was compared to the 1OG5 model to determine which ligands were present in the active site. Then, the ligands bound to the active site were selected and the docked model of all the selected ligands and the protein was saved as a PDB file. This PDB file and the positive control prepared previously was opened in discovery studio v4.5 separately. Both were compared visually to find the ligands present in the active site and then the ligand with the most similar orientation in the active site and lowest binding energy was kept and the rest of the ligands were deleted in the docked models. Then the model docked with the best ligand was saved as a PDB file.

2.1.1.12 Comparison of the docked model with that of 1OG5

The 1OG5 model was opened in PyMOL and was set as a cell-like it was done previously. Then the SWF molecule was shown as stick like it was done previously. Finally, the pdbqt format files of the docked models were opened and compared with the 1OG5 model.

2.1.1.13 Binding affinity comparison

The binding affinities for all the docked proteins which contained SNPs were compared with the binding affinity of the positive control to find the changes in the binding affinity for the presence of a SNP.

2.1.1.14 Primer designing

The online tool BatchPrimer3 v1.0 was used for primer designing for the allele-specific primers for CYP2C9*3 or rs1057910 which was selected in section 3.1.11 (You *et al.*, 2008).

For primerdesigning, the input sequence was needed to be arranged in a format that can be used by the primer designing tool. To know the exact position of the SNP, the help of the NCBI database was necessary. TheURL: https://www.ncbi.nlm.nih.gov was used in a web browser to visit the NCBI database website. Then by selecting “SNP” as the database, rs1057910 was written in the search box and the search was pressed. The results showed the SNP and the flanking sequence beside the SNP. The flanking sequence beside the SNP was copied. Then the sequence of the CYP2C9 gene which was retrieved previously was opened in a word document. The copied flanking sequence was then pasted in the navigation box of the word document to find the exact position of the SNP. Then from the instructions found in the Batchprimer3 tool, the CYP2C9 nucleotide sequence was arranged as shown in figure 2.2. The primer designing conditions were set as shown in table 2.1. The most suitable primer sets designed by the tool were selected from the result and were used for in vitro experiments in the later steps.

```
>gnl|dbSNP|rs1057910|allelePos=759|totalLen=1500|taxid=9606|mol="genomic"|class=1|alleles="A/C"|bu
ild=123

ACTTGTTCCA TCTTCTTGCC AAGCTGACCA CTTTATCACC AATAGTGATC TTCTTTGTGT CTCCTTATGG TTTTTGTTTT GAAATCTACT
TTGTCTGTTT TAAATATAGT AACTCATGCT CTTTTTTTCA TTTCCATTGG CAGGTACTGT CTCATTCAAT TCCTTTATTT TCAGCCTATG
TGTGTCTTTA TAAGTGAAGT GTGTTTCTTT TAGGCAACAG ATTAATAGGT CTTGTTTTTC CATCCAGGTC AGTAACAGGT CAGTATGTCT
TTTGATTGGA GATTTTATTC CATTTACATT CAGTGTTATT ATTGATAAGT AAGGACTTAC CCATGCCCCT TTGTTATTTG TTTTCTGGTT
GTTTTGTGGA CTTCTCTTCC TTCTTTCATT TCTTCCTGTC TTCCTTTATT GAAGAGAATT TTCTCCACTT ATATGTGTAC AGATTTTTCT
TAATATCTGG TTTATGGCAG TTACACATTT GTGCATCTGT AACCATCCTC TCTTTAAGTT TGCATATACT TCCAGCACTA TAATTTAAAT
TTATAATGAT GTTTGGATAC CTTCATGATT CATATACCCC TGAATTGCTA CAACAAATGT GCCATTTTTC TCCTTTTCCA TCAGTTTTTA
CTTGTGTCTT ATCAGCTAAA GTCCAGGAAG AGATTGAACG TGTGATTGGC AGAAACCGGA GCCCCTGCAT GCAAGACAGG AGCCACATGC
CCTACACAGA TGCTGTGGTG CACGAGGTCC AGAGATAC

M

TTGACCTTCT CCCCACCAGC CTGCCCCATG CAGTGACCTG TGACATTAAA TTCAGAAACT ATCTCATTCC CAAGGTAAGT TTGTTTCTCC
TACACTGCAA CTCCATGTTT TCGAAGTCCC CAAATTCATA GTATCATTTT TAAACCTCTA CCATCACCGG GTGAGAGAAG TGCATAACTC
ATATGTATGG CAGTTTAACT GGACTTTCTC TTGTTTCCAG TTTGGGGCTA TAAAGGTTTG TAACAGGTCC TAGTGTCTGG CAGTGTGTGT
TCTCCAGATT TATTATCTTT CTTCAAGATT GGTTTGGCTA CTCTTAGGTG CTTATATTTC CAAATAATTT TTAAAGGTAT TAGTTTGTCA
ATTTCCCAAA ACCTTGGGCT GGAATTTCTG GCAGGGTGAC ACTAAATTTA TAGGCTAGTT TGGAAAGAAC TGAATCTTGA CACGTTGAGG
CTTTCCATTC CTGAATATAA TTATGCTTCC AATTTGTTTG GGGTTTCTTT TATTTAACCA GGAATGTTGT GAATTTGTTG TCATGGCTTT
CGAGTCTTTG GTTTTCCCTA GATAATTAAT ATTTTTGTTG TAGAACATAA ATAGTTTTTA TCATTCTGAT GATGTTAATC TGTCAACTTT
GCTAAATTTA CTAGTCACTA TTCGTAATTT ATTTCTGGAT TCATTGTAAT TTCTGTGTAT ATTATACTGT ATCTGAGTTA ATATTGTTTT
ATTTCTTATT TTCCATTTCT C
```

Figure 2.2: The figure shows the arrangement of the DNA sequence necessary for the primer designing in the primer designing tool BatchPrimer3 v1.0.

Table 2.1: The table shows the conditions used for primer designing.

Conditions	Minimum	Optimum	Maximum
Primer size	20	22	30
Primer Tm	50	65	80
Max Tm difference	-	5	-
Primer GC%	20	-	80
Inner product size	100	250	400
Relative size difference between inner product sizes	0.5	-	1.5
Max #N's	-	0	-
Salt concentration	-	50	-
Max Self Complementarity	-	8	-
DNA concentration	-	50	-
Max 3' Self Complementarity	-	3	-

2.1.2 In vitro development of SNP tetra-primer ARMS PCR based SNP genotyping method

The in vitro experiment involved samples collected from volunteers from which was used for the optimization of PCR conditions with the primers designed in the earlier *in silico* experiment. The optimized PCR products were prepared for Sanger sequencing to validate the results and finally, individual DNA samples were set to be genotyped.

2.1.2.1 Sample collection

For sample collection, 50 individuals were randomly selected without any bias that might affect the result of the experiment. The saline water was used for the collection of buccal cells while avoiding invasive methods like blood sample collection for simplicity of the later steps. The process is described below:

A volunteer was first required to fill up a consent form with his/her identification and write the answer of a few questions required for statistical analysis and he/she was also required not to eat anything for an hour before the sample collection to avoid any contamination from food particles. A saline solution (0.5%) was prepared beforehand for the collection of buccal cell samples. Next, the volunteer was asked to rinse their mouth vigorously with approximately 15mL saline water for around a minute. A sterile 50mL polypropylene tube was taken and labelled appropriately with the volunteers' ID. Then the volunteers expectorated the water inside the sterile 50mL polypropylene tubes. By repeating the process, 50 samples were collected from 50 random individuals.

2.1.2.2 DNA extraction

The extraction of DNA from the buccal cells was performed using phenol-chloroform extraction. The phenol-chloroform DNA extraction process was used for the extraction of DNA from the collected buccal cell samples (Ghatak *et al*., 2013). The process started with taking a 1.5ml mouth water sample inside 1.5ml Eppendorf tubes® which were centrifuged for 5minutes at 4500rpm. Then the supernatant was removed and again another 1.5ml of the same mouth water was taken in the tubes corresponding to the sample and centrifuged for 5minutes at 4500rpm. This process was repeated a third time and in the end, the only pellet was kept in each tube. Then 500µl lysis buffer added to the pellet which contained 10mM Tris (pH8.0), 10mM EDTA, and 2.0% SDS. After that, 50µl 10% SDS and 10µl Proteinase K were added to the tube and was mixed gently by pipetting. The sample tubes were then incubated in a water bath for 2hours at 65^0C for the lysis buffer to lyse the cell to bring out the DNA into the solution. Then,Phenol:Chloroform:Isoamyl alcohol solution (25:24:1) was added to each tube and the solution was gently mixed through inversion for 3minutes. The samples were then centrifuged at 10000g for 10mins with a temperature setting of 4^0C. After centrifugation, the upper aqueous layer was transferred to a fresh microcentrifuge tube and 10µl RNase A with a concentration of 10mg/ml was added to each microcentrifuge tube and incubated at 37^0C for 30minutes. Then, Chloroform:Isoamyl alcohol (24:1) with a volume equal to the content of the microcentrifuge tube was added and centrifuged at 10000g for 10minutes in 4^0C. After centrifugation, the fresh aqueous layer was transferred to another sterile microcentrifuge tube in which chilled Isopropanol was added in twice the volume of the aqueous layer along with the one-tenth volume of sodium acetate at a concentration of 10M. The tubes were then incubated at -20^0C for 1hour. After incubation, the samples were centrifuged at 10000g for 10mins with a temperature setting of 4^0C. This time the supernatant was discarded and the pellet was air-dried. These pellets were then extracted DNA from the cell samples. The pellets were then suspended in 30µl nuclease-free water. Gel electrophoresis was performed using 3µl of the extracted DNA samples following the procedures mentioned in section **2.1.2.3.2** and the presence of DNA in the extracted samples was visualized following the procedures mentioned in section **2.1.2.3.3** to confirm the presence of DNA in the extracted samples. Finally, the extracted DNA samples were kept in a -20^oC freezer to avoid degradation of DNA samples.

2.1.2.3 PCR reaction and visualization of PCR products

Four different DNA samples were chosen for the optimization of PCR conditions. Initially, a general recipe for regular PCR was used along with general conditions for thermal cycling. The conditions were gradually optimized throughout the experiment. The PCR products were separated by gel electrophoresis and the agarose gel placed under UV rays for the visualization of separate DNA bands.

2.1.2.3.1 Initial attempt of PCR

The four different samples which were chosen previously were used for the initial attempt. The three different annealing temperatures 60^0C, 61^0C, and 62^0C were selected. These annealing temperatures were selected based on the Tm values of all 4 primers. The PCR reaction mixture was prepared sequentially as stated in table 2.2. The thermal cycler could only be set with one annealing temperature at a time. So, 3 different PCR reactions were prepared with the same four samples each time and were run with the same thermal cycler conditions shown in figure 2.3 except the annealing temperature which was different each time. The PCR tubes were set in the middle with avoiding the sides and the thermal cycler was started.

Table 2.2: The table shows the PCR reaction mixture components with their amounts and sequence at which the components were added in the initial attempt of PCR.

