Syrian Arab Republic

Ministry of Higher Education

Syrian Virtual University

Master of Bioinformatics

In silico detection for Beta Thalassemia via bioinformatics and expert systems

A Project submitted for the Master's degree of Bioinformatics

Presented by:

Student name : Rand khayata *Username : rand_175348*

Supervised by:

Dr. Yasser khadra

2023

Table of Contents

Index of Tables

Index of Figures

Abbreviation list

Abbreviation	The meaning
A	Adenine
AS-PCR	Allele-Specific PCR
β^0	Complete absence of beta globin on the affected allele
β^+	Residual production of beta globin (around 10%)
BMP	Bone morphogenetic protein
BMs	Predicted binding modes
BMT	Bone marrow transplantation
C	Cytosine
FL	Femtoliter
G	Guanine
Hb	Hemoglobin
HbA	Adult hemoglobin
HbA2	Adult hemoglobin minor
HbF	Fetal hemoglobin
HPLC	High-performance liquid chromatography
Kbp	Kilo base pairs
MCH	Mean corpuscular hemoglobin
MCV	Mean corpuscular volume
MW	Molecular weight
NTDT	Non-transfusion dependent thalassaemia
PCR	Polymerase chain reaction
Pg	Picogram
RBC	Red blood cells
SNP	Single Nucleotide Polymorphism
T	Thymine
TDT	Transfusion dependent thalassaemia
TI	β-thalassemia intermedia
TM	β-thalassemia major
UTR	Untranslated region

Summary

Background: Beta-thalassemia is one of most common autosomal recessive disorders worldwide, characterized by reduced or absent beta globin chain synthesis, resulting in reduced Hb in red blood cells (RBC), decreased RBC production and anemia. Recently, the number of thalassemia patients has been increasing in other regions worldwide, but there is a lack of comprehensive knowledge regarding the epidemiologic profile of thalassemia in these regions. This high prevalence of thalassemia makes it one of the major health problems and a priority genetic disease. A prevention program would be useful to overcome these problems, but it requires a preliminary knowledge of hemoglobin, disease pathophysiology, as well as a spectrum of globin gene mutations among different populations. Independently, new studies are needed to validate the clinical consequences of the mutations with undefined pathogenicity. Considering the absence of physiopathological knowledge relative to the newly identified mutations, the use of in silico predictors emerges as a possible tool to aid in decision-making with respect to diagnostic, preventative, and treatment measures .

Aim of the study: Providing a guide to choose the most efficient way to design a new specific-primer by applying web services on SNPs from the HbVar database to understand the relationship between phenotype and genotype in the clinical setting and investigating the effects of SNP mutations in the HBB exons and give a guideline for functional studies and prenatal diagnosis to be developed as basis for future studies , Finding alternative therapeutic molecules made from natural inducers that had fewer side effects than traditional medications for treating beta thalassemia by recognizing the particular ligands that bind to specific receptor binding sites and recognize the foremost favorable ligand with the assistance of molecular docking , and creating a fuzzy inference system to predict the severity involve in Thalassemia disease.

Results: Allele-specific primers corresponding to 37 kinds of SNP in β thalassemia were designed using primer program BatchPrimer3 v1.0. Single assays were performed for the specific amplification of each target allele using the tested primers , and all amplicons were between 100 - 300 bp fragments in which containing the corresponding SNPs by Silica tool . PCR products were detected on 2.5% agarose gel by insilico electrophoresis via SnapGene tool . The results showed set of primers selected that have a unique sequence within the template DNA, optimal melting temperatures , appropriate primer length, suitable GC content and Specificity for allele pairs for each SNPs . These primers result refers to 8 SNPs in studied dataset . Also , showed set of primers selected that have a Specificity for only one allele for each SNPs . These primers result refers to 21 SNPs in

studied dataset. For the simulated results of electrophoresis on agarose gel , All 18 bands that showed were identical with Silica results . SOPMA tool was used to predicted Secondary structures of HBB protein , The sequence length was 147 amino acids ,a considerable prediction was observed in mainly alpha helix classes of protein by 62.59% and less prediction for beta sheet (extended strand) by 9.52% , in addition random coils and beta turn were found by 21.77% , 6.12 % respectively. Molucular docking via SwissDOCK was used to predict the molecular interactions that may occur between a target protein and a small molecule. The results supported usage of Coptidis Rhizome for the treatment of thalassemia and its related conditions. Among the bioactive compounds, Berberine was determined to be best for treatment of thalassemia compared to the already available drugs available with binding energy of -7.57 Kcal/mol because a better docking score corresponds to Low Binding Energy. Fuzzy Inference System was developed in order to analyze the severity of Thalassemia disease using Fuzzy Logic Toolbox in Matlab by developing 26 if-then rules. The fuzzy system result shows that the selected inputs such as MCH, MCV and HGB are suitable for the study. , It was found that our program matched the doctor's diagnosis in 608 cases perfectly from 646 cases . This results with an accuracy of about 94.11 % which is more stable than the results obtained from the similar contents in (Thakur et al., 2016) with accuracy 83% . Based on the results presented in the confusion matrix , it was discovered that there were 608 correct classifications (175 for low, 405 for moderate and 28 for high risk-along the diagonal) .

Conclusion: Single nucleotide polymorphisms (SNPs) have been proposed as the next generation of markers to identify loci associated with complex diseases and their therapeutic treatment . Low-cost genotyping tools are absolutely necessary for effective personalized medicine ,so the in silico analysis like AS-PCR methods are quick, excellent and inexpensive strategies and require minimal instruments that are found in most laboratories to be developed for massive implementation into clinical laboratories . we hope that identify the mechanisms responsible for fetal hemoglobin control, since reactivation of fetal hemoglobin can provide major therapeutic benefits to people affected by β-hemoglobinopathies .

Keywords : Beta thalassemia **,** SNP , AS-PCR , SOPMA , SwissDOCK , Fuzzy inference system

Research gap:

Recently, the number of thalassemia patients has been increasing in other regions worldwide, including in North America, Northern Europe, and Northeast Asia, due to an increase in migrant populations, but there is a lack of comprehensive knowledge regarding the epidemiologic profile of thalassemia in these regions. This high prevalence of thalassemia makes it one of the major health problems and a priority genetic disease. In contrast, the treatment of thalassemia is entirely different in less developed countries, where most of the patients with this disease require safe transfusion and chelation that are not universally available. Treatment of β-thalassemia still represents a significant drain of the country's resources due to the disease's major complications. Thalassemia encompasses serious diseases with complex pathophysiology that is difficult to explain since it is considered a group of defects with similar clinical effects, still not a single disorder.in addition, a common confounding factor in hemoglobin electrophoresis is a concomitant iron deficiency that masks an underlying beta-thalassemia minor The resultant electrophoresis pattern appears normal.

So, the research gab is a lack of preliminary knowledge of hemoglobin, disease pathophysiology, as well as a spectrum of globin gene mutations among different populations. Independently, new studies are needed to validate the clinical consequences of the mutations with undefined pathogenicity. Considering the absence of physiopathological knowledge relative to the newly identified mutations, the use of in silico predictors emerges as a possible tool to aid in decision-making with respect to diagnostic, preventative, and treatment measures.

Aim of the study :

• Providing a guide to choose the most efficient way to design a new specific-primer by applying web services on SNPs from the HbVar database that cause (β^{0} or β^{+}) genotyping in the exon regions of HBB gene because in silico analysis is faster and easier to execute, yields more results, and costs less, thus making it more efficient , so the application of a combined molecular approach with clinical data and efficient bioinformatics tools will enable understanding the relationship between phenotype and genotype in the clinical setting and investigating the effects of SNP mutations in the HBB exons and give a guideline for functional studies and prenatal diagnosis to be developed as basis for future studies. In addition , allowing couples at risk to make informed decision on their reproductive choices by population screening associated with genetic counseling and effective prevention eventually decreases the number of severely affected patients worldwide , So diagnostic methods based on single nucleotide polymorphism (SNP) biomarkers are essential for the real adoption of personalized medicine .

• Recognizing the particular ligands that bind to specific receptor binding sites and recognize the foremost favorable ligand with the assistance of molecular docking to explore the behavior of small molecules in the binding site of the targeted protein which is used as a very important tool for drug discovery . As a result, the goal of this research was to find alternative therapeutic molecules made from natural inducers that had fewer side effects than traditional medications for treating beta thalassemia .

• creating a fuzzy inference system to predict the severity involve in Thalassemia disease because thalassemia and hemoglobinopathy genotype interpretation and phenotype determination as well as their clinical symptoms and electrophoresis are one of the most complex issues of Hematology. Therefore, establishing a logical digital relationship between the above-mentioned items can contribute to the understating of this issue.

1- Literature review

The term thalassemia is derived from the Greek, thalassa (sea) and haima (blood) [1] . Beta-thalassemia is one of most common autosomal recessive disorders worldwide, characterized by reduced or absent beta globin chain synthesis, resulting in reduced Hb in red blood cells (RBC), decreased RBC production and anemia. Most thalassemias are inherited as recessive traits [1] . Within the red blood cell precursors, when the beta globin chains are reduced or absent, the unassembled alpha chains precipitate and lead to oxidative damage of the cell membrane, thereby resulting in apoptosis (ineffective erythropoiesis) [2].

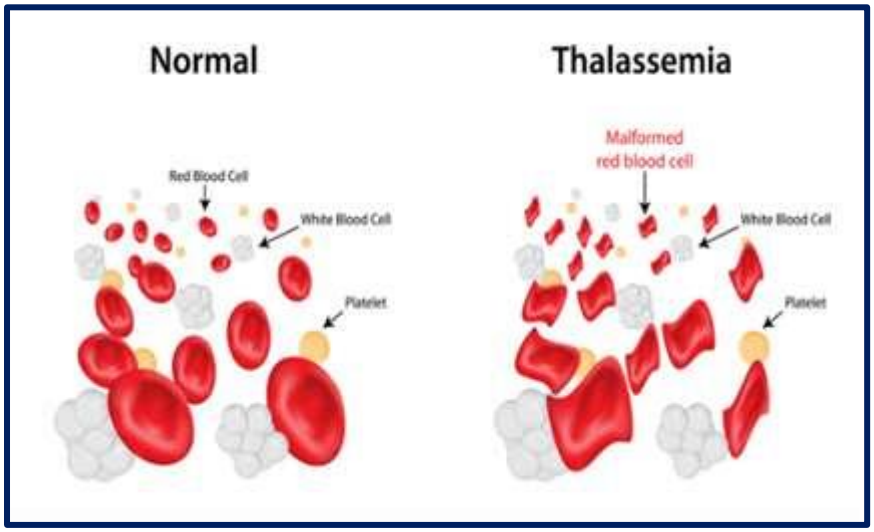


Figure 1 : Abnormal red blood cells in thalassemia

1-1 : Epidemiology

High prevalence is present in populations in the Mediterranean, Middle-East, Transcaucasus, Central Asia, Indian subcontinent, and Far East. It is also relatively common in populations of African descent.[2] However, as a result of mass migrations of populations from high-prevalence areas, thalassemias are now encountered in most countries, including the United States, Canada,

Australia, South America, and North Europe [5] .It has been estimated that about 1.5% of the global population (80 to 90 million people) are carriers of beta-thalassemia, with about 60,000 symptomatic individuals born annually, the great majority in the developing world. The total annual incidence of symptomatic individuals is estimated at 1 in 100,000 throughout the world and 1 in 10,000 people in the European Union. [1]

Figure 2 : The prevalence of thalassemia in world

1-2 : Pathophysiology

The main types of thalassemia reported on the basis of the type of globin chains are affected, grouped as α, β, δβ, γδβ, δ, γ and εγδβ thalassemia [13].

During early gestation, embryonic hemoglobins (ζ2ε2, α2ε2, ζ2γ2) predominate in erythroid cells in the yolk sac. For the remainder of fetal life, fetal hemoglobin (HbF [α2γ2]) is the main component of red cells produced initially by the spleen and liver and later by the bone marrow. The key switch from γ-globin to β-globin gene expression begins around week 12 of gestation and is completed by 6 months of age, after which the majority

(>95%) of hemoglobin in red cells is adult hemoglobin (HbA [α2β2]), with minor concentrations of HbA2 (α2δ2) and HbF.2 [16] .

Table 1 : Different types of hemoglobin at various developmental stages of human.

Developmental stages	***Hemoglobin***
Embryonic	Hb Grower1 (ζ2ε2)
	Hb Grower2 (α2ε2)
	Hb Portland 1 (ζ2β2)
	Hb Portland 2 (ζ2γ2)
Fetal	HbF (α2γ2)
Adult	**HbA (α2β2)**
	HbA2 (α2δ2)

α : alpha - β : beta – γ : gamma – δ : delta - ε : epsilon – ζ : zeta

Beta-thalassemia is caused by the reduced (beta^{+}) or absent (beta^{0}) synthesis of the beta globin chains of the hemoglobin (Hb) tetramer, which is made up of two alpha globin and two beta globin chains (alpha2beta2) [2] It is working in combination with heme to transport oxygen in the blood. It is iron containing protein, synthesized inside immature erythrocyte in the red bone marrow. The globin polypeptides bind heme molecule, which in turn allows the hemoglobin in erythrocytes to bind oxygen reversibly and transport it from the lungs to other part of body [13]

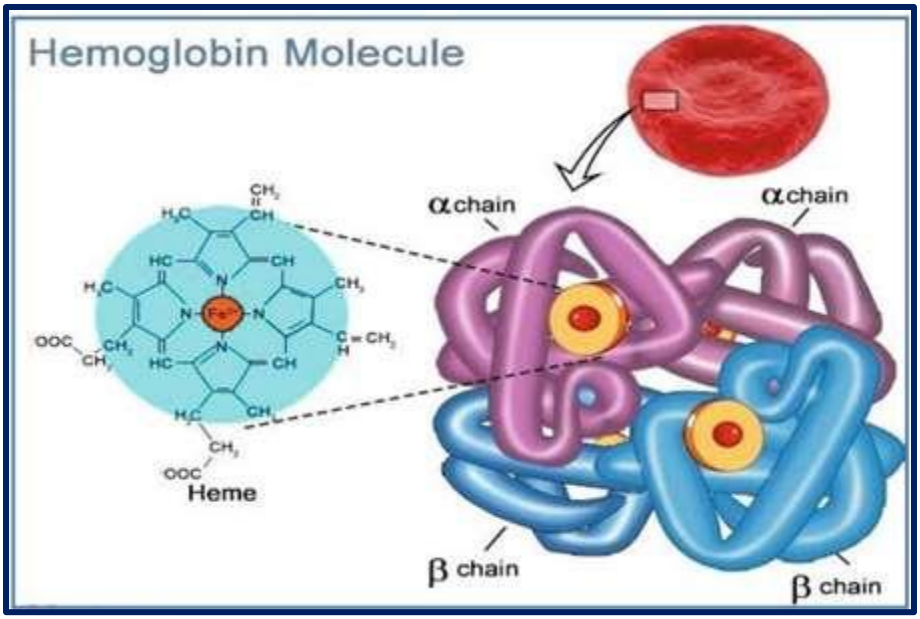


Figure 3 : Structure of hemoglobin

In the alpha (α)-thalassemia, there is reduced production or absence of α-globin subunits, whereas in the beta (β)-thalassemia, there is reduced production of β-globin subunits. The β-thalassemia can be clinically classified according to the degree of severity i.e [13] , β-thalassemia major (TM), also referred to as "Cooley's anemia" and "Mediterranean anemia"; β-thalassemia intermedia (TI); and thalassemia minor, called "β-thalassemia carrier," "β-thalassemia trait," or "heterozygous β-thalassemia." [5] Apart from the rare dominant forms, subjects with TM are homozygotes or compound heterozygotes for β^0 or β^+ genes, subjects with TI are mostly homozygotes or compound heterozygotes, and subjects with thalassemia minor are mostly heterozygotes [5] .