Components		Concentration	Volume
Standard Taq Reaction Buffer (contains $MgCl_2$)		10X	2.5µl
dNTP		10mM	0.5 µl
Primer	Outer Forward	10mM	0.25 µl
	Outer Reverse	10mM	0.25 µl
	Inner Forward	10mM	0.25 µl
	Inner Reverse	10mM	0.25 µl
Taq Polymerase		5000U/ml	0.2 µl
Nuclease Free H_2O		-	18.8 µl
Template DNA		-	2.0 µl
Total		-	25 µl

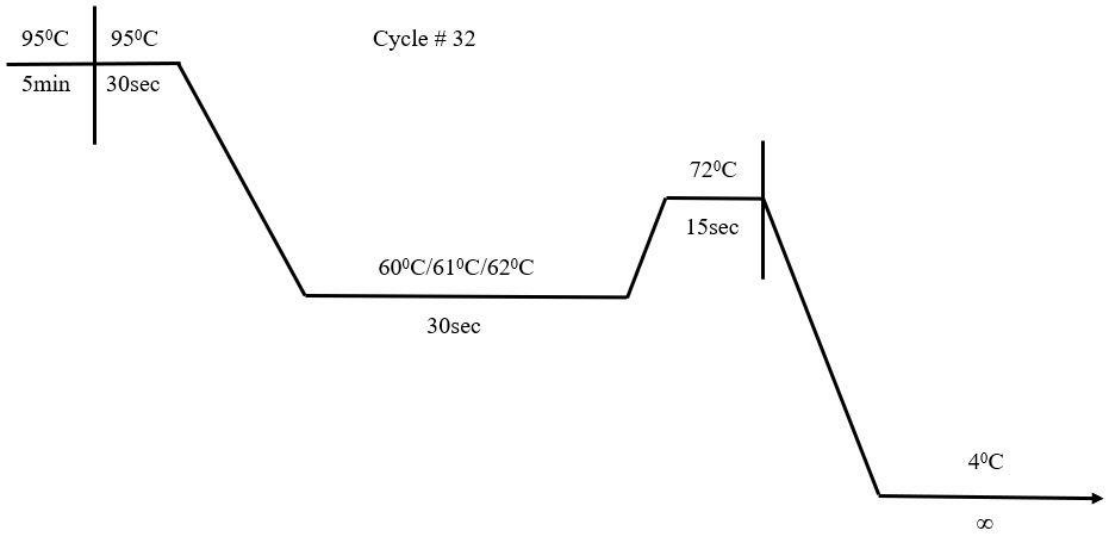


Figure 2.3: The figure shows the thermal cycler conditions used for the initial attempt of PCR some of which were later changed during optimization.

2.1.2.3.2 Separation of DNA bands through gel electrophoresis

After the reaction was completed, 2% agarose powder was mixed with 40mL TAE buffer in a conical flask which was repeatedly heated with an oven in a short duration for uniform mixing of the solution. The gel electrophoresis apparatus was cleaned with 70% ethanol. Then 1µl ethidium bromide was added in the agarose gel and mixed properly. The gel was poured in the electrophoresis apparatus which had a comb placed on it and the gel was left to solidify. After the gel was solid, the comb was removed and TAE buffer was added to submerge the gel. The PCR products were then mixed with a loading dye in a 5µl:1µl ratio and was loaded into the wells. A ladder and negative control were added in 2 empty wells. The voltage was then set to 80volt and the gel electrophoresis was then run for 60minutes.

2.1.2.3.3 Visualization of DNA bands

After the completion of gel electrophoresis, the gel was carefully removed from the electrophoresis apparatus and placed on the UV transilluminator and the safety lid was closed. The transilluminator was switched on which caused UV rays to pass through the gel. The DNA bands which got bound with ethidium bromide in the gel electrophoresis were illuminated and the bands could be observed.

2.1.2.3.4 Optimizing the PCR conditions

The yield of the PCR product was very low in the initial attempts. So, for the next attempt, the annealing temperature was reduced to 59^0C, and PCR was performed with the previous recipe along with the previous thermal cycler conditions. After the result was obtained which can be seen in section 3.2.1, the recipe of the PCR reaction mixture was optimized for higher yield as shown in table 2.3. The annealing temperature was kept at 60^0C which was decided at the best T_A from the initial attempt. However, a few other thermal cycler conditions were changed as shown in Figure 2.4. After the completion of the PCR reaction, the PCR products were run in 2% agarose gel under 80volt for 60minutes following the procedures of section **2.1.2.3.2**. The DNA bands were visualized under UV rays after gel electrophoresis was completed by following section **2.1.2.3.3**.

Table 2.3: The primer volumes, template volume and nuclease-free water volume were optimized for better PCR reaction results.

Components		Concentration	Volume
Standard Taq Reaction Buffer (contains $MgCl_2$)		10X	2.5µl
dNTP		10mM	0.5 µl
Primer	Outer Forward	10mM	0.5 µl
	Outer Reverse	10mM	0.5 µl
	Inner Forward	10mM	0.5 µl
	Inner Reverse	10mM	0.5 µl
Taq Polymerase		5000U/ml	0.2 µl
Nuclease Free H_2O		-	16.8 µl
Template DNA		-	3.0 µl
Total		-	25 µl

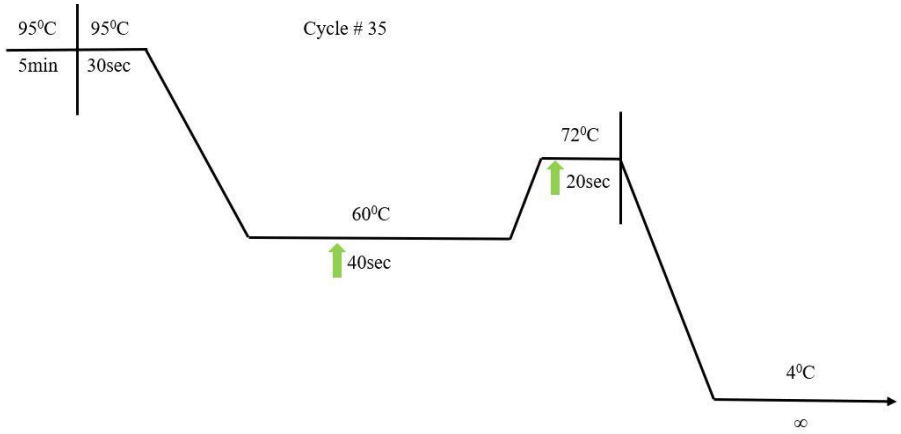


Figure 2.4: The time for the annealing and extension were changed along with the number of cycles to obtain better PCR results.

2.1.2.3.5 Confirming the PCR conditions

After the result from optimization of the PCR reaction was somewhat satisfactory, for the finalization of the PCR conditions ten different samples were selected. The reaction mixture shown in table 2.3 was used along with the thermal cycler conditions mentioned in Figure 2.4. The PCR products were run in 2% agarose gel under 80volt for 60minutesfollowing the procedures of section **2.1.2.3.2**. The DNA bands were visualized under UV rays after gel electrophoresis was completedfollowing section **2.1.2.3.3**.

2.1.2.4 Genotype detection with ARMS-PCR

ARMS PCR uses 4 primers where the DNA bands in the gel electrophoresis are used to determine the genotype. In this experiment, the forward inner primer was designed for the mutant allele and the inner reverse primer was designed for the wild type allele. These 2 primers give separate bands combined with the outer primers depending on the allele present in the gene. The bands under the UV rays were checked. All the PCR products had to have the band formed by the 2 outer primers. The band sizes were checked to determine the genotype. If the band formed by reverse primer was present only, the genotype was homozygous wild type. If the band formed by forward-reverse primer was present only, the genotype was homozygous mutant type. If both bands were present, the genotype was heterozygous. By following this, the genotype of the samples was determined.

2.1.2.5 Validation of PCR result

The validation of the PCR result was supposed to be done using the Sanger sequencing method. However, it was not possible because the labs were closed due to the current pandemic situation. Otherwise, 10 PCR product samples were set to be validated using the sanger sequencing method.

Chapter 3
Results

3.1 *In silico* analysis results

3.1.1 Selection of SNPs in CYP2C9 for bioinformatic analysis

The SNPs were selected by 2 methods. The first one was through a literature review where the SNPs rs1057910 and rs1799853 were selected due to having an extensive study on these SNPs and also the ability of the SNPs to reduce the activity of the CYP2C9 protein. The other method which was followed was to find the amino acids in the active site which interact with the ligand and then finding possible SNPs in those amino acids which can change the amino acid in a way that the tertiary structure of the protein would be affected. As said in the methodologies section 2.1.1.1, the table 3.1 shows the list of all amino acids present in the active site of the protein which was found in the UniProt database and also the list of all SNPs present in the CYP2C9 protein which was obtained from the Geneview report from NCBI database. The table 3.2 shows the list of selected SNPs, the position of the SNPs in the amino acids of the protein, and the change in amino acid due to the presence of SNP.

Table 3.1: This table shows the amino acids in the active site which interact with through various type of bonds and interactions. The position of the amino acid in the protein is also shown in the table.

Name of the amino acid in active site	Position of the amino acid in the protein	Type of bond formed with the ligand
Phe	100	Hydrogen
Leu	102	Electrostatic
Thr	364	Van-der-waals
Ser	365	Van-der-waals
Arg	97	Van-der-waals
Gly	98	Van-der-waals
Ala	103	Van-der-waals
Val	113	Van-der-waals
Phe	114	Van-der-waals
Leu	208	Van-der-waals
Asn	217	Van-der-waals
Leu	366	Van-der-waals
Pro	367	Van-der-waals
Phe	476	Van-der-waals

Table 3.2: The table shows the selected SNPs along with their position in the protein and change in the nature of the amino acid due to the presence of the SNP.

Name of the SNP	Position of the SNP in protein	Change in amino acid	Change in amino acid side chain due to the presence of SNP
rs1057910	359	I>L	Isomeric change in the side chain
rs1799853	144	R>C	Positively charged side chain changed to polar uncharged side chain
rs1240526815	98	G>R	Nonpolar aliphatic side chain changed to positively charged side chain
rs762239445	98	G>D	Nonpolar aliphatic side chain changed to negatively charged side chain
rs368462490	208	L>F	Nonpolar aliphatic side chain changed to aromatic side chain
rs749300923	367	P>S	Nonpolar aliphatic side chain changed to polar uncharged side chain
rs760340421	367	P>H	Nonpolar aliphatic side chain changed to positively charged side chain

3.1.2 DNA and amino acid sequence of the CYP2C9 gene

CYP2C9 is a gene present in the p-arm of chromosome 10 which is 51434bp in size. The gene has a 2561bp mRNA transcript which is later translated into a precursor protein of 490 amino acid in size. The various information of the gene is shown in the table 3.3.

Table 3.3: The table shows information regarding the CYP2C9 gene.

CYP2C9 Summary	
Official Symbol	CYP2C9 (provided by HGNC)
Official Full Name	cytochrome P450 family 2 subfamily C member 9
Gene ID	1559
Gene type	protein-coding
RefSeq status	Reviewed
Accession Number	NC_000010.11
Organism	*Homo sapiens*
Location	10q23.33
Size of the Gene	51434 bp
Exon count	9
Size of the mRNA	2561 bp
mRNA Accession Number	NM_000771.4
Size of the Protein	490 aa

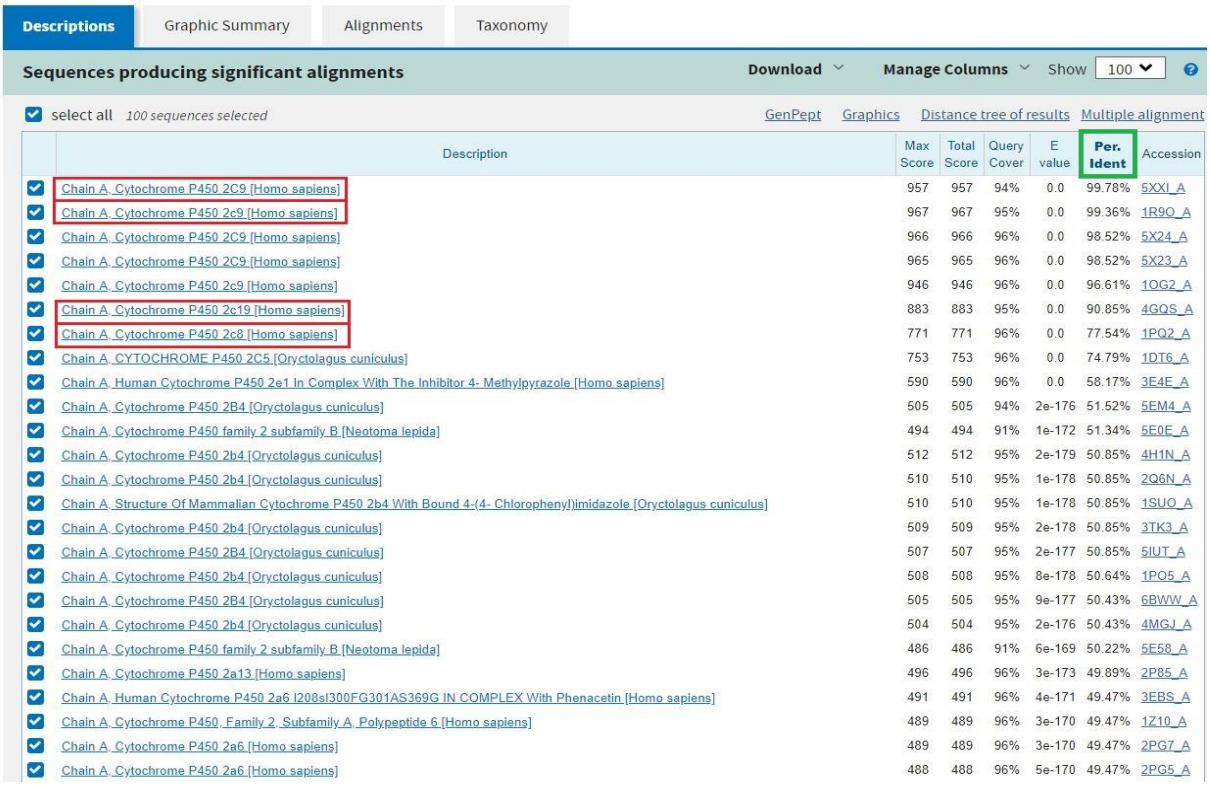
The amino acid sequence of the protein contains 490 amino acids. Various information and the sequence of the protein are shown in the table 3.4.

Table 3.4: The table shows the amino acid sequence of the CYP2C9 and some related information.

CYP2C9 Protein Summary	
Size	490 aa
Locus	NP_000762
Accession Number	NP_000762.2
Organism	*Homo sapiens*
Signal Peptide Position	1..25
Amino acid Sequence	>NP_000762.2 cytochrome P450 2C9 precursor [Homo sapiens] MDSLVVLVLCLSCLLLLSLWRQSSGRGKLPPGPTPLPVIGNIL QIGIKDISKSLTNLSKVYGPVFTLYFGLKPIVVLHGYEAVKEA LIDLGEEFSGRGIFPLAERANRGFGIVFSNGKKWKEIRRFSLM TLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILG CAPCNVICSIIFHKRFDYKDQQFLNLMEKLNENIKILSSPWIQI CNNFSPIIDYFPGTHNKLLKNVAFMKSYILEKVKEHQESMDM NNPQDFIDCFLMKMEKEKHNQPSEFTIESLENTAVDLFGAGT ETTSTTLRYALLLLLKHPEVTAKVQEEIERVIGRNRSPCMQD RSHMPYTDAVVHEVQRYIDLLPTSLPHAVTCDIKFRNYLIPK GTTILISLTSVLHDNKEFPNPEMFDPHHFLDEGGNFKKSKYF MPFSAGKRICVGEALAGMELFLFLTSILQNFNLKSLVDPKNL DTTPVVNGFASVPPFYQLCFIPV

3.1.3 Finding homologous protein sequence using BLAST

Homologous protein modelling requires reference models that are used to model a new protein. Thus, finding homologous proteins through blastp is an important step in modelling new proteins. After using the blastp using the “Protein Data Bank (pdb)” as the database, the result was sorted by the “Per. Identification” as shown in the green box Figure 3.1. The red boxed models in figure 3.1 were selected to be used in homologous modelling which has the following IDs: 1PQ2, 1R9O,5XXI and 4GQS. Along with these proteins, the protein model 1OG5 was selected which is a model bound with warfarin and Heme-C as ligands. Among the models of the proteins that were selected, not all of them had high similarity with the query sequence. This helps with the accuracy of protein modelling.



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Manage Columns — Show 100

select all 100 sequences selected — GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Chain A, Cytochrome P450 2C9 [Homo sapiens]	957	957	94%	0.0	99.78%	5XXI_A
Chain A, Cytochrome P450 2c9 [Homo sapiens]	967	967	95%	0.0	99.36%	1R9O_A
Chain A, Cytochrome P450 2C9 [Homo sapiens]	966	966	96%	0.0	98.52%	5X24_A
Chain A, Cytochrome P450 2C9 [Homo sapiens]	965	965	96%	0.0	98.52%	5X23_A
Chain A, Cytochrome P450 2c9 [Homo sapiens]	946	946	96%	0.0	96.61%	1OG2_A
Chain A, Cytochrome P450 2c19 [Homo sapiens]	883	883	95%	0.0	90.85%	4GQS_A
Chain A, Cytochrome P450 2c8 [Homo sapiens]	771	771	96%	0.0	77.54%	1PQ2_A
Chain A, CYTOCHROME P450 2C5 [Oryctolagus cuniculus]	753	753	96%	0.0	74.79%	1DT6_A
Chain A, Human Cytochrome P450 2e1 In Complex With The Inhibitor 4- Methylpyrazole [Homo sapiens]	590	590	96%	0.0	58.17%	3E4E_A
Chain A, Cytochrome P450 2B4 [Oryctolagus cuniculus]	505	505	94%	2e-176	51.52%	5EM4_A
Chain A, Cytochrome P450 family 2 subfamily B [Neotoma lepida]	494	494	91%	1e-172	51.34%	5E0E_A
Chain A, Cytochrome P450 2b4 [Oryctolagus cuniculus]	512	512	95%	2e-179	50.85%	4H1N_A
Chain A, Cytochrome P450 2b4 [Oryctolagus cuniculus]	510	510	95%	1e-178	50.85%	2Q6N_A
Chain A, Structure Of Mammalian Cytochrome P450 2b4 With Bound 4-(4- Chlorophenyl)imidazole [Oryctolagus cuniculus]	510	510	95%	1e-178	50.85%	1SUO_A
Chain A, Cytochrome P450 2b4 [Oryctolagus cuniculus]	509	509	95%	2e-178	50.85%	3TK3_A
Chain A, Cytochrome P450 2B4 [Oryctolagus cuniculus]	507	507	95%	2e-177	50.85%	5IUT_A
Chain A, Cytochrome P450 2b4 [Oryctolagus cuniculus]	508	508	95%	8e-178	50.64%	1PO5_A
Chain A, Cytochrome P450 2B4 [Oryctolagus cuniculus]	505	505	95%	9e-177	50.43%	6BWW_A
Chain A, Cytochrome P450 2b4 [Oryctolagus cuniculus]	504	504	95%	2e-176	50.43%	4MGJ_A
Chain A, Cytochrome P450 family 2 subfamily B [Neotoma lepida]	486	486	91%	6e-169	50.22%	5E58_A
Chain A, Cytochrome P450 2a13 [Homo sapiens]	496	496	96%	3e-173	49.89%	2P85_A
Chain A, Human Cytochrome P450 2a6 I208sI300FG301AS369G IN COMPLEX With Phenacetin [Homo sapiens]	491	491	96%	4e-171	49.47%	3EBS_A
Chain A, Cytochrome P450, Family 2, Subfamily A, Polypeptide 6 [Homo sapiens]	489	489	96%	3e-170	49.47%	1Z10_A
Chain A, Cytochrome P450 2a6 [Homo sapiens]	489	489	96%	3e-170	49.47%	2PG7_A
Chain A, Cytochrome P450 2a6 [Homo sapiens]	488	488	96%	5e-170	49.47%	2PG5_A

Figure 3.1: The figure showed the result of the BLAST search which was sorted using the Per. Identification highlighted in the green box. The red box highlighted the proteins selected for the homologous protein modelling in the next steps.

3.1.4 Retrieving the 3D models of the reference proteins

The 3D models of the proteins with the IDs 1OG5, 1PQ2, 1R9O,5XXI, and 4GQS which were selected from blastp were retrieved from the RCSB PDB data bank. The 3D models of the selected proteins are shown in Figure 3.2.

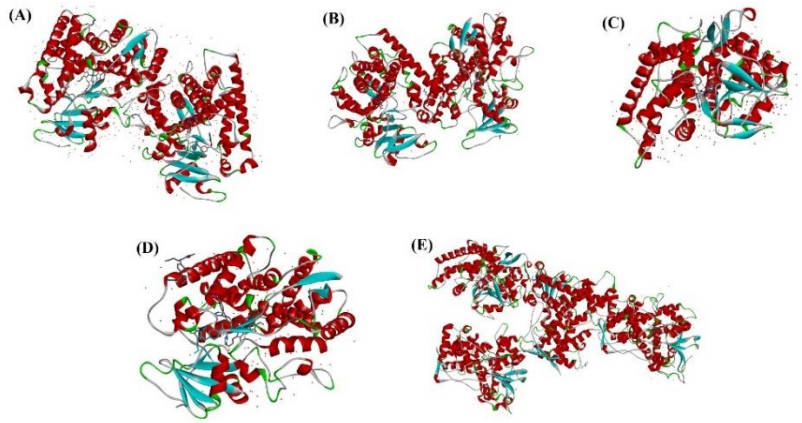


Figure 3.2: The figure shows the models of the proteins with IDs 1OG5(A), 1PQ2(B), 1RO9(C), 5XXI(D), and 4GQS(E). The variations in the models could be identified easily where some proteins have multiple isoforms whereas others have only one.

3.1.5 Homology based modelling of protein 3D structures

There were 7 SNPs that were selected in previous steps which are the topic of study in this experiment. It was required to model 7 proteins each containing one of those 7 SNPs. So first, the amino acid sequence was changed according to the SNP to obtain 7 different CYP2C9 sequences each having 1 of the selected SNP. The changed amino acid sequences were used as query by the EasyModeller4.0 tool to design homologous proteins. The EasyModeller4.0 modelled 3 proteins for each query sequence by taking the selected 3D models as reference. As a result, 21 protein models were obtained in total for the 7 SNPs. The models are all similar and could not be distinguished visually and so a modelled protein is shown as an example in Figure 3.3.

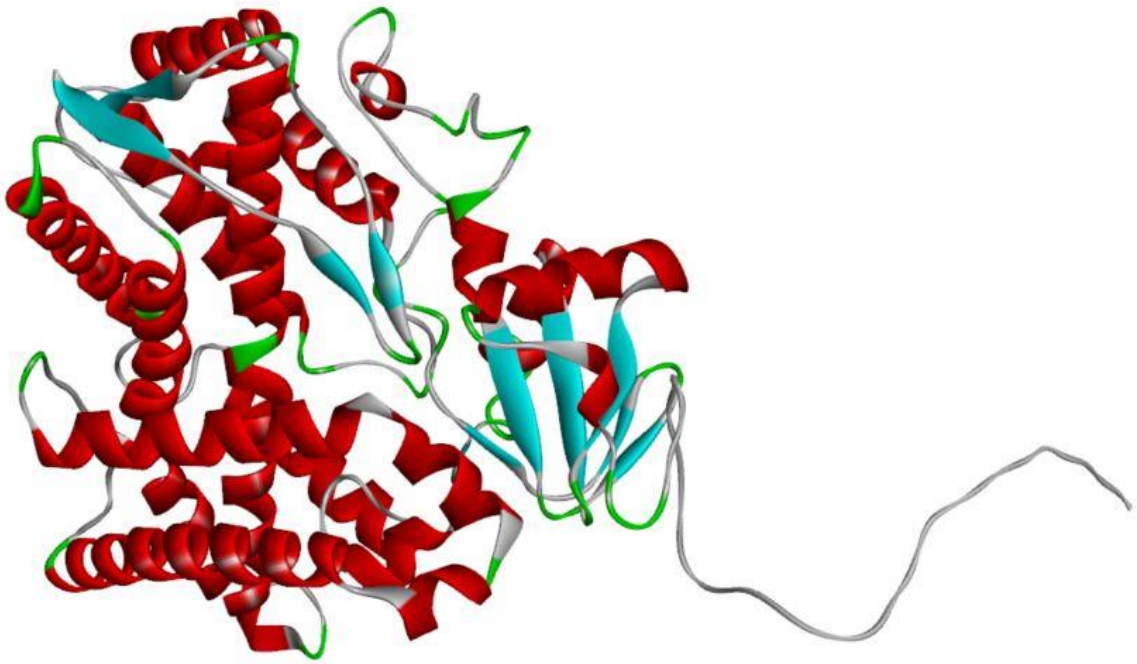


Figure 3.3: The figure shows one of the 3 models designed with a query sequence containing the SNP rs1057910. The various parts of the proteins like alpha helix, beta sheets, turns, etc. can be recognized.