Most pathologic significance in beta-thalassemia major and intermedia, the relative excess alpha chains form insoluble alpha chain inclusions that cause marked intramedullary hemolysis. This ineffective erythropoiesis leads to severe anemia and erythroid hyperplasia with bone marrow expansion and extramedullary hematopoiesis. Biochemical signaling from marrow expansion involving the bone morphogenetic protein (BMP) pathway inhibits

hepcidin production causing iron hyperabsorption. Inadequately treated patients and transfusion-dependent patients are at risk for end-organ damaging iron overload. Hepatosplenomegaly from extramedullary hematopoiesis and ongoing hemolysis also causes thrombocytopenia and hepatic dysfunction.

Beta-thalassemia minor causes microcytosis with, at most, mild anemia as a result of reduced HbA synthesis. Individuals with beta-thalassemia minor have one unaffected beta-globin gene, so they can still produce sufficient hemoglobin to supply the body's regular demand without causing significant erythroid hyperplasia.

Beta-thalassemia can also coexist with other hemoglobinopathies (hemoglobin S, C, and E, for example) and cause variably clinically significant anemias in the heterozygous beta-thalassemia carrier [6].

Table 2 : Common genotypes and basic classification of beta thalassemia

Genotypes	***Name***	***Phynotype***
β/β	Normal	None
β/β^0	Beta thalassemia trait	Thalassemia minor: asymptomatic, mild microcytic hypochromic anemia
β/β^+		
β^+/β^+ β^+/β^0	Beta thalassemia intermedia	Variable severity Mild to moderate anemia Possible extramedullary hematopoiesis Iron overload
β^0/β^0 β^+/β^+ β^+/β^0	Beta thalassemia major (Cooley's Anemia)	Severe anemia Transfusion dependence Extramedullary hematopoiesis Iron overload

1-3 : Clinical Features

The phenotypes of homozygous or genetic heterozygous compound beta-thalassemias include thalassemia major and thalassemia intermedia.

1-3-1: β-thalassemia minor : Carriers of thalassemia minor are usually clinically asymptomatic but sometimes have a mild anemi

1-3-2: β-thalassemia intermedia: Individuals with thalassemia intermedia present later than thalassemia major, have milder anemia and by definition do not require or only occasionally require transfusion. As a result of ineffective erythropoiesis and peripheral hemolysis, thalassemia intermedia patients may develop gallstones, which occur more commonly than in thalassemia major

1-3-3: β-thalassemia major: the clinical picture of thalassemia major is characterized by growth retardation, pallor, jaundice, poor musculature, genu valgum , hepatosplenomegaly, leg ulcers, development of masses from extramedullary hematopoiesis, and skeletal changes resulting from expansion of the bone marrow. Skeletal changes include deformities in the long bones of the legs and typical craniofacial changes [1] . Individuals with thalassemia major usually come to medical attention within the first two years of life and require regular RBC transfusions to survive [1]. However, patients who have undergone transfusions may develop complications related to iron overload, depending on their compliance with chelation therapy [5].

1-4 Diagnosis :

Several procedures have been proposed for beta-thalassemia carrier screening. The cheapest and simplest is based on MCV and MCH determination, followed by HbA2 quantitation for subjects showing microcytosis (low MCV) and reduced Hb content per red blood cell (low MCH) [2] .

1-4-1 Thalassemia minor: is characterized by reduced MCV and MCH, with increased Hb A2 level [1] .

1-4-2 Thalassemia intermedia : is characterized by Hb level between 7 and 10 g/dl, MCV between 50 and 80 fl and MCH between 16 and 24 pg [1] .

1-4-3 Thalassemia major : is characterized by reduced Hb level (<7 g/dl), mean corpuscolar volume (MCV) > 50 < 70 fl and mean corpuscolar Hb (MCH) > 12< 20 pg [1] .

TM is suspected in infants or children less than 2 years old with severe microcytic anemia, mild jaundice, and hepatosplenomegaly [5] .

we include MCV and MCH determination and Hb chromatography by HPLC, which can quantitate HbA2 and HbF and can detect the most common Hb variants (HbS, HbC, and HbE) that may result in a Hb disorder by interacting with beta-thalassemia [2] .

Table 3 : Hemoglobin electrophoresis of thalassemia types

Thalassemia type	**Hb A**	**Hb A2**	**Hb F**
Beta thalassemia minor	90%	3% to 10%	3% to 10%
Beta thalassemia intermediate	50% to 70%	3% to 8%	20% to 40%
Beta thalassemia major	0%	3% to 8%	>90%

[9]

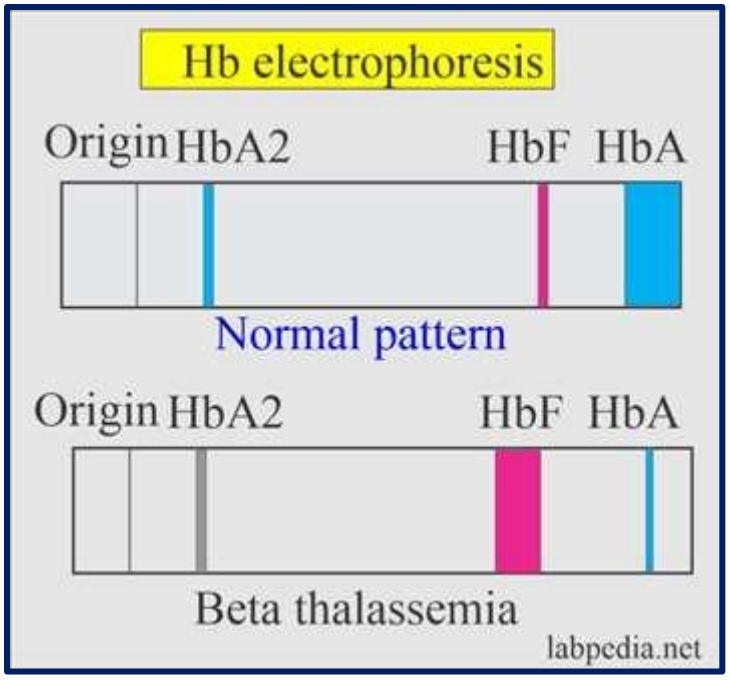


Figure 4 : Hemoglobin electrophoresis of thalassemia

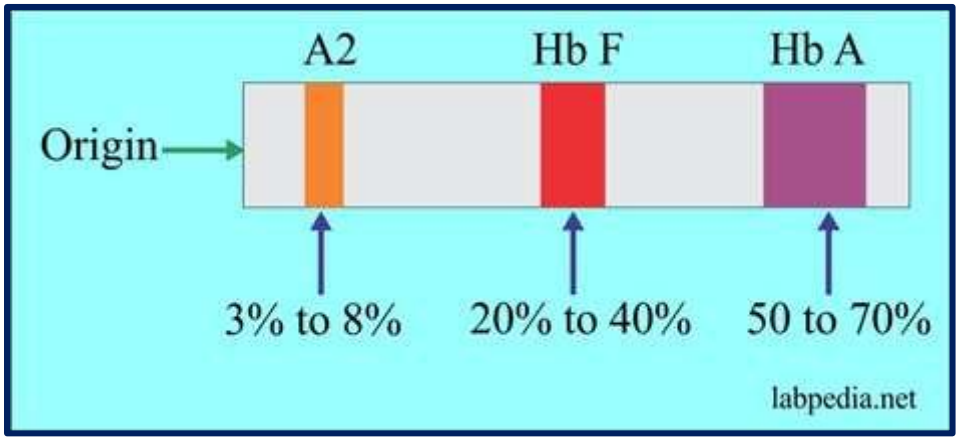


Figure 5 : Thalassemia intermedia on electrophoresis

1-4-4 : Prenatal diagnosis

Because of the high carrier rate for HBB mutations in certain populations and the availability of genetic counseling and prenatal diagnosis, Population screening associated with genetic counseling is extremely useful by allowing couples at risk to make informed decision on their reproductive choices.

When the hematological analysis indicates a beta-thalassemia carrier state, molecular genetic testing of HBB can be performed to identify a disease-causing mutation. If both partners of a couple have the HBB disease-causing mutation, each of their offspring has a ¼ risk of being affected. Through genetic counseling and the option of prenatal testing, such a couple can opt to bring to term only those pregnancies in which the fetus is unaffected [2] .

Newborn screening is an important way to identify thalassemia, especially in high-risk populations, before symptoms appear. Common methods of diagnosis in the newborn and later life are the Hb separation techniques, such as gel-based electrophoresis (especially isoelectric focusing), high-performance liquid chromatography, and capillary electrophoresis [15] .

1-4-5 : Expert systems for thalassemia disease diagnosis :

Diseases should be treated well and on time. If they are not treated on time, they can lead to many health problems and these problems may become the cause of death. These problems are becoming worse due to the scarcity of specialists, practitioners and health facilities. In an effort to address such problems, studies made attempts to design and develop expert systems which can provide advice for physicians and patients to facilitate the diagnosis and recommend treatment of patients [32] .The fuzzy control can be applied in cases where the control processes are too complex to analyze by conventional quantitative techniques or the available sources of information are interpreted qualitatively, inexactly, or uncertainly [31] . Having so many factors to detect Thalassemia makes doctor's work difficult. So, experts require an accurate tool that considering these risk factors and give some certain result in uncertain terms. Because of uncertainty involve in the diagnosis of Thalassemia disease a new method for Thalassemia disease diagnostic problem solving based on fuzzy inference system is constructed [27] .The constructed system is an efficient attempt to solve the Thalassemia disease problem. The proposed model detects Thalassemia on the basis of both Thalassemia Symptoms and CBC Test which are useful for

identifying carriers of the Thalassemia trait. The results of this work can facilitate laboratory work by reducing the time and cost [27].

Lotfi A. Zadeh in 1965 (Zadeh et al., 1965) of the University of California at Berkelely published fuzzy sets. Then Mamdani in 1975 applied the fuzzy logic in a practical application to control an automatic steam engine. In 1976 MYCIN, an early expert system, or artificial intelligence (AI) program, for treating blood infections was developed [28] . Fuzzy modeling owns some distinctive advantages, compared to traditional mathematical modeling, such as the mechanism of reasoning in human comprehensible conditions, the capability of captivating linguistic information from human experts and combining it with numerical data and the ability of approximating complex nonlinear functions with simple models.

Fuzzy expert systems are oriented toward numerical processing. It takes numbers as input, and then translates the input numbers into linguistic terms like Small, Medium and large(fuzzification) . Then the task of Rules is to map the input linguistic terms onto similar linguistic terms describing the output. Finally, the translation of output linguistic terms into an output number is done (Defuzzification) [32].

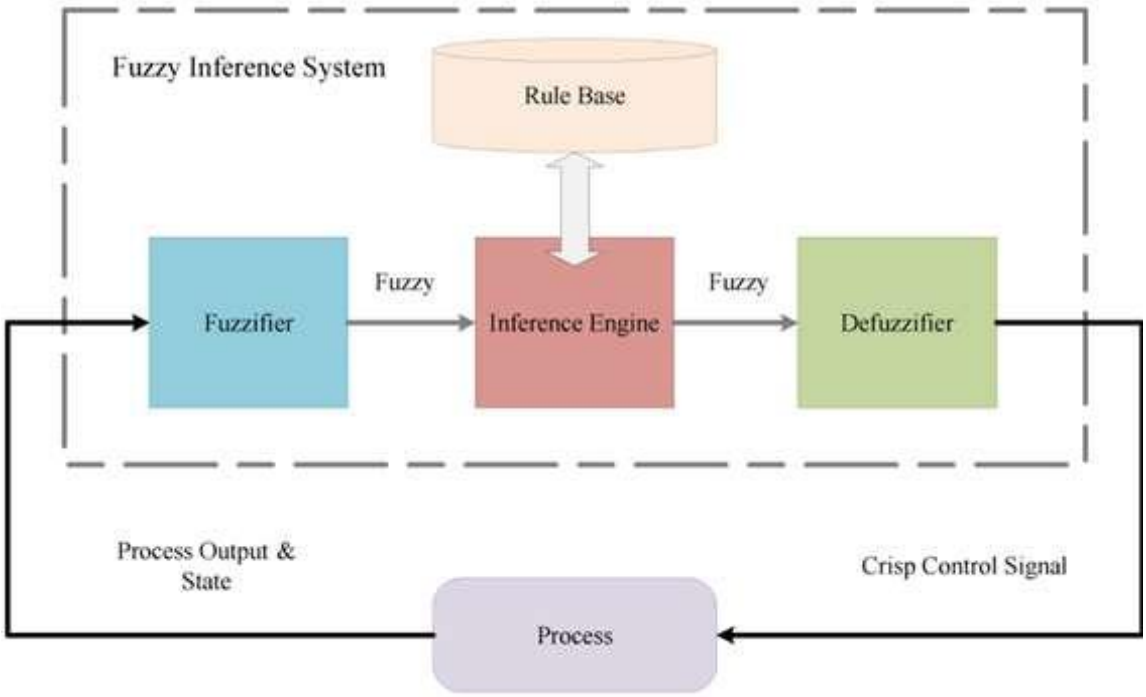


Figure 6 : Fuzzy logic controller

The basic framework used in this approach involves a presentation of the relationship being modeled by a collection of fuzzy IF–THEN rules [31]. These rules are based on the knowledge and experience of a human expert within that domain. Fuzzy rules are of the general form:

$$if \rightarrow antecedent(s)\ then\ consequent(s)$$

There are two major types of fuzzy rules, Mamdani fuzzy rules and Takagi-Sueno (TS) fuzzy rules , general Mamdani fuzzy rule can be expressed as:

$$IF\ v_1\ is\ S_1 AND\ .....AND\ v_m\ is\ S_m$$
$$THEN\ z_1\ is\ W_1,...,z_p\ is\ W_p$$

where $v_i, i = 1,..,M$, is an input variable and $z_j, j = 1,..,P$, is an output variable. S_i and W_j are input and output fuzzy sets respectively.

The features of the Fuzzy Mamdani method can be summarized as: It is intuitive, It has widespread acceptance and it is well suited to human input [32].

In previous studies , a Fuzzy Inference System was designed to diagnose the severity of the Thalassemia disease of a patient and Define linguistic Variables by using Fuzzy Logic, there was 3 input variables and 1 output variable [27], then the authors presented an improvement of the previous study (Thakur et al., 2016), such that the obtained results are more stable which included 26 rules instead of 15 rules [28]. In this study, we will design a Fuzzy Inference System of Dataset for various thalassemia patients in Iraq in 2022 from Hematology Center (Thalassemia) in Ibn Al-Baladi Hospital to diagnose the severity of the Thalassemia disease .

In classification, the confusion matrix can be used to evaluate performance of the method. Confusion matrix show the number of samples that were correctly and incorrectly diagnosed from classification model compared to the actual results in the data [42].

Table 4 : The confusion matrix

Class		Predicted	
		Positive	Negative
Expected	Positive	TP	FN
	Negative	FP	TN

Based on Table 4, there are four conditions for measuring performance. True Positive (TP), it means the number of samples having thalassemia disease which are correctly diagnosed. False Negative (FN), it means the number of samples having thalassemia disease which are incorrectly diagnosed. False Positive (FP), it means the number of non-thalassemia samples which are incorrectly diagnosed. True Negative (TN), it means the number of non-thalassemia samples which are correctly diagnosed [42] . According to the value in confusion matrix table, we can calculate the value of accuracy, precision, and recall . The sum of the values of the cells across provides the number of actual cases in the training dataset while the sum of the columns provide the number of predicted cases in the training dataset. The cells located on the diagonal are the correct classifications (true positives/negatives) while other cells are the misclassifications/incorrect classifications (false positives/negatives) [43] .

1-5 : Etiology

The globin gene clusters show variability in their base composition and are organized into alpha globin gene cluster and beta globin cluster [16] .

The α-globin gene cluster is located on the short arm of Chromosome 16 (16p13.3). The cluster includes three protein coding functional genes (α1, α2, and ζ2) and spans about 30 kb, and the genes are arranged in order as per their expression during developmental stages in human beings [16].

The beta globin (HBB) gene maps in the short arm of chromosome 11 specifically on the short arm of the chromosome at position 15.5 [11] .

NCBI Reference Sequence: NC_000011.10 for HBB gene
https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10?from=5225464&to=5227071&report=fasta&strand=true [1]

[1] This link was validated on 4/2023

Figure 7 : NCBI Reference Sequence for HBB gene

In a region also containing the delta globin gene, the embryonic epsilon gene, the fetal A-gamma and G-gamma genes, and a pseudogene (ψB1). The five functional globin genes are arranged in the order of their developmental expression [2] .The HBB gene, which spans 1.6 Kb, contains three exons and both 5’ and 3’ untranslated regions (UTRs).The HBB is regulated by an adjacent 5’ promoter in which a TATA, CAAT, and duplicated CACCC boxes are located. A major regulatory region, containing also a strong enhancer, maps 50 Kb from the beta globin gene [2] .

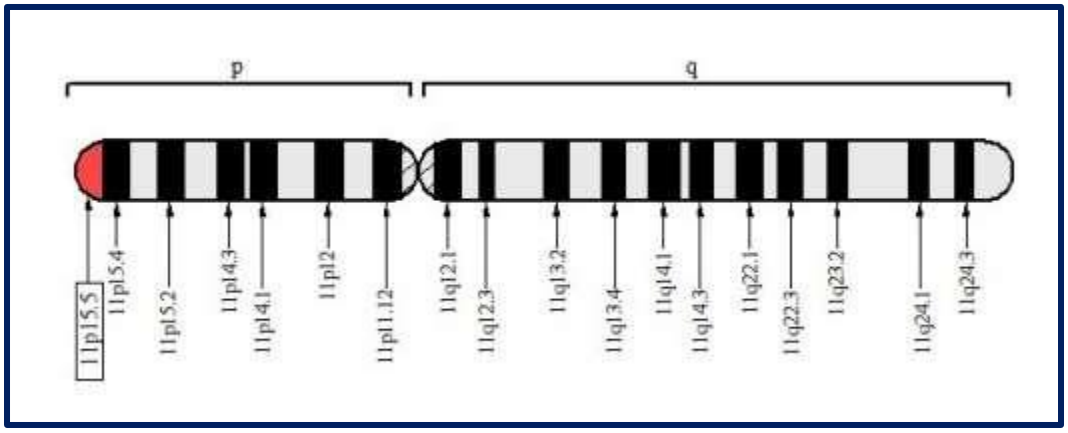


Figure 8 : Location of HBB gene on the chromosome 11

This region, dubbed locus control region (LCR), contains four (HS-1 to HS-4) erythroid specific DNAse hypersensitive sites (HSs), which are a hallmark of DNA-protein interaction [2]. The LCR appears to interact with a combination of transcription factors at the onset of erythroid maturation in such a way as to enhance access of the transcriptional machinery and other transcriptional factors to the promoters, enhancers, and silencers within the gene complex. LCR function is absolutely required for expression of globin genes at the extraordinary high levels needed for normal hemoglobin synthesis [16].

Beta^{0}-thalassemias, characterized by the complete absence of beta chain production result from deletion, initiation codon, nonsense, frameshift, and splicing mutations, especially at the splice-site junction. On the other hand, beta^{+}-thalassemias, characterized by reduced production of the beta chains, are produced by mutations in the promoter area (either the CACCC or TATA box), the polyadenylation signal, and the 5’ or 3’ UTR or by splicing abnormalities. According to the extent of the reduction of the beta chain output, the beta+-thalassemia mutations may be divided into severe, mild, and silent [1] .

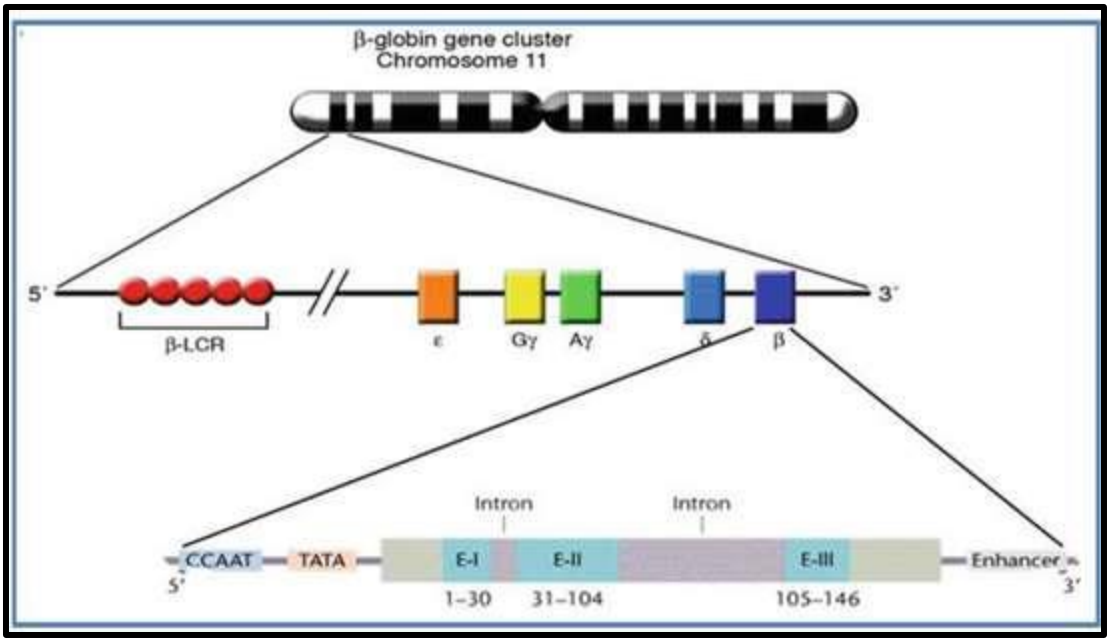


Figure 9 : Chromosome localization and structure beta globin gene

1-5-1 : Single nucleotide polymorphisms (SNPs) In causing β-thalassemia

During the past few years there has been a rapid increase of knowledge in the field of genetic control of hemoglobin synthesis in health and disease, which led to reviving the interest in thalassemia and associated disorders of hemoglobin (Hb) production [17]. More than 200 mutations have been so far reported; the large majority are point mutations in functionally important regions of the beta globin gene . Deletions of the beta globin gene are uncommon. The beta globin gene mutations cause a reduced or absent production of beta globin chains [1] . Point mutations affecting the beta globin expression belong to three different categories: mutations leading to defective beta-gene transcription (promoter and 5' UTR mutations); mutations affecting messenger RNA (mRNA) processing (splice-junction and consensus sequence mutations, polyadenylation, and other 3' UTR mutations); and mutations resulting in abnormal mRNA translation (nonsense, frameshift, and initiation codon mutations) [2]. One important reason for the variability in the expression pattern of the genes is due to

change in the protein sequence caused by a type of mutations known as SNPs [11] .

Diagnostic methods based on single nucleotide polymorphism (SNP) biomarkers are essential for the real adoption of personalized medicine. Single nucleotide polymorphisms (SNPs) have been proposed as the next generation of markers to identify loci associated with complex diseases and their therapeutic treatment [1]. The main initiative behind SNP-related work is that genetic differences between people to be used to predict phenotypes and phylogeny [22]. Single Nucleotide Polymorphism (SNP) can be defined as a nucleotide variation in which a single base change occurs that may or may not lead to a phenotypic change. It is a variation in the DNA sequence in which a single nucleotide (A, T, G or C) differs between members of a biological species or paired chromosomes in an individual [11]. There are two types of nucleotide base substitutions resulting in SNPs: A transition substitution occurs between purines (A, G) or between pyrimidines (C, T). This type of substitution constitutes two thirds of all SNPs. A transversion substitution occurs between a purine and a pyrimidine [22] . In humans these variations occur at a frequency of more than 1% i.e. occurs in about every 3000 base pairs of the genome. SNPs occur both in the coding as well as the non-coding region of the DNA [11] . Understanding of SNPs can help in identifying the cause of β-thalassemia AND analysis of various SNPs involved in the gene sequence variations helps in discovering newer engineered methods which can be successfully used to reduce the severity of hemoglobin opathies includes β-thalassemia and sickle cell anemia [11].

SNPs in the non-coding region do not hinder the translation of standardproteinshence report normal functional protein but SNPs in the coding region may alter the functionalities of the translated proteins which may lead to a noticeable phenotypic change. **Synonymous SNPs** are the third base pair changes in the codon which does not lead to any change in the amino acid; asper the Wobble hypothesis, a single amino acid can be coded by multiple codons, any change in the third base pair does not lead to any change in the protein structure. **Nonsynonymous SNPs** are the mutations

leads to the formation of abnormal protein sequences occurs due to either missense mutation or nonsense mutation. In case of **missense type**, the change in a base pair of the amino acid leads to the change in the codon, forms different amino acids which inturn change the entire protein sequence and the functionality of the protein may be altered . In case of **nonsense type** of mutation, the change in the single base pair leads to the formation of a stop codon yields an incomplete non functional protein upon translation [11]. Mutations that create a premature translation termination codon (nonsense codon) account for the most common forms of thalassemia, in terms of numbers of patients affected. These mutations create translation stop signals prematurely, so that the complete beta globin polypeptide is never made. In the most common type of thalassemia, globin fragments are highly unstable, resulting in the accumulation of protein synthesized from the mutated gene (i.e., beta (0) thalassemia) . As a result, patients homozygous for this defect cannot make any beta chains and suffer from a severe form of beta thalassemia [13]. **Latent SNPs** are the variations which occur in the coding and regulatory region, do not cause any harm at the present state but may transform from harmless to harmful under certain stress conditions [11].

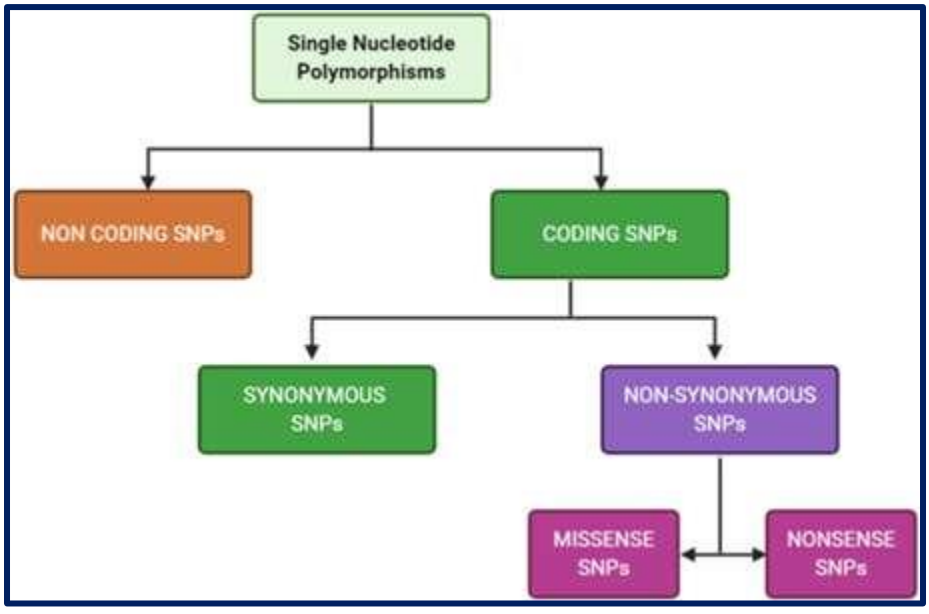


Figure 10 : Types of SNPs

Change in charge properties of amino acids due to mutation results in the change of overall integrity and 3 dimensional (3D) confirmations of the proteins and hence it altered functionality. The SNPs causing change in charge properties of amino acids results in β-thalassemia was discussed in the Figure 11 [11] .

Accession number	Position	Amino acid		Property	
		Normal sequence	Mutated sequence	Normal sequence	Mutated sequence
rs28933077	475	Cysteine	Glycine	Polar uncharged	Non polar
rs33986703	102	Lysine	Glutamine	Positive	Polar uncharged
rs33986703	102	Lysine	Glutamate	Positive	Negative
rs33959855	117	Glutamate	Lysine	Negative	Positive
rs33959855	117	Glutamate	Glutamine	Negative	Polar uncharged
rs33941844	370	Leucine	Glutamine	Non polar	Polar uncharged
rs33941844	370	Leucine	Arginine	Non polar	Positive
rs34579351	334	Aspartate	Glycine	Negative	Non polar

Figure 11 : Change in charge properties of amino acid due to mutation

1-5-2 : Molecular diagnosis of modifying genes

Distinction between TM and TI is currently based on clinical criteria and it often, therefore, takes at least four years of follow up before classification can be confirmed. Variant genotyping of genetic modifiers may possibly help in the early prediction of the type of thalassemia the patient will develop later. If further validated, this prediction tool of severity may have implications not only for genetic counseling but also for therapeutic decision making concerning [7] .

Polymerase-chain-reaction (PCR) technology has been used for more than a decade to detect point mutations or deletions in chorionic-villus samples,

enabling first-trimester, DNA-based testing for thalassemia . Commonly occurring mutations of the HBB gene are detected by a number of polymerase chain reaction (PCR)-based procedures [3] .