3.1.6 Selection of the best protein model

The 3 models for each SNP modelled by EasyModeller4.0 are different in various aspects like dope profile, Ramachandran plot, etc. The usage of all 3 models for each SNP increases

the complexity of the process and all models might not have been modelled correctly. As a result, the best model out of the 3 for each SNP was selected using the SAVES server. The criteria were set to ERRAT, Verify3D, and Procheck out of various criteria which are shown in figure 3.4 based on which the best model was selected. Moreover, the main Ramachandran plot was also taken into concern while selecting the best model for each SNP. The selected models visually look similar to the model shown in figure 3.3.

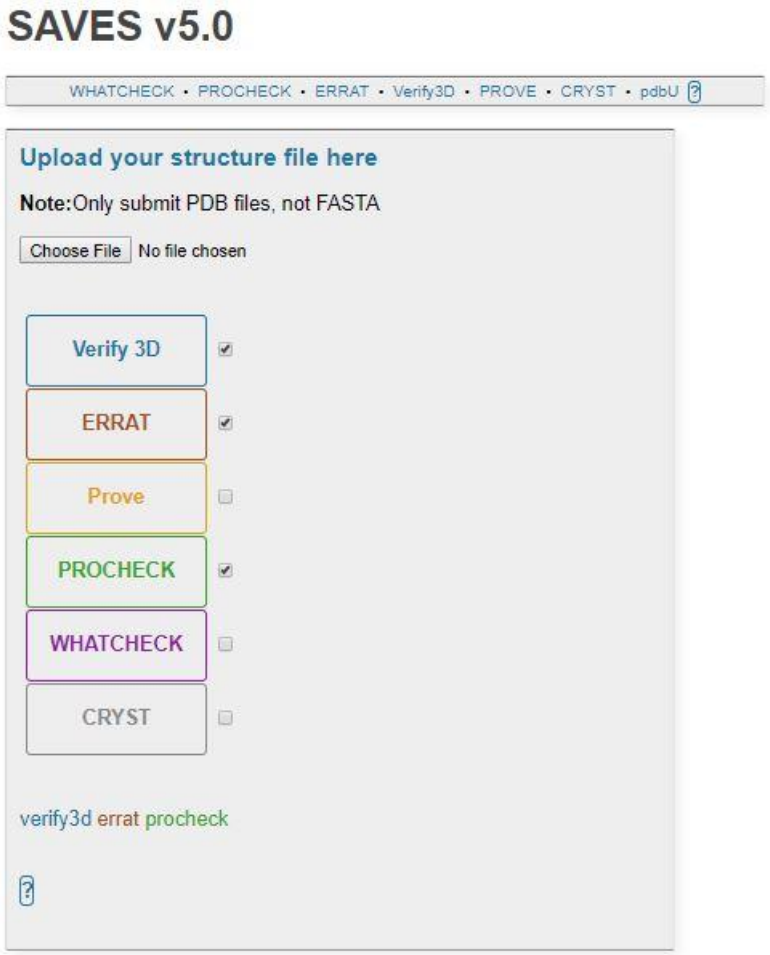


Figure 3.4: The figure shows the website for the SAVES server where the different criteria for checking the quality of the proteins can be seen. The Verify3D, ERRAT, and Procheck was selected among the 6 criteria to determine the best of the 3 proteins designed by EasyModeller for each SNP

3.1.7 Ligand 3D structure

The interaction between protein and ligand with bioinformatic tools require 3D models of both. Previously, the 3D models of proteins each containing a specific SNP were designed using EasyModeller4.0. Then searching with the InChi key of the ligand, the original 3D structure was obtained from the PubChem. If the 3D structure were not available, the 3D

model could also be designed using a different tool. The 3D structure of the ligand is shown in figure 3.5.

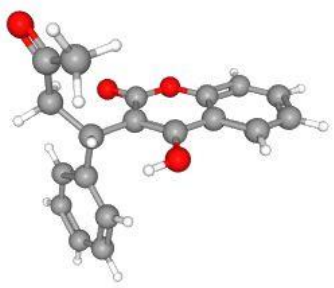


Figure 3.5: The figure shows the 3D model of the ligand S-warfarin where the carbons, oxygen, and hydrogen atoms are shown through the grey, red and white balls respectively.

3.1.8 Positive control preparation for protein-ligand interaction analysis

A positive control was required for the comparison of models to determine whether the docking was successful. The positive control was prepared by docking a clean established model with the ligand and compared with an established model to know whether the ligand-bound with the protein in the active site. The docking results provided the binding affinity data which was used to compare with the binding affinity data of other designed models. The binding affinity data is shown in the figure 3.6. The comparison of the positive control with an established model is shown in the figure 3.7 where the 9 different positions of the ligand were obtained from which the best possible orientation of the ligand with protein was the first docked model.

Ligand	Binding Affinity
cleaned_1og5_54688261_mmff94_E=55.15	-9.3
cleaned_1og5_54688261_mmff94_E=55.15	-8.7
cleaned_1og5_54688261_mmff94_E=55.15	-8.1
cleaned_1og5_54688261_mmff94_E=55.15	-8.1
cleaned_1og5_54688261_mmff94_E=55.15	-8
cleaned_1og5_54688261_mmff94_E=55.15	-7.8
cleaned_1og5_54688261_mmff94_E=55.15	-7.4
cleaned_1og5_54688261_mmff94_E=55.15	-7.3
cleaned_1og5_54688261_mmff94_E=55.15	-7.3

Figure 3.6: The figure shows the binding affinity data for the positive control which was obtained through the docking in PyRx tool

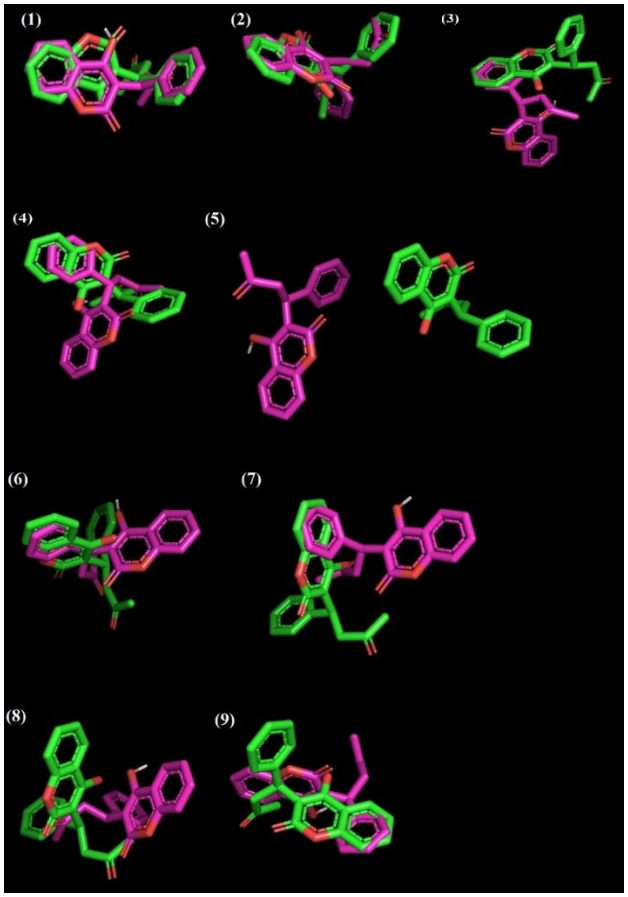
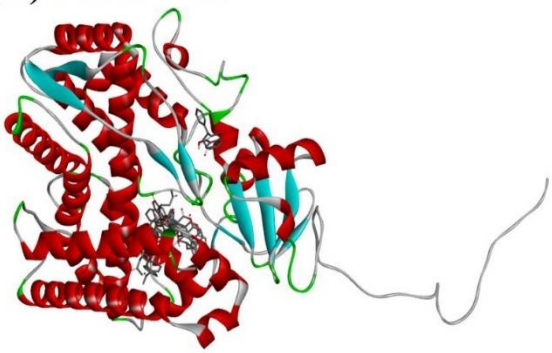


Figure 3.7: The figure shows the overlapping of the different ligand positions of positive control shown in pink color with the ligand of the 1OG5 model shown in green color. The numbering in the position of the ligands of docking is correspondent from top to bottom of the affinity data in the figure 3.6. This figure shows the (1) ligand present in the active site showing the most similarity with the 1OG5 model with the lowest binding affinity of -9.3kcal/mol

3.1.9 Protein-ligand docking using PyRx

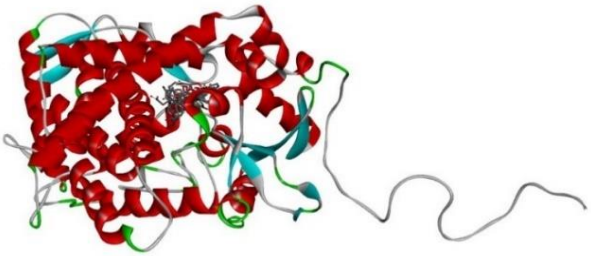
The PyRx tool used the 7 different protein models to dock with the ligand molecule. The energy minimization of the ligand molecule was done first to make the molecule more stable before docking. The results of the PyRx tool have shown the binding affinity of the ligand with each protein molecule. The tool tried to dock the ligand in 9 places in a protein with the binding affinity of the ligand for each place which is shown in the figure 3.8. The binding affinity data is shown in the figure 3.9.

(A) rs1799853

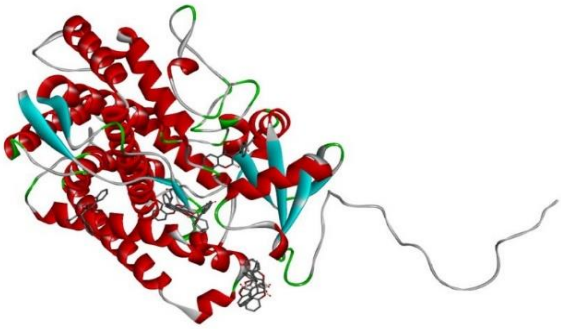


(B) rs368462490

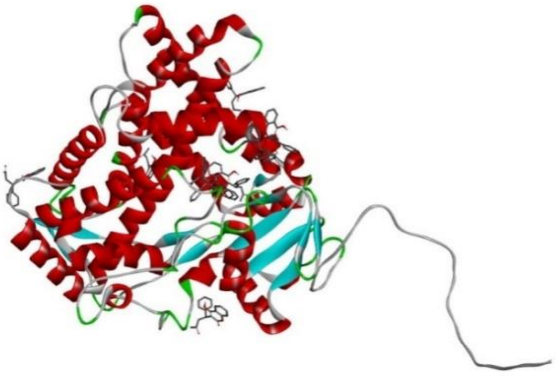
(C) rs1240526815



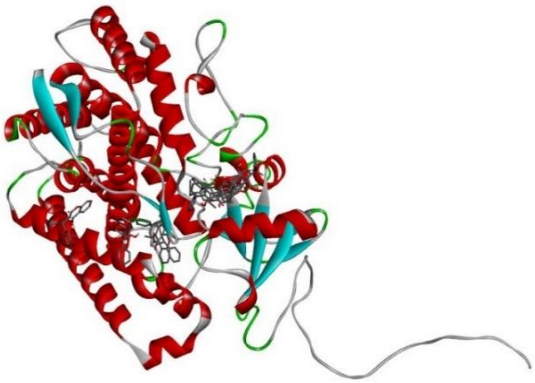
(D) rs760340421



(E) rs1057910



(F) rs762239445



(G) rs749300923

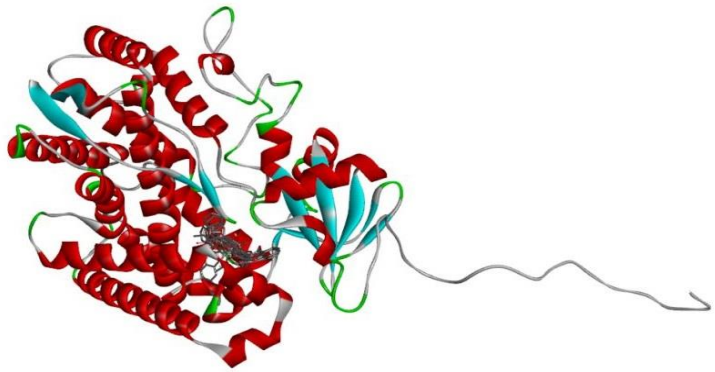


Figure 3.8: The figure shows protein-ligand docking done through the use of the tool PyRx. It can be observed in the figure that not all ligands in a model are bound to the active site of the protein. However, there are protein models in which the active site interacted with all the ligands as well.