There are many computational tools available to assist with critical bioinformatics issues related to primer design. These resources allow the user to define parameters and criteria that need to be taken into account when designing primers. Following the initial in silico selection, a primer pair should be further tested in vivo for their amplification efficiency and robustness [21]. One remarkable program is **BatchPrimer3**, https://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi [1] based on the Primer3 algorithm, which incorporates a specific module to choose the best primer pairs for AS-amplification [20] . BatchPrimer3 was designed as a web application consisting of a set of CGI programs written in Perl, which can run on different operating systems, such as Solaris, Linux, Mac OS or Windows with an Apache HTTP server and Perl interpreter program. BatchPrimer3 adopted the Primer3 core program as a major primer design engine to choose the best primer pairs. It is a high throughput web application for PCR and sequencing primer design [30] . BatchPrimer3 v1.0 implements several types of primer designs including generic primers, SSR primers together with SSR detection, and SNP genotyping primers (including single-base extension primers, allele-specific primers, and tetra-primers for tetra-primer ARMS PCR), as well as DNA sequencing primers [30] .

In most SNP detection platforms, SNP detection requires previous PCR amplification of the genomic region that flanks the SNP site [30] .Deletion alpha^{0}- or alpha^{+}-thalassemias are detected by PCR using **two primers flanking** the deletion breakpoint, which amplify a DNA segment only in presence of specific deletions [2]. Primer extension is the most commonly used approach to SNP genotyping because it can be used in a wide variety of high-throughput detection platforms, i.e., electrophoresis [30] .

[1] This link was validated on 4/2023

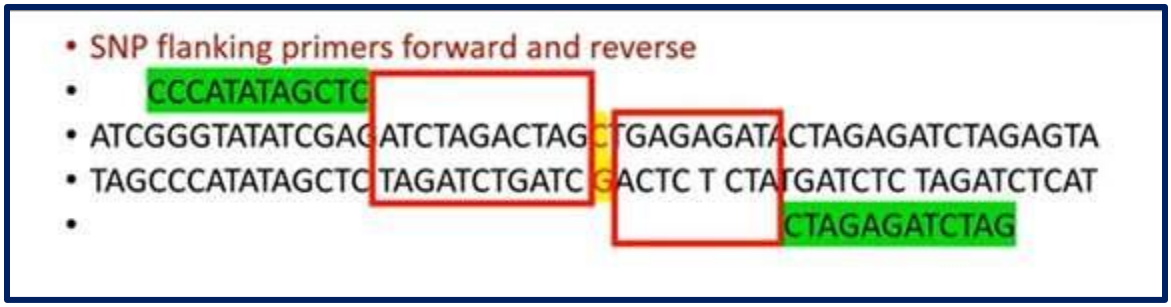


Figure 12 : SNP flanking primers

For SNP flanking primer or SNP genotyping primer design, the SNPs or alleles in sequences need to be converted to IUB/IUPAC codes ,and the sequence file follows the NCBI dbSNP FASTA format [30] .

Table 5 : IUB/IUPAC Code of a SNP

IUB/IUPAC Code of a SNP	*R*	*Y*	*S*	*W*	*K*	*M*	*B*	*D*	*H*	*V*	*N*
Alleles of a SNP	G/A	T/C	G/C	A/T	G/T	A/C	C /G / T	A /G / T	A/ C / T	A / C / G	A / C / G /T

The genotyping principle is based on an effective primer extension by the polymerase when the 3 terminal base of the primer matches its target, whereas extension is inefficient or nonexistent when the terminal base is mismatched [20] . AS-PCR approach is to use primer mixtures combined to genotyping determination based on product size or amplification kinetics. SNPs can be genotyped using AS primers with the last nucleotide at the 3' end of a primer corresponding to the site of the SNP . Thus, an AS primer is specific to one of two alleles of a SNP at the 3' end of primers and specifically amplifies one of the two alleles. If a common reverse primer is used in the reaction, the reaction is called Allele-Specific PCR (AS-PCR) [30] . AS-PCR is also known as amplification refractory mutation system (ARMS) . This technique is a quick and dependable genotyping protocol that requires minimal instruments found in most laboratories .

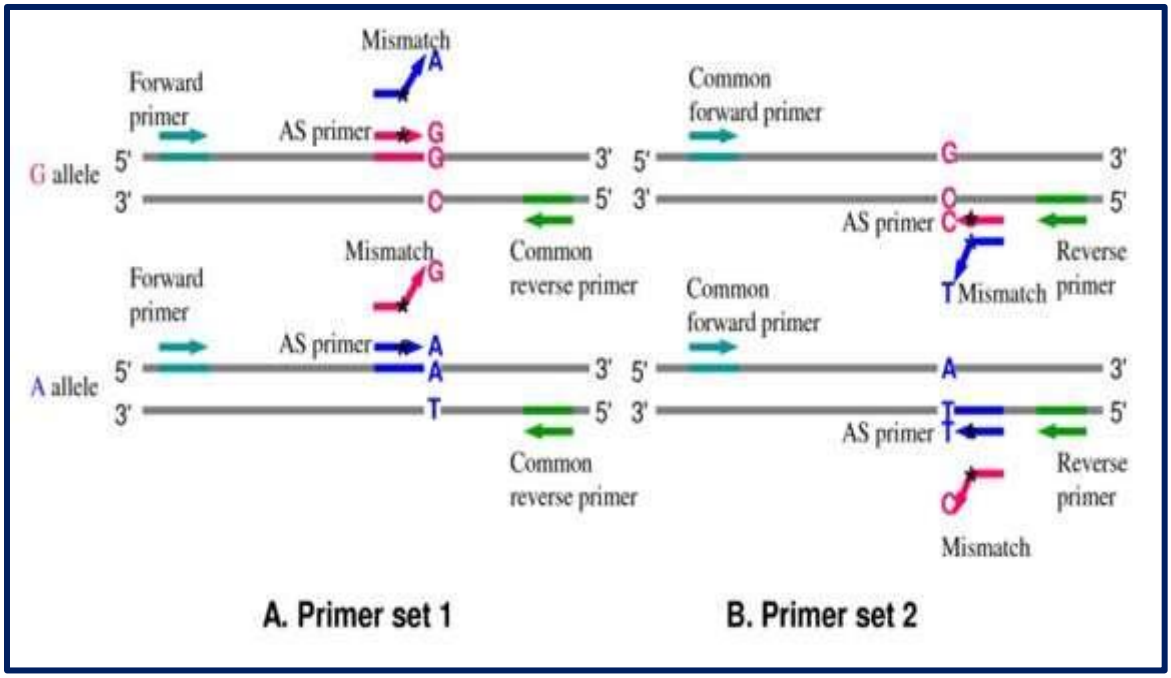


Figure 13 : Primer design of allele-specific (AS) primers

Once resulting AS primer set is obtained, these primers can be verified for their global uniqueness by running through an in silico PCR . the genotyping results can be observed by simply comparing the length of PCR products [33]. **SnapGene (**https://www.snapgene.com/free-trial [1]) is the most popular cloning tool for a reason. It's fast, smart and extremely user-friendly and is the primary tool utilized to facilitate students' understanding of gene cloning by allowing them to design primers, generate the results of Gibson assembly or ligation cloning, and simulate polymerase chain reaction (PCR), restriction digestion, and agarose gel electrophoresis . Intuitive technology identifies design flaws in cloning procedures so they can be corrected and simulate standard PCR using your own primers, or allow SnapGene to design them automatically Specialised cloning tools ensure fast accurate construct design for all major molecular cloning techniques and Clear visual schematics let you see exactly how your construct will be put together and Visualise exactly

[1] This link was validated on 5/2023

what you will see in the lab with SnapGene's empirically based gel simulation algorithm and Flexible configuration of all gel elements, including number of lanes, % agarose, running time and a full set of MW markers, recording and identifing your band of interest with detailed fragment information for each lane [37] .

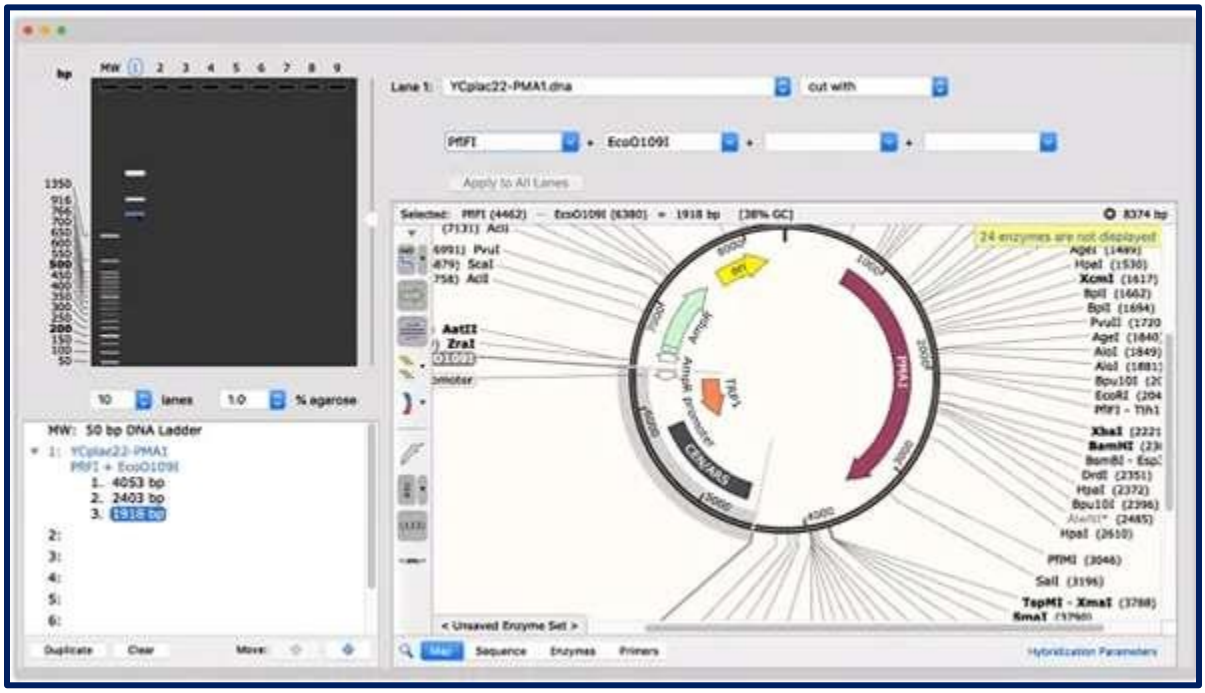


Figure 14 : Web interface of SnapGene

In previous studies , 2071 primer pairs were designed based on SNPs from Brassica oleracea , rapeseed (Brassica napus L.) and sesame (Sesamum indicum). High polymorphism percent (75%) of the designed primers indicated it is a general method and can be applied in other species [19] and the variations in the HBB gene in the 1,000 Genomes database were analyzed to describe the mutation frequencies in the different population groups, and to investigate the pattern of pathogenicity , Twenty different mutations were found in 209 healthy individuals [29] . In addition ,Thai β0-thalassemia/HbE disease genome-wild association studies (GWAS) data of 487 patients were analyzed by SNP interaction prioritization algorithm to find predictive SNPs for disease severity [46].

These studies need to reference which include SNP primers database for all Mutations that have been discovered ,so in this study we will provide a guide to choose the most efficient way to design a new specific-primer by applying BatchPrimer3 on SNPs from the HbVar database that cause (β^0 or β^+) genotyping in the exon regions of HBB gene , so the application of a combined molecular approach with clinical data and efficient bioinformatics tools will enable understanding the relationship between phenotype and genotype in the clinical setting and investigating the effects of SNP mutations in the HBB exons and give a guideline for functional studies and prenatal diagnosis to be developed as basis for future studies.

1-5-3 : Secondary structure prediction of protein

Of all the molecules found in living organisms, proteins are the most important as they are the biological workhorses that carry out vital functions in every cell. With the advent of various sequencing techniques, amino acid sequences for a number of proteins have been determined. However, three-dimensional structural information obtained through X-ray crystallography, nuclear magnetic resonance, and other experimental methods are available only for around 10% of these protein sequences. Hence, computational prediction of protein structures has become important with the rapid growth of database of protein sequences. Such an attempt for the use of computational methods for protein structure prediction based on only primary structure information started over 40 years ago. The prediction of protein secondary structure is an important step in modeling the tertiary structure of a protein which indeed is essential for the functional annotation of the protein [38]. The secondary structure arises from the hydrogen bonds formed between atoms of the polypeptide backbone. The hydrogen bonds form between the partially negative oxygen atom and the partially positive nitrogen atom. Most proteins have segments of their polypeptide chains that are either coiled or folded in patterns that contribute to the protein's shape. Many of these coils and folds repeat so often that they have been

given names. Two folds that are extremely common in biochemistry are the alpha-helix and the beta-pleated sheet [41] . Some regions of the protein chain do not form regular secondary structure and are not characterized by any regular hydrogen bonding pattern. These regions are known as random coils and are found in two locations in proteins: Terminal arms (both at the N-terminus and the C-terminus of the protein); Loops which are unstructured regions found between regular secondary structure elements.

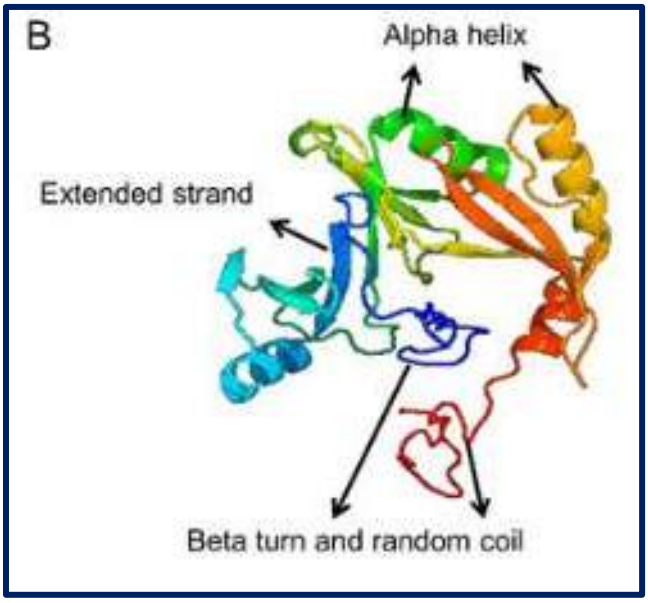


Figure 15 : Secondary structure of protein

Recently, we have described a new method called **SOPMA** (self-optimized prediction method with alignment) https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html [1] to predict the secondary structure of a given protein. Briefly, this method: (i) builds a limited database of protein sequences with their known secondary structures; (ii) predicts the secondary structure of all the proteins of the

[1] This link was validated on 5/2023

database using a similarity algorithm; (iii) determines the prediction parameters that maximize the accuracy of the prediction; (iv) applies the prediction parameters to the given protein [36] . According to this method, short homologous sequence of amino acids will tend to form similar secondary structure. So it has a whole database consist of 126 chains of non-homologous proteins. If the user enters an unknown protein, it will search against a collection of proteins in the database that have some similar properties and evolutionary history [38] .