(A) rs1799853

Ligand	Binding Affinity
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-7.9
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-7.8
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-7.5
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-7.4
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-7
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-7
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-6.9
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-6.6
rs1799853_energy_minimized_model_54688261_mmff94_E=55.15	-6.5

(B) rs368462490

Ligand	Binding Affinity
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-9.1
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-8.7
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-8.6
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-8.4
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-8.3
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-7.7
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-7.7
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-7.7
CYP2C9_rs368462490_trimmed_model_1_54688261_mmff94_E=55.15	-7.3

(C) rs1240526815

Ligand	Binding Affinity
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-8.1
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.7
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.7
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.6
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.6
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.4
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.4
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.3
rs1240526815_energy_minimized_model_54688261_mmff94_E=55.15	-7.2

(D) rs760340421

Ligand	Binding Affinity
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-8.1
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-7.8
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-7.7
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-7.3
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-6.9
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-6.9
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-6.5
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-6.4
rs760340421_energy_minimized_model_54688261_mmff94_E=55.15	-6.3

(E) rs1057910

Ligand	Binding Affinity
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-7.9
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-7.2
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-6.6
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-6.3
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-6
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-6
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-6
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-5.9
rs1057910_energy_minimized_model_54688261_mmff94_E=55.15	-5.9

(F) rs762239445

Ligand	Binding Affinity
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-8.1
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-7.1
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-7.1
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-7
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-6.4
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-6.4
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-6.2
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-6.1
rs762239445_energy_minimized_model_54688261_mmff94_E=55.15	-6.1

(G) rs749300923

Ligand	Binding Affinity
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-9.1
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-8.7
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-8.6
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-8.4
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-8.2
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-8.2
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-8
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-7.9
rs749300923_energy_minimized_model_54688261_mmff94_E=55.15	-7.8

Figure 3.9: The figure shows the binding affinity data of all the interactions between ligands and proteins showed in figure 3.6 for each model containing a specific SNP.

3.1.10 Binding affinity data comparison and selection of the SNP

The binding affinity data of different models were obtained during the docking in PyRx software which uses the unit in negative kcal/mol. From the data, it was found that the positive control had a binding affinity of -9.3kcal/mol. It could be observed from previous data that the models of the SNPs rs1057910 and rs1799853 had the highest change with both having a binding affinity of -7.9kcal/mol. This 15% change in the binding affinity is the highest among all the models designed with each carrying a specific SNP. As a result, these 2 SNPs were initially selected for the in vitro experiment. However, the ARMS-PCR can be used to screen only one SNP at a time as multiplexing is not possible in this technique. Thus, one of the SNPs had to be selected from the two SNPs rs1057910 and rs1799853. For the selection of the SNP, the help of a literature review was taken where rs1057910 was much more prevalent in various research related to drug dosing. Moreover, in a study done in Bangladesh, the prevalence of rs1799853 was not found in the samples. Thus, rs1057910 was finally selected for the in vitro experiment.

3.1.11 Allele-specific primer designing

The search result of rs1057910 in the NCBI database which in Figure 3.10 is showing the flanking sequence of the rs1057910 SNP "5′-AGATGCTGTGGTGCACGAGGTCCAGAGA TAC-3′". It was used to find the SNP in the retrieved CYP2C9 DNA sequence.

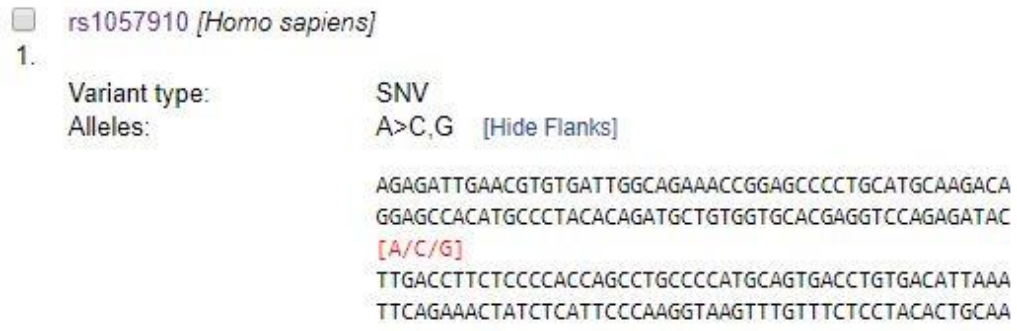


Figure 3.10: The figure shows the search result of the SNP in NCBI which is showing the flanking sequence of the SNP rs1057910

The table 3.5 contains the selected allele-specific primer sets for the SNP rs1057910 along with additional information like the melting temperature (Tm), GC content, product size, etc. The 3′ end forward inner primer was designed for the mutant allele with a product size of 145bp whereas the reverse inner primer was designed for the wild type allele with a product

size of 81bp. This indicates that the band size of 81bp meant a sample had a wild type allele and a band size of 145bp meant a sample had a mutant allele. There is a deliberate mismatch at the fourth position from the 3′ end which could be observed and it increases the sensitivity of the ARMS-PCR process. The outer primers had a product size of 184bp which had present in every sample. The length of the primers was 22bp except the inner reverse primer which was 21bp in size. The closeness of the Tm value could also be observed which made it easy to set an annealing temperature and made the optimization of the PCR process easier.

Table 3.5: The table shows the primers designed for ARMS PCR and various information regarding the primers.

Index	Primer type	Orientation	Start	Length	Tm	GC %	SNP	Position	Primer Seq	Prod Size
1	Outer primer	Forward	699	22	66.92	54.55			ATGCAAGACAGGAGCCACATGC	184
2	Outer primer	Reverse	882	22	65.81	45.45			TTGGGGACTTCGAAAACATGGA	
3	Inner primer	Forward	738	22	68.72	63.64	C	759	GTGCACGAGGTCCAGAGATGCC	145
4	Inner primer	Reverse	781	21	67	61.9	A	759	GCTGGTGGGGAGAAGGTCGAT	81

3.2 Results of in vitro development of SNP tetra-primer ARMS PCR basedSNP genotyping method

3.2.1 Optimization of PCR conditions

The initial attempt for PCR with 60^0C annealing temperature resulted in a very faint band for 2 samples of both 81bp and 145bp whereas no bands for the other of the 2 samples as shown in figure 3.11. The negative control showed no bands which proved that the process had no contamination.

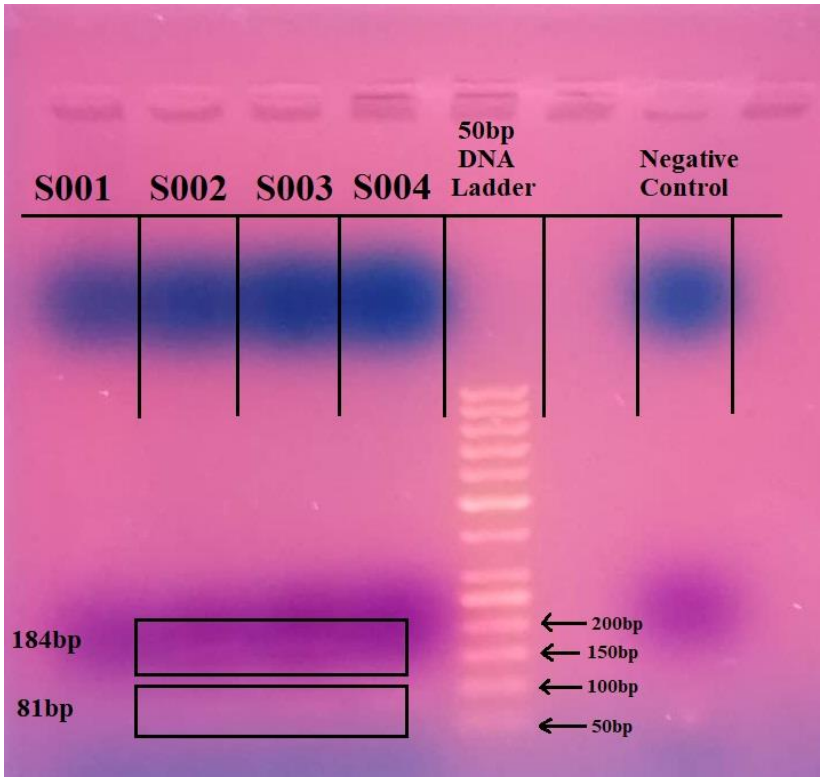


Figure 3.11: The gel electrophoresis figure shows that faint bands of 184bp and 181bp were found for the samples S002 and S003 but other samples did not show results from the PCR reaction performed with 60^0C annealing temperature.

So next the annealing temperature was set to 61^0C to see whether the bands become more distinct. However, this time there were no DNA bands found in any lane for any sample. There were very hazy lines that could be seen around the 50bp to 100bp range. However, it could not be ensured whether they were the desired products from the PCR. Moreover, as the 61^0C annealing temperature did not give results, the PCR reaction with 62^0C annealing temperature was not performed. The Figure 3.12 shows the result of the PCR with 61^0C annealing temperature.

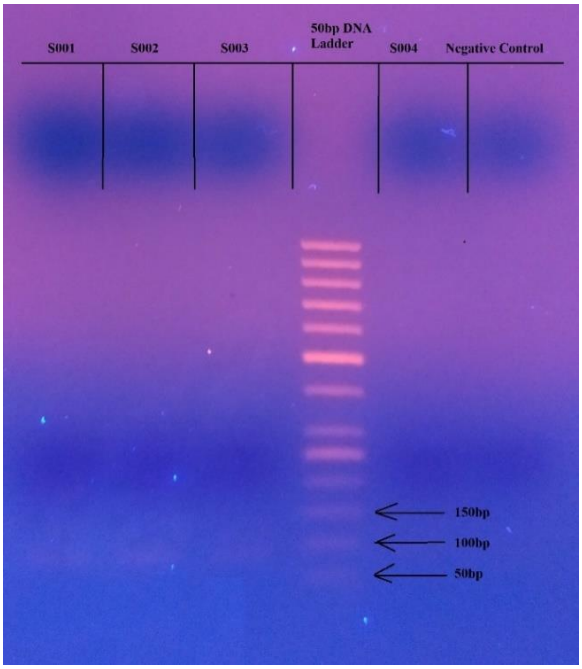


Figure 3.12: The gel electrophoresis result shows that no DNA bands were obtained after the annealing temperature was increased to 61^0C whereas other conditions remained unchanged.

The previous result after increasing the annealing temperature was not satisfactory. As a result, the annealing temperature was decreased to 59^0C to see whether distinct DNA bands could be obtained. However, the result was similar to the result of the previous PCR reaction where no DNA bands could be obtained. The Figure 3.13 shows the result of the PCR reaction.