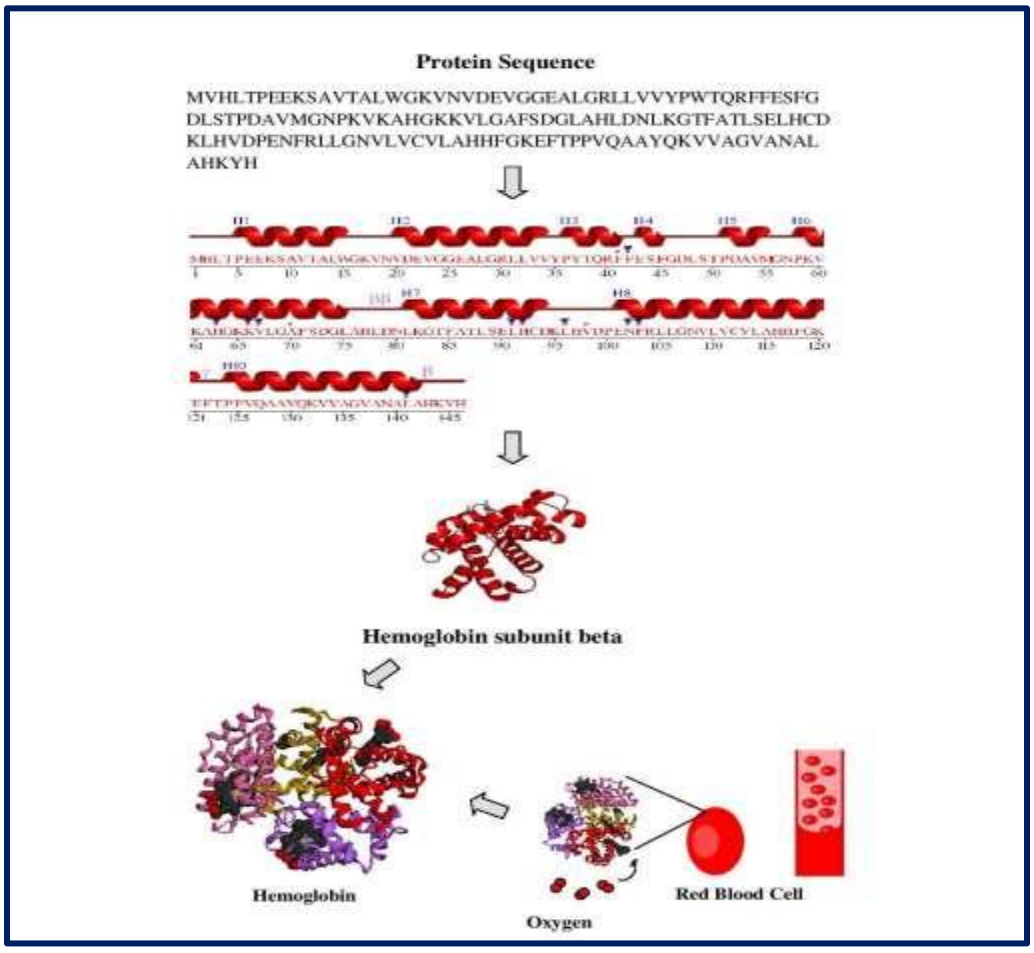


Figure 16 : Protein sequence of HBB

In previous studies, the molecular epidemiology characteristics of Hb variants, along with associated structural and functional predictions in the Yunnan province population of Southwestern China were investigated. Based on bioinformatics and structural analysis, as well as protein modeling, the pathogenesis and type of Hb genetic mutations were characterized [47]. Also, Secondary structure prediction was achieved for two common mutations with online tools. The mutations were also common among the countries neighboring Iran, which are responsible for 71% to 98% of mutations. In particular, This is valuable for the analysis of panel-based studies in the region that may have no access to advanced genetic technologies [44]. In addition ,the secondary and tertiary structures of common Hb Q variants using bioinformatics tool were investigated . Differences in secondary and tertiary structure of various Hb Q variants have been observed in the present study [45] .

In this study , we will predict secondary structure of HBB protein by SOPMA tool . The study will provide valuable data for better understanding of these uncommon hemoglobinopathies . And the recent advances in molecular biology will allow us to simulate mutations on the basis of their known sequences. Also , a greater understanding of this will help to shed light on the pathogenesis of these disorders.

1-6: Management of thalassemia and treatment

A comprehensive review of the management of thalassemia major and thalassemia intermedia has been published by Thalassemia International Federation and is available at the Thalassemia International Federation Web site [2] .

In the most severe forms of beta thalassaemia the anemia is so severe that unless it is corrected regularly by blood transfusion the patient will die early in life (mostly in infancy). The condition is then known as transfusion dependent thalassaemia or TDT. Other cases may be able to survive with

occasional or no blood transfusions, known as non-transfusion thalassaemia or NTDT [12] .

1-6-1 Management of thalassemia minor

Patients with beta thalassemia minor are usually asymptomatic and are often monitored without treatment [14] .

1-6-2 Management of thalassemia intermedia:

Patients with beta thalassemia intermedia require no transfusions or may require episodic blood transfusions during certain circumstances (infection, pregnancy, surgery) [14] .

Treatment of individuals with thalassemia intermedia is symptomatic . As hypersplenism may cause worsening anemia, retarded growth and mechanical disturbance from the large spleen, splenectomy is a relevant aspect of the management of thalassemia intermedia. Risks associated with splenectomy include an increased susceptibility to infections mainly from encapsulated bacteria (Streptococcus Pneumoniae, Haemophilus Influenzae and Neisseria Meningitidis) and an increase in thromboembolic events. Prevention of post-splenectomy sepsis includes immunization against the above mentioned bacteria and antibiotic prophylaxis as well as early antibiotic treatment for fever and malaise [1]. Recently promising results have been obtained with platelet derived growth factor. Since patients with thalassemia intermedia have a high risk of thrombosis, exacerbated by splenectomy. Supplementary folic acid can be prescribed to patients with thalassemia intermedia to prevent deficiency from hyperactive bone marrow [1].

1-6-3 : Management of thalassemia major :

Thalassemia major is treated with red blood cell transfusion. The aim of transfusion is mainly to suppress erythroid expansion. It also serves to mitigate symptoms of anemia and to inhibit gastrointestinal iron absorption [6].

The most widely used drug is Hydroxyurea. Hydroxyurea could be a cell-cycle specific agent which blocks DNA synthesis by inhibition of the ribonuclease reductase, the enzyme which converts ribonucleotides to deoxyribonucleotides. It's been seen that chronic daily low dose administration of hydroxyurea will enhance gamma globin synthesis, increase red cell production, and partially or substantially correct the anemia in patients with homozygous beta-thalassemia. Hydroxyurea is also one of the most costeffective options available in the market [23] , but there are worries about negative side effects, such as long-term carcinogenesis . Hence, it is necessary to develop good inhibitor for beta thalassemia from natural inducer to treat beta thalassemia a without any side effects [26] .

The most common secondary complications are those related to transfusional iron overload, which can be prevented by adequate iron chelation [2] ,Later iron overload-related complications include involvement of the heart (dilated myocardiopathy or rarely arrythmias), liver (fibrosis and cirrhosis), and endocrine glands (diabetes mellitus, hypogonadism and insufficiency of the parathyroid, thyroid, pituitary, and, less commonly, adrenal glands) Other complications are hypersplenism, chronic hepatitis (resulting from infection with viruses that cause hepatitis B and/or C), HIV infection, venous thrombosis, and osteoporosis [1], As the body has no effective means for removing iron, the only way to remove excess iron is to use iron binders (chelators), which allow iron excretion through the urine and/or stool [1] .

Deferoxamine continues to be the most common iron-chelating agent in use, but it has several limitations: the need for parenteral administration (which is painful and reduces compliance), side effects, and cost (which is prohibitive in underdeveloped countries) [3] . Bone marrow transplantation (BMT) remains the only definitive cure currently available for patients with thalassemia. If BMT is successful, iron overload may be reduced by repeated phlebotomy, thus eliminating the need for iron chelation [1] . Therapies under investigation are the induction of fetal hemoglobin with pharmacologic compounds and stem cell gene therapy [2] .

1-6-4 Gene therapy :

Thalassemias typically affect only the mRNAs for production of the beta chains (hence the name). Since the mutation may be a change in only a single base (single-nucleotide polymorphism), on-going efforts seek gene therapies to make that single correction [6]. The possibility of correction of the molecular defect in hematopoietic stem cells by transfer of a normal gene via a suitable vector or by homologous recombination is being actively investigated . The most promising results in the mouse model have been obtained with lentiviral vectors [1] . Small interfering RNA is the basis for a new strategy to augment transduced β-globin expression. Small interfering RNA corresponding to transcripts of BP1 (a protein that negatively regulates β-globin expression by binding to its upstream region) enhanced β-globin promoter activity in erythroid cells [3] .

1-6-5: Molecular docking

After 2000, all of these developments led to a significant trend in decreasing cardiac mortality, which was previously reported to cause 71% of deaths for individuals with TM. Recent studies have shown that despite geographic differences, most individuals with transfusion-dependent thalassemia have normal cardiac iron; however, a significant proportion have simultaneous liver iron overload and the number of patients who die from liver disorders now exceeds that of individuals who die from cardiac diseases in some European countries. In particular, the risk for hepatocellular carcinoma has progressively increased secondary to liver viral infection, iron overload, and longer survival [5].

Clearly, these types of therapies remain highly experimental; accordingly, their clinical potential remains uncertain. Nevertheless, these methods may be useful molecular approaches for the development of new therapies for thalassemia [3]. Linkage analysis and genome-wide association studies have greatly contributed to results, and next generation sequencing might further improve prediction ability and eventually guide the development of new therapies [4] . In novel therapeutics molecules, in silico drug discovery is an

emerging and effective alternative for identification . Various bioactive compounds have been discovered to be effective for a wide range of therapeutic applications [26] .

Programs based on different algorithms were developed to perform molecular docking studies, which have made docking an increasingly important tool in pharmaceutical research. The docking process involves two basic steps: prediction of the ligand conformation as well as its position and orientation within these sites (usually referred to as pose) and assessment of the binding affinity [24] .Molecular docking programs like AUTODOCK 4 perform a search algorithm in which the conformation of the ligand is evaluated recursively until the convergence to the minimum energy is reached [25] . Docking studies revealed that the best hit molecule for for beta thalassemia was the drug Indicaxanthin and the bioactive Berberine from CoptidisRhizome .The Coptidis Rhizome is commonly used in the Traditional Chinese Medicine (TCM) for treating various diseases and contain bioactive compounds such as alkaloids, phenylpropanoids, flavonoids and other compounds, All these bioactive compounds tend to have pharmacological activities, especially berberine is known to have anti-pathogenic and antibacterial effects [23] .

The majority of patients are thought to be receiving plant-based medication treatment in order to maintain a healthy lifestyle with fewer side effects. Many studies have extracted bioactive chemicals from natural sources, but more research is needed to discover their molecular interactions, such as computational analyses. As a result, in the current investigation, we used molecular docking to explore the efficacy of several bioactive compounds [26] .

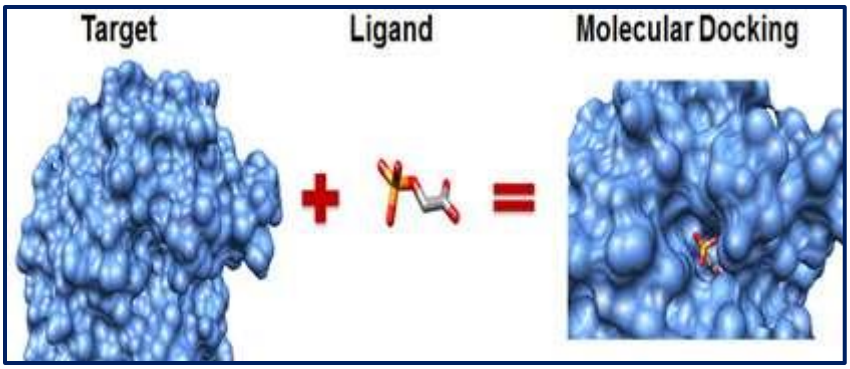


Figure 17 : Molecular docking

A large interest for docking web servers has emerged recently, such as **SwissDock (**http://www.swissdock.ch/[1]). With the SwissDock web site, we aim at extending the use of protein-small molecule docking software far beyond experts in the field by providing convenient answers to many of the difficulties . First, manually curated protein structures can be downloaded from the web site, and original PDB files can be prepared through ad hoc scripts . Second, the docking software is easily accessible through either a web browser or a programmatic interface. Third, predicted binding modes (BMs) can be viewed online with a simple embedded applet or analyzed in more details thanks to a seamless integration with the UCSF Chimera molecular viewer , with the help of the online documentation and the user community. In brief, several docking parameters are adjusted in order to reach the desired docking time and exhaustiveness of the search: the number of sampled BMs, the number of minimization steps that are performed to relax the ligand and the number of predicted BMs . This docking result web page features a Jmol applet for the visualization of the predicted BMs within the web browser [40].

SwissDock is based on the docking software EADock DSS . Its algorithm consists of the following steps. First, a large number of BMs (typically from 5000 to 15 000) are generated, either in a user-defined box (local docking) or in the vicinity of the target cavities of the entire protein surface (blind

[1] This link was validated on 6/2023

docking). Simultaneously, their CHARMM energies are estimated on a grid. Then, BMs with the most favorable energies are ranked . This unique combination of features allows accurate docking assays to be carried out within minutes [40] .

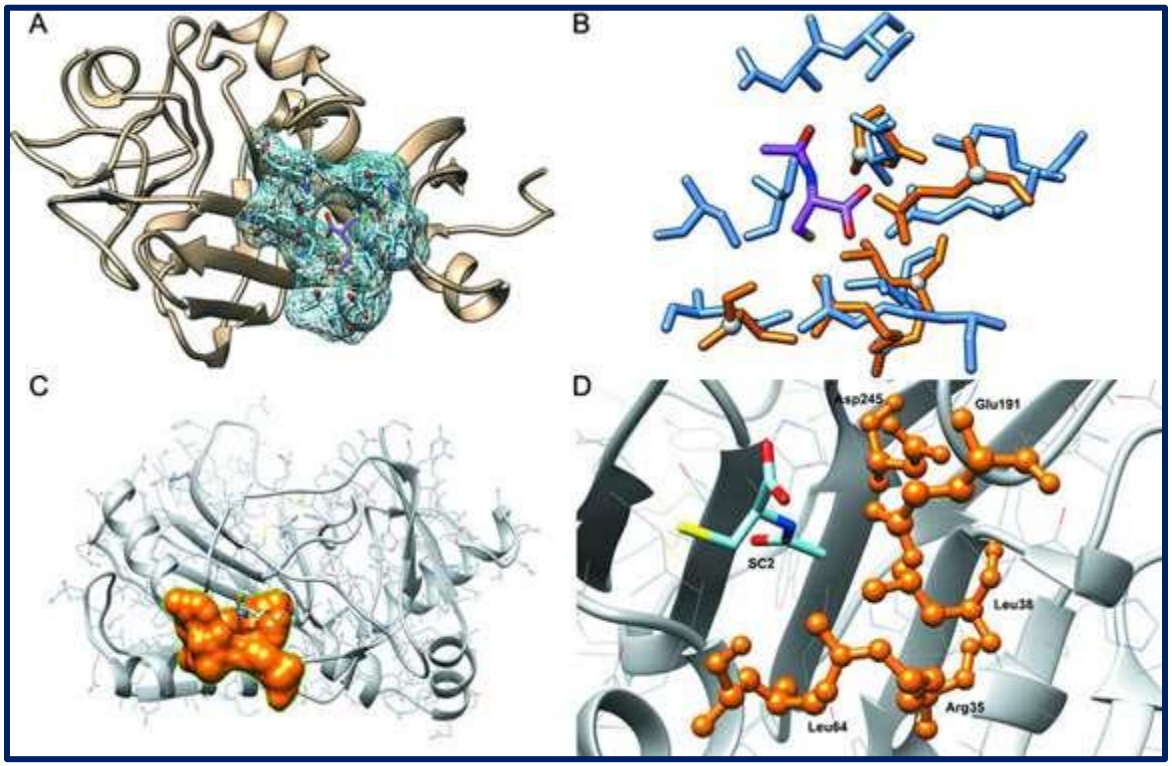


Figure 18 : Molecular docking prediction performed by SwissDock.