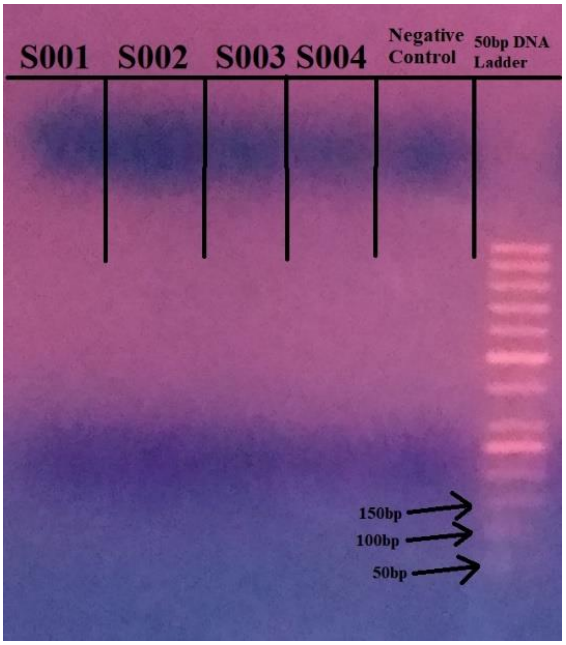


Figure 3.13: The gel electrophoresis result shows that no DNA bands were obtained after the annealing temperature was decreased to 59^0C whereas other conditions remained unchanged.

The previous PCR reactions showed that only the 60^0C annealing temperature had results. So, for the PCR reaction, 60^0C annealing temperature was used. The conditions for PCR reaction mixture and thermal cycler were changed according to table 2.3 and figure 2.4 respectively. The preparation was taken much more cautiously to avoid any type of contamination. The result was similar to the first attempt where 2 samples gave faint bands of both 81bp and 184bp size whereas the other 2 samples did not show any DNA bands. There was no DNA band in the negative control lane which proved the absence of contamination. The DNA bands were a bit more distinct which was the difference between this result and the previous ones. The Figure 3.14 shows the results of this PCR reaction.

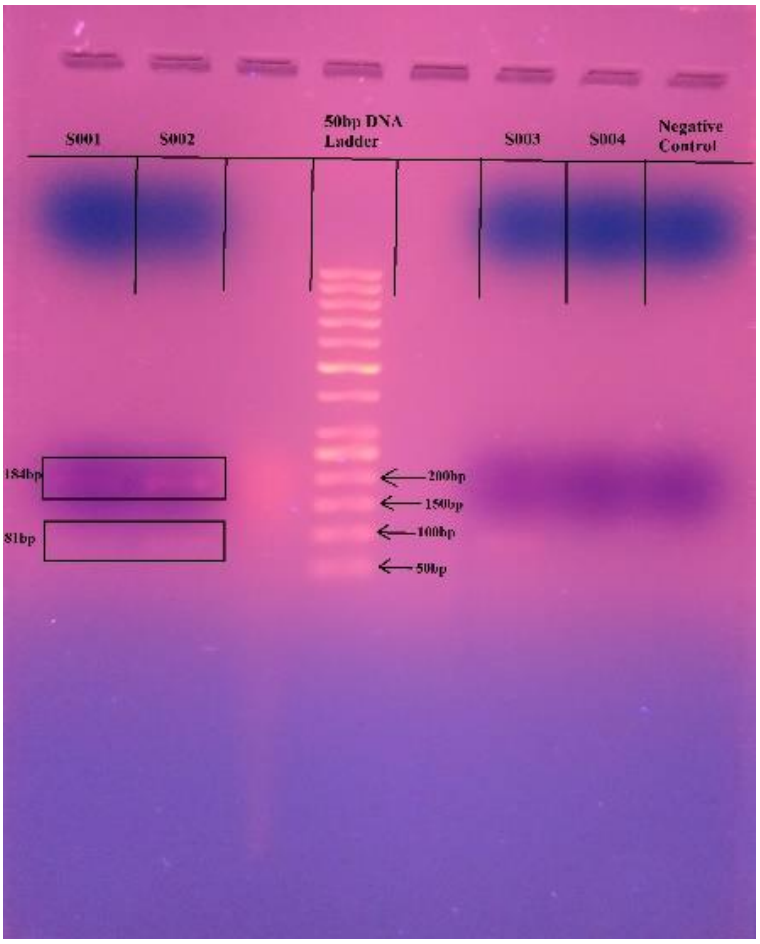


Figure 3.14: The result of the gel electrophoresis performed with 60^0C annealing temperature with slight changes in the PCR reaction recipe and thermal cycler conditions showed bands in 184bp and 81bp.

A new PCR reaction was set up to confirm the PCR conditions of the previous result. This time 10 samples were taken including the previous 4 samples and the last PCR mixture and thermal cycler conditions were used as stated in table 2.3 and figure 2.4. The result this time showed DNA bands in 8 samples in which 6 samples showed faint bands of 81bp size and all

8 samples showed a distinct band in 184bp whereas no samples showed the band of 145bp in size. There were 2 samples with no DNA bands are the samples which did not show any results in the previous PCR reactions as well. The result of the PCR reaction is shown in Figure 3.15.

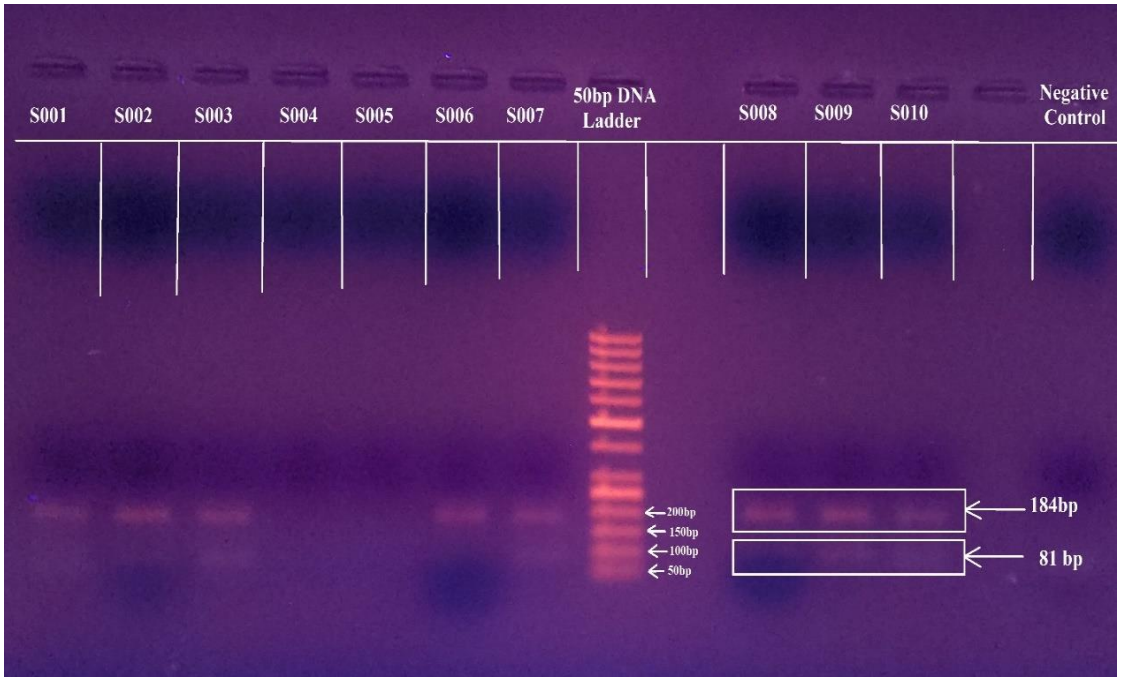


Figure 3.15: The result of the gel electrophoresis performed with the same condition but with more newly extracted samples showed better results than any other previous ones. In this PCR reaction, the results showed the DNA bands of 184bp and 81bp size which confirmed the presence of wild type allele. Two samples named S004 and S005 did not show any results which are discussed in the discussion section.

The results in the last PCR reaction gave a somewhat favorable result. Due to this, the PCR mixture and thermal cycler conditions were finalized to be used for the rest of the samples that were collected for genotyping.

3.2.2 Genotype detection with ARMS-PCR

The genotype detection is done based on the DNA bands that were found in the PCR results. In this experiment, the inner reverse primer was designed to detect wild type alleles and it produced an 81bp DNA band in gel electrophoresis. On the other hand, the inner forward primer was designed to detect mutant allele and it produced a 145bp DNA band in gel electrophoresis. The outer primers givea 184bp band which is not related to the detection of genotype and can only determine the presence of the gene in the sample. In the last PCR reaction with 10 samples, all of the samples with DNA bands had either a 184bp band or a

81bp band. The 6 samples which did not show any 145bp band and only showed 81bp band can be said to have the genotype homozygous wild type. The genotype of the two samples which did not show any bands except 184bp cannot be determined. Moreover, no mutant allele was present in those 10 samples and so no heterozygous or homozygous mutant genotypes were found.

Chapter 5
Discussions

4.1 General discussions

Drug dosage requirement is an important factor for any diseased patients due to the inter-individual drug response caused by genetic variations. A certain dose of a drug can be life-saving when used by an individual but the same dose of that drug can also create complications to another individual which can sometimes be life-threatening. Pharmacogenetics study these variations in the drug response among individuals caused by genetic variations. The drug response in most cases is affected by a single nucleotide change in the gene associated with drug metabolism. This single nucleotide changes also known as single nucleotide polymorphisms or SNPs which is present throughout the DNA in various positions. SNPs occur in almost every 1000 nucleotides and so the total number of SNPs in the human genome are in millions. These SNPs can be unique to a certain individual or can be present in a significant number of individuals in a population. However, not all SNPs are significant as there are more non-coding regions present in the human genome. When an SNP is present in the gene or a regulatory region of a gene, that SNP can play a significant role in altering the activity of the protein that is translated from that particular gene. The result can be a genetic disease or an alteration in drug response. The SNP is even more significant when it is present in the active site of a protein that has direct interaction with a drug. A drug is meant to be present in the body for a short period of time and then metabolized by a specific protein to be excreted out of the body. However, due to the alteration of protein activity, the drug can be metabolized at a slower rate which will result in the drug being present in the body much longer than intended. This can result in creating complications making a patient's condition worse rather than healing him. For this reason, the clinically significant SNPs have been a topic of interest among scientists and researchers. There are various methods for finding a SNP in the genome of an individual like RFLP (Restriction Fragment Length Polymorphism), SNP microarray, TaqMan assay, Sanger sequencing, ARMS-PCR, Allele-Specific PCR, etc. These techniques each require specific reagents and materials. In this experiment, the ARMS-PCR method was used to genotype the clinically significant SNP in the gene CYP2C9 due to its simplicity having the required only minimum reagents and materials which could be obtained in any molecular biology and diagnostics lab.

There were a few objectives of this experiment which was already mentioned earlier and the first objective was to determine a significant SNP in the CYP2C9 gene in relation to the drug warfarin. This drug is being used over six decades for the prevention and treatment of various complications in cardiovascular diseases (Dunmount *et al*., 2013) (Baillargeon *et al.*, 2012).

The gene CYP2C9 has a great number of SNPs throughout the gene which can be found in the NCBI database and the method was described in Section 2.1.1.1. However, as said earlier that all SNPs are not significant, all these SNPs found in the NCBI database are not equally significant. As a result, the significant SNP had to be determined before going to the in vitro experiment where genotyping of the SNP was involved. So, the first approach of the literature review was taken. Studies show that CYP2C9*2 (rs1799853) and CYP2C9*3 (1057910) are poor metabolizers of warfarin and have importance in the dosage requirement (Steward *et al*., 1997) (Takahashi & Echizen, 2001). The importance of the 2 SNPs could be easily understood from the different studies. After the selection of the 2 SNPs, another approach was taken to determine a clinically significant SNP in the CYP2C9 gene which is the use of bioinformatic analysis. As previously mentioned, a SNP in the active site of a protein can lead to significant changes in drug response. Thus, the possible SNPs in the active site were found out and were selected for the *in silico* analysis along with the two SNPs which were previously selected through the literature review. The analysis was performed and found CYP2C9*3 as a significant SNP in drug response. The previous studies on this variant of the gene showed that the SNP was responsible for the poor metabolism of warfarin. This can cause the drug to stay inside the body for a longer period of time than necessary. Warfarin is an anticoagulant which prevents the formation of the blood clot. When this drug is used in a normal dose, different complications caused due to the blood clot formation from a disease can be treated. However, when the drug stays inside the body for a longer period of time due to the presence of the SNP, the drug can cause adverse effects. In different studies, the drug was found to be responsible for hypersensitivity, jaundice, coumarin induced skin necrosis, hemorrhagic effects, etc. which can have serious effects depending on the patients (Baillargeon *et al.*, 2012) (Juurlink, 2007) (Mc & Swinson, 2012). So, a normal dose might be detrimental to the individuals having this SNP and a different dose might be necessary to be prescribed to these patients. There are studies in the populations of the United Kingdom, Brazil, Egypt among different populations which concluded that the variants of the CYP2C9 gene are important in the dose requirement of warfarin among others (Parra *et al.*, 2015) (Shahin *et al.*, 2011) (Bourgeois *et al.*, 2016). In the Bangladeshi population, a study conducted on 87 heart valve replacement patients showed that 5.75% of the patients had the CYP2C9*3 variant which is a very significant number of individuals whereas no patients had the CYP2C9*2 variant (Chowdhury *et al.*, 2017). Thus, the CYP2C9*3 variant of the gene with respect to the population of our country is very significant.

The genotyping of a SNP can be done using various methods as mentioned earlier. It was also mentioned that each technique requires specific reagents and instruments. For example, RFLP requires the use of PCR amplification of the specific region then the usage of restriction enzymes which requires very specific conditions to maintain, and finally separation of the DNA bands in gel electrophoresis before laser detection. This process was used in different studies including the study in the Bangladeshi population (Chowdhury *et al.*, 2017). This process is very tedious and required skilled personnel to perform the whole technique. Moreover, the whole process is very expensive and takes a long time to complete. There are other processes like the TaqMan assay which was used in the study of CYP2C9 in the Egyptian population (Shahin *et al.*, 2011). This process requires fluorescent probes and instruments which is expensive. Another process is using a molecular beacon which can be automated and is a swift process to detect a SNP but the technique requires real-time PCR which is not available everywhere. Another process that can be considered one of the best ways to detect SNP is sequencing. The Sanger sequencing method was the first commercial DNA sequencing method which is very useful in sequencing short DNA segments but the sequencing process requires expertise and a sequencer machine that is not available in every laboratory (Lee *et al.*, 1992). The Next Generation Sequencing is another sequencing method that is fast and inexpensive compared to other SNP detection techniques. However, it still requires special instruments and expensive in developing countries like Bangladesh. These various requirements make a routine diagnosis of the SNP very hard in different laboratories of our country. After taking all this into account, ARMS-PCR was selected in this study for the detection of SNP. The only limitation of this technique is that the detection of multiple SNPs using a single PCR tube makes the process very complex and the PCR might fail. However, the detection of a single SNP can be done with high specificity. The process only requires thermal cycler and gel electrophoresis apparatus which is present in every molecular biology and diagnostics lab. The reagents required are also readily available in these labs. This makes the process fast, inexpensive, easy, and not very high expertise is necessary. These all advantages are very suitable to be used in any laboratory of our country which is why it was selected.

After the selection of the SNP and the process was completed, the process was optimized and was going to be used on the 50 samples that were obtained from volunteers. However, the validation of the results through sequencing and genotyping of the samples could not be

completed due to the current pandemic situation and so the study was written with all the data available from the portion of the experiment that is completed.

Overall in this experiment, it was tried to find the most significant SNP in the gene CYP2C9 and tried to confirm the prevalence of the SNP in a highly variable sample from different origins in the population. Moreover, a fast and expensive method for the diagnosis of the SNP was tried to be established in this experiment which is most suitable to be used in our country.

4.2 *In Silico* analysis

The *in silico* analysis was performed for the determination of the most significant SNP in the CYP2C9 gene which also has been mentioned earlier. The SNPs for the analysis were selected based on the hypothesis that a change in amino acid in the active site of the protein can cause a change in the protein-ligand interaction. However, as all the changes in amino acid were not always significant, a selection of the SNPs present in the amino acid in the active site was done as shown in the table 3.1. The other two SNPs which are CYP2C9*2 and CYP2C9*3 were analyzed due to the significant number of cohort studies performed on them where they found these variants of the gene to be a poor metabolizer of the drug warfarin. These two SNPs are in the 144^{th} and 359^{th} positions in the amino acid chain which are not present in the active site but due to the extensive research and evidence on these SNPs, they were selected. After the selection, the next steps were carried out in the analysis.

In the *in silico* analysis, the objective was to determine whether there is any change in the protein-ligand interaction with the presence of a SNP. To observe the change, it was required to have a 3D model of the proteins which contain the SNPs. These models are not available in databases and were required to be modelled. The EasyModeller4.0 was selected for modelling which is a validated tool used in various works by different researchers and was cited many times. The tool uses similar protein models as a template to design a new protein from a sequence. Thus, homologous protein models were obtained using the sequence of the CYP2C9 protein in a database before starting the modelling of new proteins. The most similar proteins were selected including the model used so the new models could still contain the conserved residues and do not deviate too much from the wild type protein. The tool was set in a way that 3 different protein models were obtained after each modelling was completed. After the protein modelling was completed, the quality of the three modelled proteins was to be checked as all the proteins might not be of good quality and only one best

protein from 3 selected protein models were to be selected for the rest of the experiments. The categories selected to determine the quality of the proteins were Verify3D, ERRAT, and Procheck. The Verify3D determines the compatibility of the protein model with its amino acid sequence and comparing the results with good protein structures to determine the quality of the query protein (Luthy *et al.*, 1992). The ERRAT statistically examines the regions of proteins to distinguish the correct and incorrect regions which are based on the atomic interactions in proteins by comparison with a validated database of proteins by using a quadratic error function (Coloyos & Yeates, 1993). The Procheck examines the residue by residue geometrical structure of the protein to determine its quality (Laskowski *et al.*, 1993). By using these 3 categories, the best protein model for each SNP was selected.

The observation protein-ligand interaction through docking was the next part of the analysis. To dock the protein and ligand, it was required to make both the molecules structurally stable and to do this the energy minimization step was performed to reduce the free energy of the molecules in the protein and the ligand. PyRx was used to dock ligand and protein and the tool tried to dock the ligand in 9 places in the molecule in the total region of the protein. The binding affinity data were obtained from the software which was saved. The positive control was also prepared through this as stated in the methodology. The positive control easily coincided with the wild type model 1OG5 but the other models could not coincide. It might have happened due to the difference in the algorithm used to model the proteins with the wild type model or maybe the geometric space used to model the proteins might be different from the original model due to which the modelled proteins could not be overlapped on the original protein. However, when the modelled proteins were checked visually, the ligands were found in the active site of the protein and thus the best docking could be selected as most of the time ligand docked on a site which was not the active site of the protein. Then, the binding affinity data were compared where it was found that the variants CYP2C9*2 and CYP2C9*3 had a 15% change in the binding affinity which was the highest among all the selected variants. Although the binding affinity data from a bioinformatic tool cannot be 100% valid, the data still contains significance and used in important processes like drug designing. Then from these last two SNPs, the rs1057910 was selected due to the prevalence of it in Bangladesh which was found in a study, and the other SNP was not found. Thus, the CYP2C9*3 or rs1057910 was selected for the in-vitro experiment.

The allele-specific primer designing for the ARMS-PCR technique to detect the CYP2C9*3 was done by using the BatchPrimer3 v1.0. The setting of the primer was optimized in a way

so that the designed primer retains all the ideal conditions for a primer. The ideal conditions were also selected when the final primer was selected. In ideal conditions, the length of the primer has to be 18-25bp and the GC content has to be between 40% to 60% (Dieffenbach *et al.*, 1993). The selected primers had a length of 22bp for all except the inner reverse primer which was 21bp and they were all within the ideal range. The GC content was also almost within the ideal range with a few percentages more for the inner primers. This was due to one of the major limitations in the ARMS-PCR technique which is the region of the inner primers is fixed. As a result, the designed primers have GC content corresponding to the region around the specific allele. However, the slight increase in the GC content was not a major factor that could alter the result of the PCR and so it was acceptable. Another major parameter that could potentially cause the failure of the PCR in the difference in melting temperature (Tm) between the primers which should be $<5^0$C. The highest difference in the designed primer was approximately 3^0C which was well within the ideal range. The other factors were also checked to be within the ideal range and as a result, the primers were found suitable to be used ARMS-PCR technique.

4.3 Optimization of PCR conditions

The optimization of PCR started with setting the annealing temperature for the thermal cycler which was decided to be 60^0C, 61^0C, and 62^0C based on the annealing temperature was based on the Tm of the primers. The regular laboratory PCR recipe was used to prepare the PCR mixture and regular thermal cycler conditions were used for all the temperatures except annealing temperature. The initial attempt was taken using all the conditions mentioned in table 2.2 and figure 2.3. Due to the unavailability of gradient PCR, the PCR reaction had to be done separately for each annealing temperature. The result of the initial attempt with 60^0C annealing temperature was not so satisfactory as the bands obtained were very faint in lane 2 and 3 whereas there was no result in lane 1 and 4. This could have happened due to various reasons. The possible reasons might be the amount of sample used in gel electrophoresis might be very small which resulted in a low concentration of the PCR product. Another reason can be the very low concentration of the template DNA used in the PCR or the time for the annealing might have been insufficient for which the concentration of the final product was very low. The other significant reason can be that the annealing temperature of 60^0C was not optimum. Thus, the next attempt was taken with 61^0C annealing temperature which resulted in a failure. This might have been caused due to the unsuitability of the annealing temperature and so increasing the temperature more would have given the same

result and so the annealing temperature of 62^0C was not used. The next attempt was taken using an annealing temperature of 59^0C and the result was also a failure. These results somewhat showed the suitability of the annealing temperature of 60^0C in the PCR reaction. So, by keeping that annealing temperature fixed, other conditions were changed as shown in table 2.3 and figure 2.4 to see whether the PCR reaction showed any results. The result showed faint bands in lane 1, 2 and 3 but no band in lane 4. However, the result this time was a bit clearer than the previous PCR reaction with 60^0C annealing temperature. This somewhat showed that the concentration of the initial DNA template might have been the reason for the faint bands. Thus, 10 newly extracted DNA samples were used to confirm the PCR condition. This time most of the samples showed more distinct bands except 2 samples in lane 4 and 5 which might have been caused due to the failure to add all reagents properly or the initial DNA template might have been degraded through any means. After the PCR reaction result was somewhat satisfactory, the conditions were finalized to be used for genotyping the rest of the DNA samples.

4.4 Genotyping individual samples

There were 50 samples that were obtained from random individuals with their consent which included the 10 samples used in the last PCR. The samples were highly variable due to each individual originating from different parts of the country and genotyping of these samples was necessary to assess the prevalence of the allele and comparing the data with the previous study by other researchers. However, it was not possible due to the closing of all the educational institutes of the country which included our university. The 10 samples which were used in the PCR reaction had results in 8 samples where 2 bands of 184bp and 81bp could be observed. The 81bp band confirmed the presence of the wild type allele in the sample and the absence of 145bp band showed that no mutant allele was found in any of the 10 samples. This data did not show the full scenario of the samples and to observe the prevalence of the SNP in the population, more samples were required to be genotyped which could be compared with other studies.

Conclusion

The pharmacogenetic studies show an important relation between the drug and a gene which should be considered highly before prescribing a drug to a patient. The use of this information should be more widespread for the treatment procedure to be much better and the end goal for these researches is towards personalized medicine. The use of genotyping can help the treatment procedure to be more controlled and better for the patients.

In this study, there were some incomplete parts at the end but the important aspects were mostly covered. Most of the objectives were covered except assessing the prevalence of the SNP in the population but with the PCR conditions optimized, this data can be obtained in the future.

The ARMS-PCR method which is exclusively designed for the genotyping of SNP. This process requires reagents that are readily available in the lab and also very time efficient. The cost of diagnosis using this method is very affordable and thus this technique is encouraged to be used in genotyping. The overall procedure of genotyping of the CYP2C9*3 was established in the study and it is hoped that further studies would be conducted in the future to complete the remaining objectives that could not be fulfilled.