In this study , we will perform molecular docking between HBB protein and four different ligands via SwissDock .

2- dataset

To perform this study, data from the HbVar database https://globin.bx.psu.edu/cgi-bin/hbvar/query_vars3 were used. HbVar is the oldest and most appreciated database of hemoglobin variants and thalassemia mutations established in 2001 . It is a locus specific database,

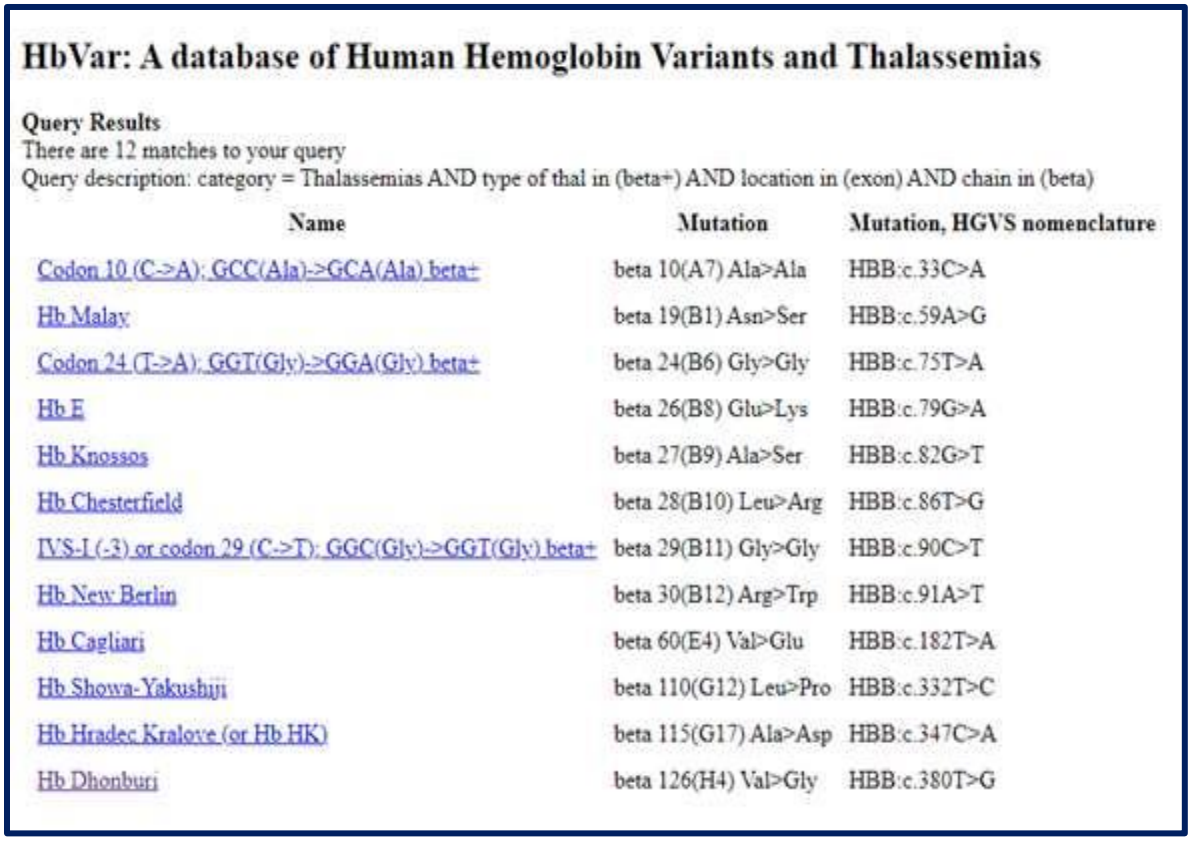
which was developed as a combined academic effort to keep a record of hemoglobin variants, new data entries, updates, and corrections. It provides high quality and up-to-date information on the genomic variations, associated phenotypic and hematological effects, pathology, frequency of different mutations, ethnic prevalence, and references [34] .

Using this database, we identified Mutations of the thalassemia that cause (β^0 or β^+) genotyping in the exon regions of HBB gene and picked up (only single nucleotide substitutions) for in silico investigation .

Query Results
There are 126 matches to your query
Query description: category = Thalassemias AND type of thal in (beta0) AND location in (exon) AND chain in (beta)

Name	Mutation	Mutation, HGVS nomenclature
Initiation codon ATG->GTG beta0	beta Initiation codon Met>Val	HBB:c.1A>G
Initiation codon ATG->AGG beta0	beta Initiation codon Met>Arg	HBB:c.2T>G
Initiation codon ATG->ACG beta0	beta Initiation codon Met>Thr	HBB:c.2T>C
Initiation codon ATG->ATA beta0	beta Initiation codon Met>Ile	HBB:c.3G>A
Initiation codon ATG->ATC beta0	beta Initiation codon Met>Ile	HBB:c.3G>C
Initiation codon ATG->ATT beta0	beta Initiation codon Met>Ile	HBB:c.3G>T
Codon 15 (G->A); TGG(Trp)->TAG(stop codon) beta0	beta 15(A12) Trp>Stop	HBB:c.47G>A
Codon 15 (G->A); TGG(Trp)->TGA(stop codon) beta0	beta 15(A12) Trp>Stop	HBB:c.48G>A
Codon 17 (A->T); AAG(Lys)->TAG(stop codon) beta0	beta 17(A14) Lys>Stop	HBB:c.52A>T
Codon 22 (G->T); GAA(Glu)->TAA(stop codon) beta0	beta 22(B4) Glu>Stop	HBB:c.67G>T
Codon 26 (G->T); GAG(Glu)->TAG(stop codon) beta0	beta 26(B8) Glu>Stop	HBB:c.79G>T
IVS-I (-2) or codon 30 (A->G); AG'GTTGGT->GG'GTTGGT Probably beta0	beta 30(B12) Arg>Gly	HBB:c.91A>G
Hb Monroe	beta 30(B12) Arg>Thr	HBB:c.92G>C
IVS-I (-1) or codon 30 (G->A) AG'GTTGGT->AA'GTTGGT beta0	beta 30(B12) Arg>Lys	HBB:c.92G>A
Hb Medicine Lake	beta 98(FG5) Val>Met AND beta 32(B14) Leu>Gln	HBB:c.[295G>A;98T>A]
Codon 35 (C->A); TAC->TAA (Tyr->Term codon) beta0	beta 35(C1) Tyr>Stop	HBB:c.108C>A
C837 (TGG>TAG)	beta 37(C3) Trp>Stop	HBB:c.113G>A
Codon 37 (G->A); TGG(Trp)->TGA(stop codon) beta0	beta 37(C3) Trp>Stop	HBB:c.114G>A
Codon 39 (C->T); CAG(Gln)->TAG(stop codon) beta0	beta 39(C5) Gln>Stop	HBB:c.118C>T
Codon 43 (G->T); GAG(Glu)->TAG (stop codon) beta0	beta 43(CD2) Glu>Stop	HBB:c.130G>T
C859 (AAG>TAG)	beta 59(E3) Lys>Stop	HBB:c.178A>T
Codon 61 (A->T); AAG(Lys)->TAG(stop codon) beta0	beta 61(E5) Lys>Stop	HBB:c.184A>T
Codon 90 (G->T); GAG(Glu)->TAG(stop codon) beta0	beta 90(F6) Glu>Stop	HBB:c.271G>T
Codon 112 (T->A); TGT(Cys)->TGA(stop codon) beta0	beta 112(G14) Cys>Stop	HBB:c.339T>A
Codon 121 (G->T); GAA(Glu)->TAA(stop codon) beta0 (dominant beta-thal trait)	beta 121(GH4) Glu>Stop	HBB:c.364G>T

Figure 20 : Dataset for β^0 mutants in exons of HBB



HbVar: A database of Human Hemoglobin Variants and Thalassemias

Query Results
There are 12 matches to your query
Query description: category = Thalassemias AND type of thal in (beta+) AND location in (exon) AND chain in (beta)

Name	Mutation	Mutation, HGVS nomenclature
Codon 10 (C->A); GCC(Ala)->GCA(Ala) beta+	beta 10(A7) Ala>Ala	HBB:c.33C>A
Hb Malay	beta 19(B1) Asn>Ser	HBB:c.59A>G
Codon 24 (T->A); GGT(Gly)->GGA(Gly) beta+	beta 24(B6) Gly>Gly	HBB:c.75T>A
Hb E	beta 26(B8) Glu>Lys	HBB:c.79G>A
Hb Knossos	beta 27(B9) Ala>Ser	HBB:c.82G>T
Hb Chesterfield	beta 28(B10) Leu>Arg	HBB:c.86T>G
IVS-I (-3) or codon 29 (C->T); GGC(Gly)->GGT(Gly) beta+	beta 29(B11) Gly>Gly	HBB:c.90C>T
Hb New Berlin	beta 30(B12) Arg>Trp	HBB:c.91A>T
Hb Cagliari	beta 60(E4) Val>Glu	HBB:c.182T>A
Hb Showa-Yakushiji	beta 110(G12) Leu>Pro	HBB:c.332T>C
Hb Hradec Kralove (or Hb HK)	beta 115(G17) Ala>Asp	HBB:c.347C>A
Hb Dhonburi	beta 126(H4) Val>Gly	HBB:c.380T>G

Figure 20 : Dataset for β^+ mutants in exons of HBB

Then, the Ensemble database https://asia.ensembl.org/index.html [1] was used to complete information about Accession number and position of each SNP on HBB gene .

Table 6 : Dataset for SNPs of thalassemia (β^0 type) in the exon regions of HBB

	Accession no	Position	HbVar Name	Normal	Mutant	Type mutation	Disease
1	rs34563000	5227021	HBB: c.1A>C / G	**A** T G Methionine	**C** T G Leucine	Missense	TM
				A T G Methionine	**G** T G Valine		
2	rs33941849	5227020	c.2T>G / C	A **T** G Methionine	A **G** G Arginine	Missense	TM /Hb SS
				A **T** G Methionine	A **C** G Threonine		TM
3	rs33930702	5227019	c.3G>C / A / T	A T **G** Methionine	A T **A /C /T** Isoleucine	Missense	TM
4	rs33930165	5227003	c.19G>T	**G** A G Glutamate	**T** A G Stop	Nonsense	TM
5	Rs334	5227002	c.20A>T	G **A** G Glutamate	G **T** G Valine	Missense	TM/ Hb E/ Hb SS
6	rs63750783	5226975	c.47G>A	T **G** G Tryptophan	T **A** G Stop	Nonsense	TM / Hb SS
7	rs34716011	5226974	c.48G>A	T G **G** Tryptophan	T G **A** Stop	Nonsense	TM
8	rs33986703	5226970	c.52A>T	**A** A G Lysine	**T** A G Stop	Nonsense	TM /Hb SS
9	rs33959855	5226955	c.67G>T	**G** A A Glutamate	**T** A A Stop	Nonsense	TM
10	rs33950507	5226943	c.79G>T	**G** A G Glutamate	**T** A G Stop	Nonsense	TM
11	rs35684407	5226931	c.91A>G	**A** G G Arginine	**G** G G Glycine	Missense	TM
12	rs33960103	5226930	c.92G>C / A	A **G** G Arginine	A **C** G Threonine	Missense	TM /Hb SS
				A **G** G Arginine	A **A** G Lysine		TM
13	rs33982568	5226784	c.108C>A	T A **C** Tyrosine	T A **A** Stop	Missense	TM
14	rs33991059	5226779	c.113G>A	T **G** G Tryptophan	T **A** G Stop	Nonsense	TM
15	rs33974936	5226778	c.114G>A	T G **G** Tryptophan	T G **A** Stop	Nonsense	TM

16	rs11549407	5226774	c.118C>T	**C** A G Glutamine	**T** A G Stop	Nonsense	TM / Hb SS
17	rs33922842	5226762	c.130G>T	**G** A G Glutamate	**T** A G Stop	Nonsense	TM /Hb SS
18	rs33969400	5226714	c.178A>T	**A** A G Lysine	**T** A G Stop	Nonsense	TM
19	rs33995148	5226708	c.184A>T / G	**A** A G Lysine	**T** A G Stop	Nonsense	TM
				A A G Lysine	**G** A G Glutamate	Missense	
20	rs33913712	5226621	c.271G>T	**G** A G Glutamate	**T** A G Stop	Nonsense	TM
				G A G Glutamate	**A** A G Lysine	Missense	
21	rs33933298	5226597	c.295G>A	**G** T G Valine	**A** T G Methionine	Missense	TM
22	rs33930977	5225703	c.339T>A	T G **T** Cysteine	T G **A** Stop	Nonsense	TM
23	rs33946267	5225678	c.364G>T/C	**G** A A Glutamate	**T** A A Stop	Nonsense	TM /Hb SS
				G A A Glutamate	**C** A A Glutamine	Missense	TM
24	rs33910569	5225659	c.383A>G	C A G Glutamine	C G G Arginine	Missense	TM
25	rs33953406	5225645	c.397A>T	**A** A G Lysine	**T** A G Stop	Nonsense	TM
26	rs34407387	5225623	c.419A>C	A **A** T Asparagine	A **C** T Threonine	Missense	Hb sagami