Bibliography

1. Ahsan Karar, Z., Alam, N., & Kim Streatfield, P. (2009). Epidemiological transition in rural Bangladesh, 1986-2006. *Global health action*, 2, 10.3402/gha.v2i0.1904. https://doi.org/10.3402/gha.v2i0.1904
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., &Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
3. Baillargeon, J., Holmes, H. M., Lin, Y. L., Raji, M. A., Sharma, G., &Kuo, Y. F. (2012). Concurrent use of warfarin and antibiotics and the risk of bleeding in older adults. *The American journal of medicine*, 125(2), 183–189. https://doi.org/10.1016/j.amjmed.2011.08.014.
4. Bangladesh Government Health bulletin. (2013). Retrieved from: https://dghs.gov.bd/images/docs/Other_Publication/HB%202013%20final%20-%20Full%20version%201March14.pdf.
5. Bangladesh Government Health bulletin. (2018). Retrieved from: https://dghs.gov.bd/images/docs/Publicaations/HB%202018%20final.pdf.
6. BIOvIA, D. S. (2015). Discovery studio modeling environment. San Diego, DassaultSystemes, Release, 4.
7. Bourgeois, S., Jorgensen, A., Zhang, E. J., Hanson, A., Gillman, M. S., Bumpstead, S., Toh, C. H., Williamson, P., Daly, A. K., Kamali, F., Deloukas, P., &Pirmohamed, M. (2016). A multi-factorial analysis of response to warfarin in a UK prospective cohort. *Genome medicine*, 8(1), 2. https://doi.org/10.1186/s13073-015-0255-y
8. Chowdhury, Z. S., Shahjin, F., Akter, F., Ahmed, M., Islam, M. S., Bin Sayeed, M. S., Islam, R., &Hasnat, A. (2017). Effect of VKORC1 and CYP2C9 polymorphisms on warfarin dose requirement in Bangladeshi population. *Pakistan journal of pharmaceutical sciences*, 30(2), 341–346.
9. Chowdhury, Z. S., Shahjin, F., Akter, F., Ahmed, M., Islam, M. S., Bin Sayeed, M. S., Islam, R., &Hasnat, A. (2017). Effect of VKORC1 and CYP2C9 polymorphisms on warfarin dose requirement in Bangladeshi population. Pakistan journal of pharmaceutical sciences, 30(2), 341–346.

10. Colovos, C., &Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein science : a publication of the Protein Society*, 2(9), 1511–1519. https://doi.org/10.1002/pro.5560020916
11. Daly, A. K., Rettie, A. E., Fowler, D. M., & Miners, J. O. (2017). Pharmacogenomics of CYP2C9: Functional and Clinical Considerations. Journal of personalized medicine, 8(1), 1. https://doi.org/10.3390/jpm8010001
12. D'Antonio, C., Milano, A., Righini, R., Onesti, C. E., Bassanelli, M., Falcone, R., Paris, I., Lauro, S., &Marchetti, P. (2014). Pharmacogenomics in lung cancer chemotherapy: a review of what the oncologist should know. *Anticancer research*, 34(10), 5241–5250.
13. DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*, 40, 82-92.
14. Dieffenbach, C. W., Lowe, T. M., &Dveksler, G. S. (1993). General concepts for PCR primer design. *PCR methods and applications*, 3(3), S30–S37. https://doi.org/10.1101/gr.3.3.s30
15. Dumont, Z., Mordasiewicz, M., Kosar, L., & Schuster, B. (2013). Warfarin: its highs and lows. *Canadian family physician Medecin de famillecanadien*, 59(8), 856–860.
16. Erichsen, H. C., &Chanock, S. J. (2004). SNPs in cancer research and treatment. *British journal of cancer*, 90(4), 747–751. https://doi.org/10.1038/sj.bjc.6601574
17. Esmon, C. T., Vigano-D'Angelo, S., D'Angelo, A., & Comp, P. C. (1987). Anticoagulation proteins C and S. *Advances in experimental medicine and biology*, 214, 47–54. https://doi.org/10.1007/978-1-4757-5985-3_4
18. Ghatak, S., Muthukumaran, R. B., &Nachimuthu, S. K. (2013). A simple method of genomic DNA extraction from human samples for PCR-RFLP analysis. *Journal of biomolecular techniques : JBT*, 24(4), 224–231. https://doi.org/10.7171/jbt.13-2404-001
19. Guex, N. and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18, 2714-2723.
20. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000). *The Protein Data Bank Nucleic Acids Research*, 28: 235-242.

21. Hosseinkhani, Z., Sadeghalvad, M., Norooznezhad, F., Khodarahmi, R., Fazilati, M., Mahnam, A., Fattahi, A., &Mansouri, K. (2018). The effect of CYP2C9*2, CYP2C9*3, and VKORC1-1639 G>A polymorphism in patients under warfarin therapy in city of Kermanshah. *Research in pharmaceutical sciences*, 13(4), 377–384. https://doi.org/10.4103/1735-5362.235165

22. Issa, N. T., Wathieu, H., Ojo, A., Byers, S. W., &Dakshanamurthy, S. (2017). Drug Metabolism in Preclinical Drug Development: A Survey of the Discovery Process, Toxicology, and Computational Tools. *Current drug metabolism*, 18(6), 556–565. https://doi.org/10.2174/1389200218666170316093301

23. Jack Hirsh, M. D. (1991). Oral Anticoagulant Drugs. *The New England Journal of Medicine*,324:1865-1875. https://doi.org/10.1056/NEJM199106273242606

24. Juurlink D. N. (2007). Drug interactions with warfarin: what clinicians need to know.*CMAJ : Canadian Medical Association journal = journal de l'Associationmedicalecanadienne*, 177(4), 369–371. https://doi.org/10.1503/cmaj.070946.

25. Kaminsky, L. S., & Zhang, Z. Y. (1997). Human P450 metabolism of warfarin. *Pharmacology & therapeutics*, 73(1), 67–74. https://doi.org/10.1016/s0163-7258(96)00140-4

26. Khanam, F., Hossain, M. B., Mistry, S. K., Afsana, K., & Rahman, M. (2019). Prevalence and Risk Factors of Cardiovascular Diseases among Bangladeshi Adults: Findings from a Cross-sectional Study. *Journal of epidemiology and global health*, *9*(3), 176–184. https://doi.org/10.2991/jegh.k.190531.001

27. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033

28. Kuntal, B. K., Aparoy, P., &Reddanna, P. (2010). EasyModeller: A graphical interface to MODELLER. *BMC research notes*, 3, 226. https://doi.org/10.1186/1756-0500-3-226

29. Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst.*, **26**, 283-291. https://doi.org/10.1107/S0021889892009944

30. Lee, L. G., Connell, C. R., Woo, S. L., Cheng, R. D., McArdle, B. F., Fuller, C. W., Halloran, N. D., & Wilson, R. K. (1992). DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic acids research*, 20(10), 2471–2483. https://doi.org/10.1093/nar/20.10.2471
31. Limdi, N. A., &Veenstra, D. L. (2008). Warfarin pharmacogenetics. *Pharmacotherapy*, 28(9), 1084–1097. https://doi.org/10.1592/phco.28.9.1084
32. Lu, D. Y., Lu, T. R., Xu, B., & Ding, J. (2015). Pharmacogenetics of cancer therapy: breakthroughs from beyond?.*Future science OA*, 1(4), FSO80. https://doi.org/10.4155/fso.15.80
33. Lüthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364), 83–85. https://doi.org/10.1038/356083a0
34. Mc, A. G., &Swinson, B. (2012). An interesting potential reaction to warfarin. *Dental update*, 39(1), 33–37. https://doi.org/10.12968/denu.2012.39.1.33
35. McDonnell, A. M., & Dang, C. H. (2013). Basic review of the cytochrome p450 system. *Journal of the advanced practitioner in oncology*, 4(4), 263–268. https://doi.org/10.6004/jadpro.2013.4.4.7
36. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/
37. Ohadoma, S. C., Lawal, B. A. S, Odoala, F. K., &Eban, L. K. Distinguishing Between Pharmacogenomics And Pharmacogenetics: How They Guide The Choice Of Drug Therapy. *European Journal Of Pharmaceutical And Medical Research*, 6(6), 234-240.
38. Olvera Lopez, E., Ballard Brian, D., Jan, A. (2020). Cardiovascular Disease. *StatPearls*. Treasure Island (FL).
39. Parra, E. J., Botton, M. R., Perini, J. A., Krithika, S., Bourgeois, S., Johnson, T. A., Tsunoda, T., Pirmohamed, M., Wadelius, M., Limdi, N. A., Cavallari, L. H., Burmester, J. K., Rettie, A. E., Klein, T. E., Johnson, J. A., Hutz, M. H., & Suarez-Kurtz, G. (2015). Genome-wide association study of warfarin maintenance dose in a Brazilian sample. *Pharmacogenomics*, 16(11), 1253–1263. https://doi.org/10.2217/PGS.15.73

40. Patel, S., Singh, R., Preuss, C. V., & Patel, N. (2020). Warfarin. *StatPearls*. Treasure Island (FL).
41. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., &Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13), 1605–1612. https://doi.org/10.1002/jcc.20084
42. Pontius, J., Richelle, J., &Wodak, S. J. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of molecular biology*, 264(1), 121–136. https://doi.org/10.1006/jmbi.1996.0628
43. Rosdi, R. A., MohdYusoff, N., Ismail, R., Soo Choon, T., Saleem, M., Musa, N., &Yusoff, S. (2016). High allele frequency of CYP2C9*3 (rs1057910) in a Negrito's subtribe population in Malaysia; Aboriginal people of Jahai. *Annals of human biology*, 43(5), 445–450. https://doi.org/10.3109/03014460.2015.1068372
44. Shahin, M. H., Khalifa, S. I., Gong, Y., Hammad, L. N., Sallam, M. T., El Shafey, M., Ali, S. S., Mohamed, M. E., Langaee, T., & Johnson, J. A. (2011). Genetic and nongenetic factors associated with warfarin dose requirements in Egyptian patients. *Pharmacogenetics and genomics*, 21(3), 130–135. https://doi.org/10.1097/FPC.0b013e3283436b86
45. Smith, C. J., Ryckman, K. K., Bahr, T. M., &Dagle, J. M. (2017). Polymorphisms in CYP2C9 are associated with response to indomethacin among neonates with patent ductus arteriosus. *Pediatric research*, 82(5), 776–780. https://doi.org/10.1038/pr.2017.145
46. Steward, D. J., Haining, R. L., Henne, K. R., Davis, G., Rushmore, T. H., Trager, W. F., &Rettie, A. E. (1997). Genetic association between sensitivity to warfarin and expression of CYP2C9*3. *Pharmacogenetics*, 7(5), 361–367. https://doi.org/10.1097/00008571-199710000-00004
47. Takahashi, H., &Echizen, H. (2001). Pharmacogenetics of warfarin elimination and its clinical implications. *Clinical pharmacokinetics*, 40(8), 587–603. https://doi.org/10.2165/00003088-200140080-00003
48. The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515. https://doi.org/10.1093/nar/gky1049
49. Trott, O., & Olson, A. J. (2010). AutoDockVina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and

multithreading. Journal of computational chemistry, 31(2), 455–461. https://doi.org/10.1002/jcc.21334

50. W.C.R.C.W. Group. (2019). World Health Organization cardiovascular disease risk charts: Revised models to estimate risk in 21 global regions, *Lancet Glob Health*, 7 (10).

51. You, F. M., Huo, N., Gu, Y. Q., Luo, M. C., Ma, Y., Hane, D., Lazo, G. R., Dvorak, J., & Anderson, O. D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC bioinformatics*, 9, 253. https://doi.org/10.1186/1471-2105-9-253

52. Zorina-Lichtenwalter, K., Lichtenwalter, R. N., Zaykin, D. V., Parisien, M., Gravel, S., Bortsov, A., & Diatchenko, L. (2019). A study in scarlet: MC1R as the main predictor of red hair and exemplar of the flip-flop effect. *Human molecular genetics*, 28(12), 2093–2106. https://doi.org/10.1093/hmg/ddz018