Table 7 : Dataset for SNPs of thalassemia (β^+ type) in the exon regions of HBB

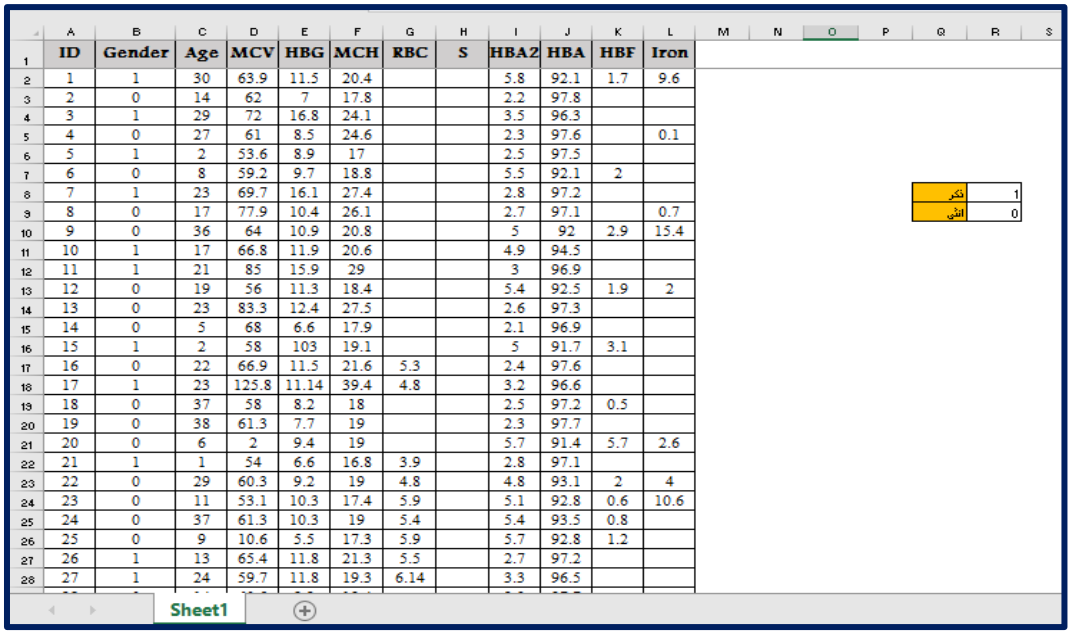
	Accession no	Position	HbVar Name	Normal	Mutant	Type mutation	Disease
1	rs35799536	5226989	HBB:c.33C>A	G C **C** Alanine	G C **A** Alanine	synonymous	TM
2	rs33972047	5226963	HBB:c.59A>G	A **A** C Asparagine	A **G** C Serine	Missense	TM / TI / Hb Malay
3	rs33951465	5226947	HBB:c.75T>A	G G **T** Glycine	G G **A** Glycine	Synonymous	TM /TI /Hb SS
4	rs33950507	5226943	HBB:c.79G>A	**G** A G Glutamate	**A** A G Stop	Nonsense	TM /TI /Hb SS/ HB E
5	rs35424040	5226940	HBB:c.82G>T	**G** C C Alanine	**T** C C Serine	Missense	TM /TI /Hb SS/ Hb Knossos
6	rs33916412	5226936	HBB:c.86T>G	C **T** G Leucine	C **G** G Arginine	Missense	Hemoglobin Chesterfield
7	rs35578002	5226932	HBB:c.90C>T	G G **C** Glycine	G G **T** Glycine	Synonymous	TM
8	rs33931779	5226710	HBB:c.182T>A	G **T** G Valine	G **A** G Glutamate	Missense	TM / Hb Cagliari
9	rs35256489	5225710	HBB:c.332T>C	C **T** G Leucine	C **C** G Proline	Missense	TM / TI / Hb Showa-Yakushiji
10	rs35485099	5225695	HBB:c.347C>A	G **C** C Alanine	G **A** C Aspartate	Missense	Hb Hradec Kralove or Hb HK
11	rs33925391	5225662	HBB:c.380T>G	G **T** G Valine	G **G** G Glycine	Missense	TM /TI /Hb Dhonburi

Hb SS : is common in African and Mediterranean populations.. The phenotype of the β-globin gene defect determines the severity of the co-inherited sickle cell mutation; β0 results in a severe disease, while β+ causes a milder clinical picture of the disease Observed when there is a variation in β chain and leads to sickle cell anemia (SCD) [4]

Hb E : is Common in India, Bangladesh, and throughout Southeast Asia , it has become the most common form of β-thalassemia detected through many newborn screening programs. It is clinically characterized by marked variability, ranging from mild asymptomatic anemia to a life-threatening disorder requiring transfusions from infancy [4].

- Second dataset was retrieved from : https://data.mendeley.com/datasets/p8rv84hrbs

This dataset contain HLPC test information for 1073 patients that collected in 2022 from Hematology Center (Thalassemia) in Ibn Al-Baladi Hospital for various thalassemia patients in Iraq containing test features (MCV- HGB- MCH- RBC- HBA2- HBA- HBF) to create a thalassemia Diagnosis system.



ID	Gender	Age	MCV	HBG	MCH	RBC	S	HBA2	HBA	HBF	Iron
1	1	30	63.9	11.5	20.4			5.8	92.1	1.7	9.6
2	0	14	62	7	17.8			2.2	97.8		
3	1	29	72	16.8	24.1			3.5	96.3		
4	0	27	61	8.5	24.6			2.3	97.6		0.1
5	1	2	53.6	8.9	17			2.5	97.5		
6	0	8	59.2	9.7	18.8			5.5	92.1	2	
7	1	23	69.7	16.1	27.4			2.8	97.2		
8	0	17	77.9	10.4	26.1			2.7	97.1		0.7
9	0	36	64	10.9	20.8			5	92	2.9	15.4
10	1	17	66.8	11.9	20.6			4.9	94.5		
11	1	21	85	15.9	29			3	96.9		
12	0	19	56	11.3	18.4			5.4	92.5	1.9	2
13	0	23	83.3	12.4	27.5			2.6	97.3		
14	0	5	68	6.6	17.9			2.1	96.9		
15	1	2	58	103	19.1			5	91.7	3.1	
16	0	22	66.9	11.5	21.6	5.3		2.4	97.6		
17	1	23	125.8	11.14	39.4	4.8		3.2	96.6		
18	0	37	58	8.2	18			2.5	97.2	0.5	
19	0	38	61.3	7.7	19			2.3	97.7		
20	0	6	2	9.4	19			5.7	91.4	5.7	2.6
21	1	1	54	6.6	16.8	3.9		2.8	97.1		
22	0	29	60.3	9.2	19	4.8		4.8	93.1	2	4
23	0	11	53.1	10.3	17.4	5.9		5.1	92.8	0.6	10.6
24	0	37	61.3	10.3	19	5.4		5.4	93.5	0.8	
25	0	9	10.6	5.5	17.3	5.9		5.7	92.8	1.2	
26	1	13	65.4	11.8	21.3	5.5		2.7	97.2		
27	1	24	59.7	11.8	19.3	6.14		3.3	96.5		

ذكر	1
انثي	0

Sheet1

Figure 21 : Dataset of HLPC test information for various thalassemia patients in Iraq

MCV stands for mean corpuscular volume with Normal Ranges (82.5- 98 fl) for adults. An MCV blood test measures the average size of your red blood cells.

HGB: Hemoglobin with Normal Ranges: (for Males 13.6 - 16.9, Females 11.9 - 14.8) Grams per deciliter

MCH stands for Mean Corpuscular Hemoglobin with Normal Ranges (27 – 32 pg), and is a calculation of the average amount of hemoglobin contained in each of a person's red blood cells

RBC: Red Blood Cell with Normal Ranges: (for Males 4.2 - 5.7, Females 3.8 - 5.0) $10^{\wedge 6}$/microliter.

HBA2: Subunits of HLPC Test with Normal Range (1-3) % for adults

HBA: Subunits of HLPC Test with Normal Range (> 97) % for adults

HBF: Subunits of HLPC Test with Normal Range (<1) % for adults

Iron: Iron in Blood cells with Normal Range (60 – 170) micrograms per deciliter

3- Tools and methods

Six tools were chosen to perform our study for the following reasons :

1- To design Allele specific primers and allele flanking primers for detecting SNPs and mutations, we found BatchPrimer3 and WASP (a Web-based Allele-Specific PCR assay designing) tools. We used **BatchPrimer3** because this tool was the only tool that create comprehensive results which validated them via PCR .

2- To perform in silico PCR , we found Silica , UCSC in silico PCR and Fast-PCR tools . Fast-PCR is offline program with only 10-day Free Trial , so we used **Silica** instead UCSC in silico PCR because this tool has colorful interface and give us not only information about amplicons and product size and primer melting temperatures but also primer binding sites .

3- To perform in silico electrophoresis for evaluating SNP primers and ensuring PCR product size results , we found DNASTAR and SnapGene tools. we used **SnapGene** tool with 30-day free trial because DNASTAR has only 14-day free trial and more difficult to use .

4- To predict secondary structure of protein, we found many tools but we choose **SOPMA** because it is online server , easy to use and you don't need to inter your E-mail and password and wait 24 hours to receive results on it .

5- To perform molecular docking between protein and ligand, we found SwissDOCK and PatchDOCK tools . PatchDOCK page wasn't working so, we choose **SwissDOCK** because it is online server, easy to use and get results in 30 minutes .

6- To diagnose the severity of the Thalassemia disease via expert system , we used **Fuzzy Inference System** from matlab platform because it has computationally efficient and well-suited to mathematical analysis .

3-1 : BatchPrimer3 :

The nucleotide sequences for the studied SNPs were retrieved from the SNP database of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). The input data were the FASTA sequence of the gene regions, SNP alleles and design constraints https://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi .

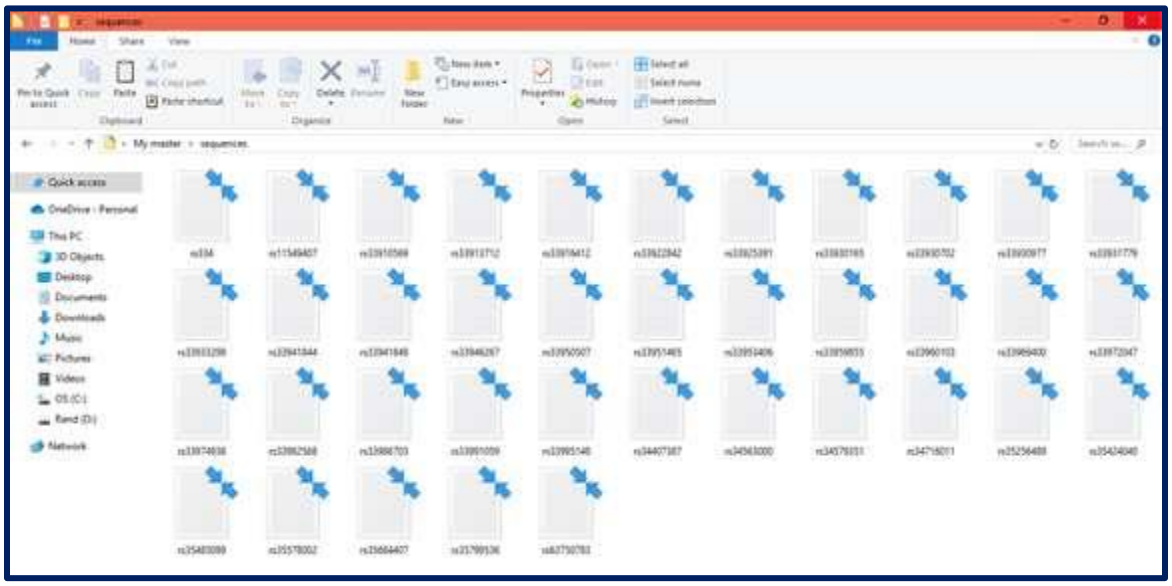


Figure 22 : DNA sequences of studied SNPs

Allele specific primers and allele flanking primers were selected from the primer type pull-up combo-box . When inputting a FASTA file or a single sequence, a header line starting with ">" is mandatory for each sequence. Alleles were indicated by IUPAC codes (G/C: S, A/T: W, G/A: R, T/C: Y, G/T: K, A/C: M).

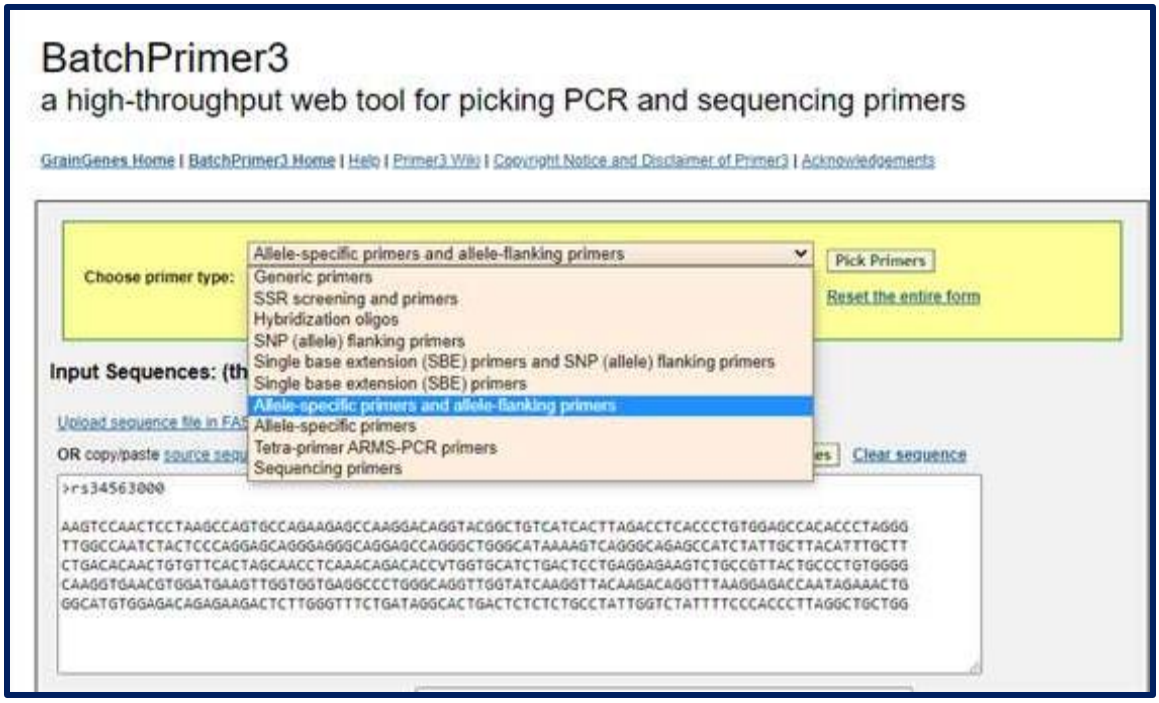
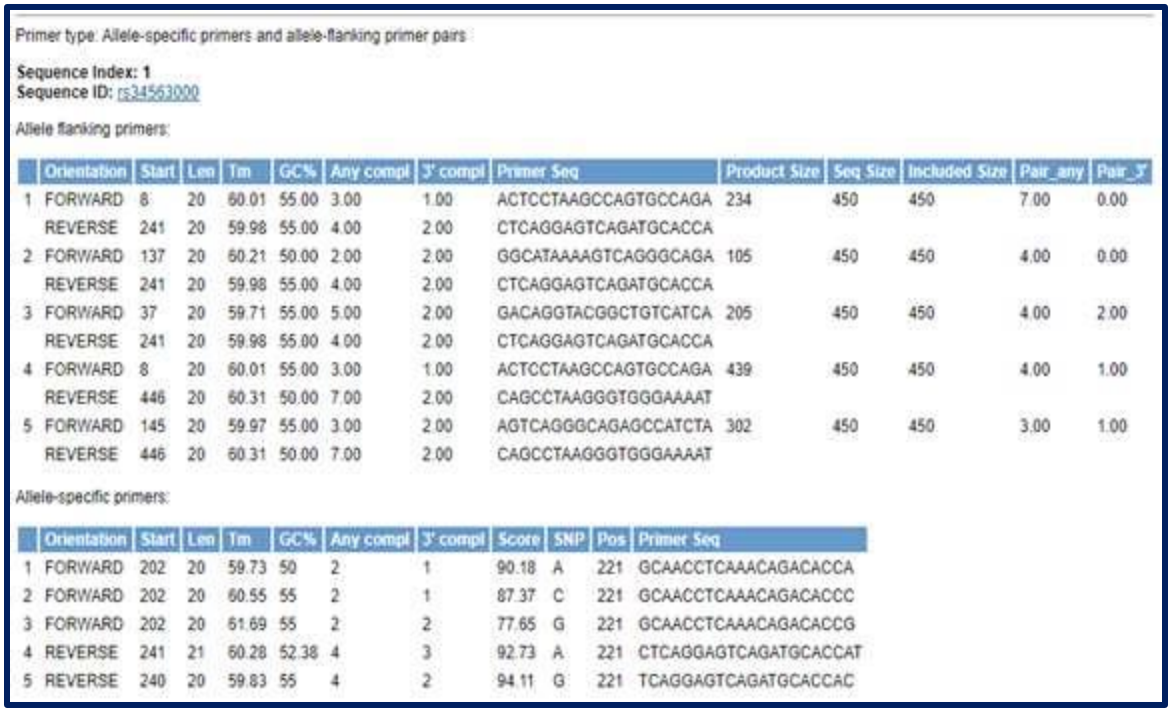


Figure 23 : Web interface of BatchPrimer3 v1.0 application

The design output is the compatible primer sequences and amplification product sequences. Information is provided about oligonucleotides, such as size, melting temperature among others. The BatchPrimer3 program produces four parts of outputs: (1) a main HTML page containing the primer design summary of all input sequences, (2) an HTML table page listing all designed primers and primer properties, (3) a tab-delimited text file with the same contents in the HTML table page, and (4) a detailed primer view page for each sequence with successfully designed primers .



Primer type: Allele-specific primers and allele-flanking primer pairs

Sequence Index: 1
Sequence ID: rs34563000

Allele flanking primers:

	Orientation	Start	Len	Tm	GC%	Any compl	3' compl	Primer Seq	Product Size	Seq Size	Included Size	Pair_any	Pair_3'
1	FORWARD	8	20	60.01	55.00	3.00	1.00	ACTCCTAAGCCAGTGCCAGA	234	450	450	7.00	0.00
	REVERSE	241	20	59.98	55.00	4.00	2.00	CTCAGGAGTCAGATGCACCA					
2	FORWARD	137	20	60.21	50.00	2.00	2.00	GGCATAAAAGTCAGGGCAGA	105	450	450	4.00	0.00
	REVERSE	241	20	59.98	55.00	4.00	2.00	CTCAGGAGTCAGATGCACCA					
3	FORWARD	37	20	59.71	55.00	5.00	2.00	GACAGGTACGGCTGTCATCA	205	450	450	4.00	2.00
	REVERSE	241	20	59.98	55.00	4.00	2.00	CTCAGGAGTCAGATGCACCA					
4	FORWARD	8	20	60.01	55.00	3.00	1.00	ACTCCTAAGCCAGTGCCAGA	439	450	450	4.00	1.00
	REVERSE	446	20	60.31	50.00	7.00	2.00	CAGCCTAAGGGTGGGAAAAT					
5	FORWARD	145	20	59.97	55.00	3.00	2.00	AGTCAGGGCAGAGCCATCTA	302	450	450	3.00	1.00
	REVERSE	446	20	60.31	50.00	7.00	2.00	CAGCCTAAGGGTGGGAAAAT					

Allele-specific primers:

	Orientation	Start	Len	Tm	GC%	Any compl	3' compl	Score	SNP	Pos	Primer Seq
1	FORWARD	202	20	59.73	50	2	1	90.18	A	221	GCAACCTCAAACAGACACCA
2	FORWARD	202	20	60.55	55	2	1	87.37	C	221	GCAACCTCAAACAGACACCC
3	FORWARD	202	20	61.69	55	2	2	77.65	G	221	GCAACCTCAAACAGACACCG
4	REVERSE	241	21	60.28	52.38	4	3	92.73	A	221	CTCAGGAGTCAGATGCACCAT
5	REVERSE	240	20	59.83	55	4	2	94.11	G	221	TCAGGAGTCAGATGCACCAC

Figure 24 : The primer design results of sequence ID (rs34563000) for allele flanking primers and allele specific primers

3-2: Silica (in silico PCR)

The aim of in silico PCR is to provide an easy way to obtain the theoretical PCR results we may expect from DNA and check quality primers . Ideally designed primer pairs will ensure the efficiency and specificity of the amplification reaction, resulting in a high yield of the desired amplicon. Important criteria such as primer-sequence, –length, and-melting temperature (Tm) are fundamental for the selection of primers and amplification of targeted nucleotide sequences from a DNA template .

After the primers were obtained , Two primers methods were selected for mutant / wild type identification or hetero / homozygosity identification : (forward allele specific primer for mutant / wild and common reverse primer from allele flanking primer pair **or** reverse allele specific primer for mutant / wild and common forward primer from allele flanking primer pair) .using

silica application (https://www.gear-genomics.com/silica/) , All primers were tested and the best result was chosen .

When choosing two PCR amplification primers, the following guidelines should be considered : Primers should be at least 18 nucleotides in length to minimize the chances of encountering problems with a secondary hybridization site on the vector or insert ,the optimal melting temperatures for primers in the range 52-58° C, generally produce better results than primers with lower melting temperatures, primers with melting temperatures above 65° C should also be avoided because of potential for secondary annealing , a good working approximation of this value (generally valid for oligos in the 18–30 base range) can be calculated using the formula of Wallace et al. (1979), Tm = 2(A+T) + 4(G+C), Primers should have a GC content between 45 and 60 percent , a "G" or "C" is desirable at the 3' end but the first part of this rule should apply, this GC clamp reduces spurious secondary bands , primers must be chosen so that they have a unique sequence within the template DNA that is to be amplified [35] .

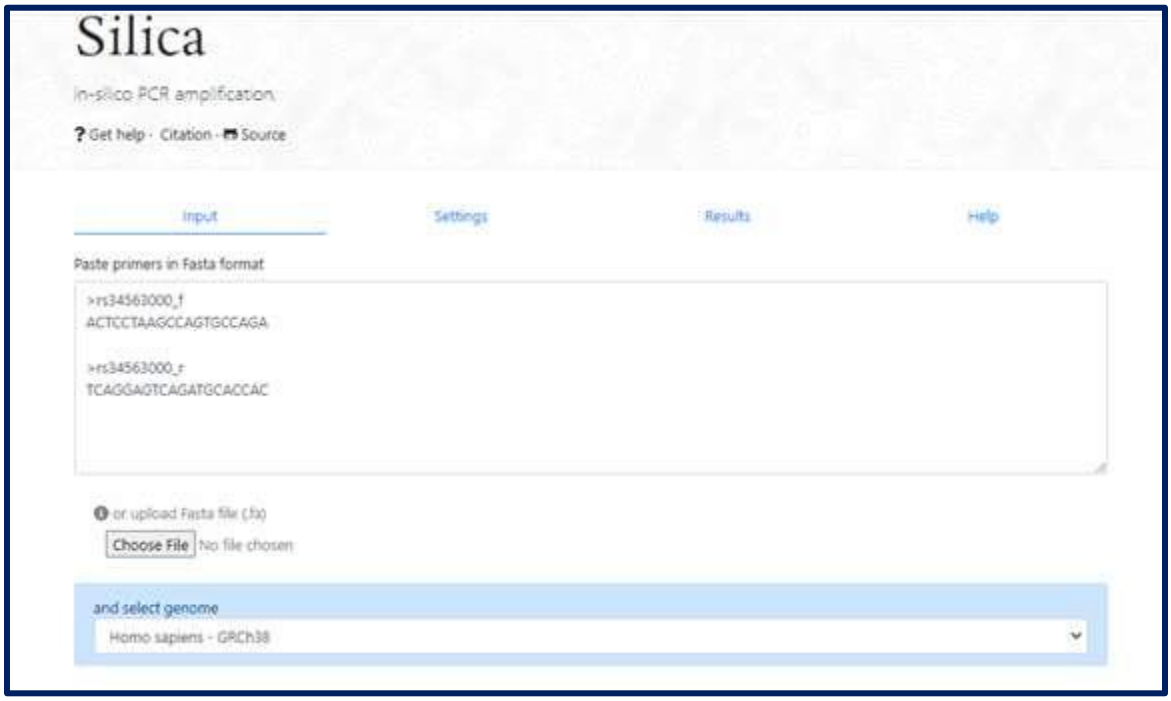


Figure 25 : Web interface of Silica application

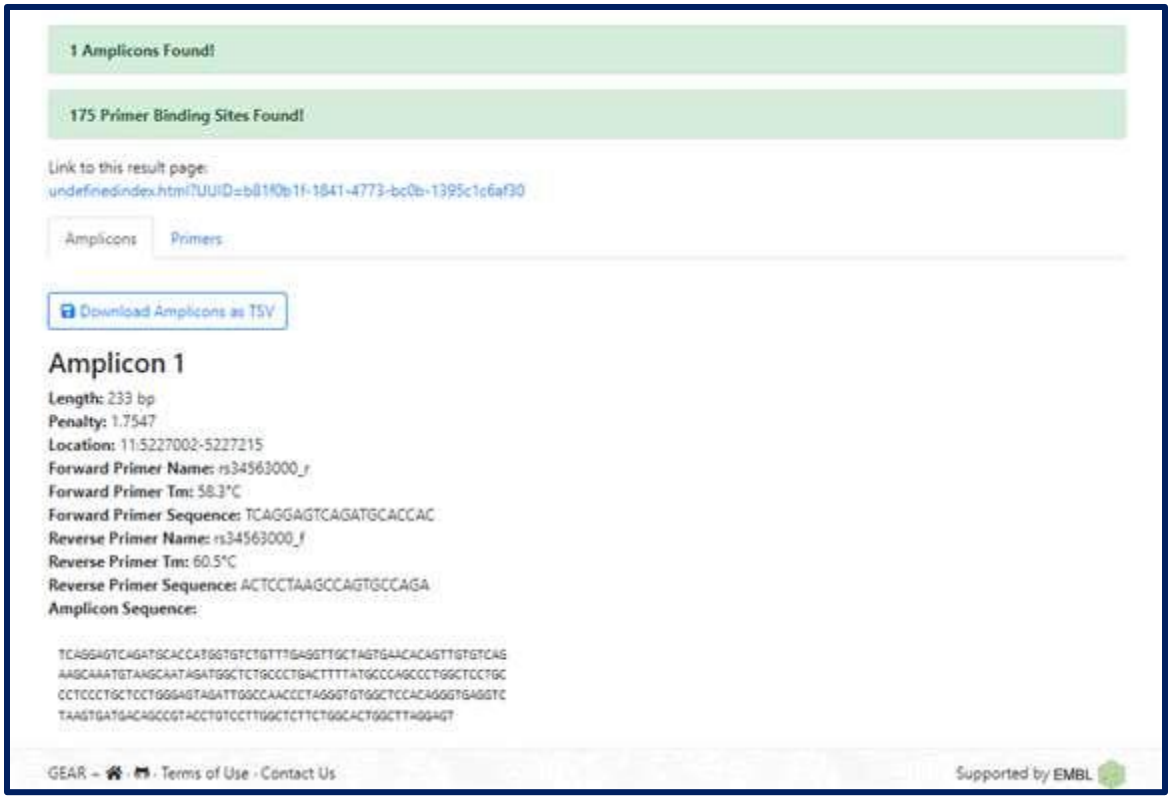


Figure 26 : The Silica application result

3-3: SnapGene (in silico electrophoresis)

It allows the simulation of Agarose gel, PCR Amplification, Restriction digestion. It is a large collection of MW markers.

The sequence of HBB gene was inserted in SnapGene tool (https://www.snapgene.com/free-trial) ,Then the 36 candidated primers results (each wild type with common primer) was added to evaluate them and ensure that PCR product size results was the same .

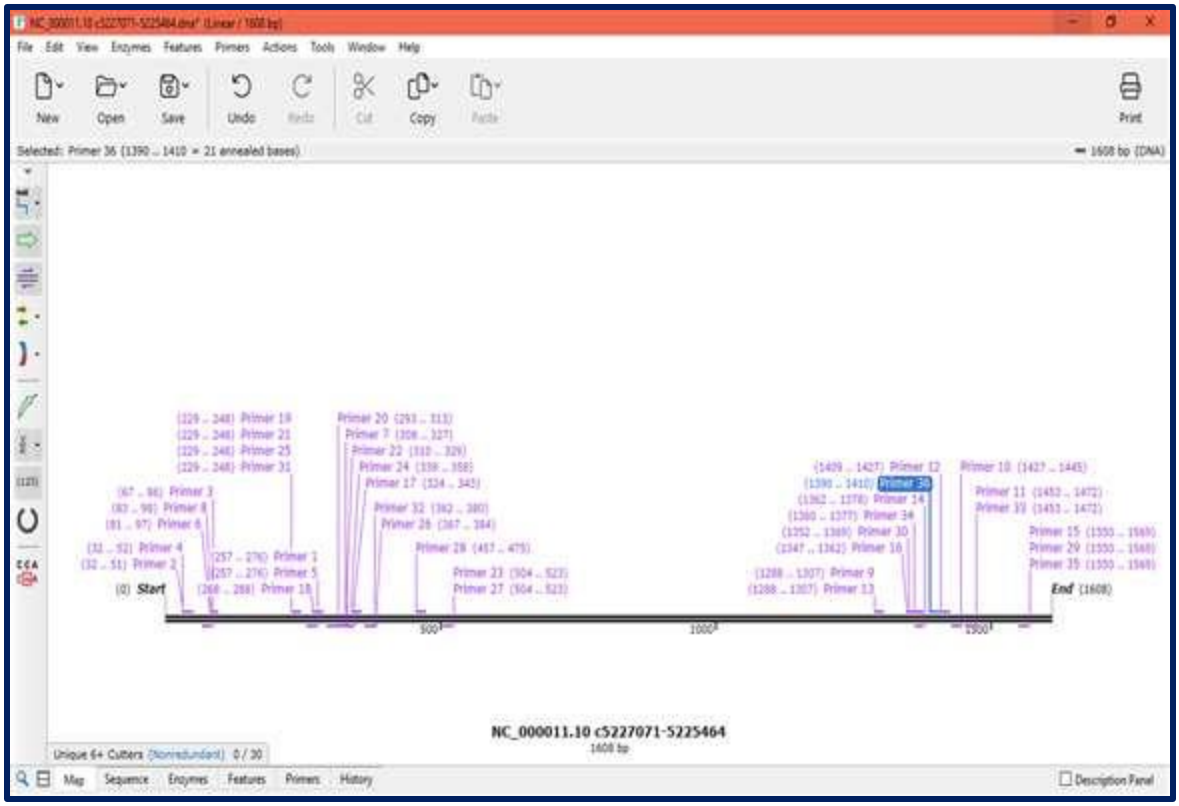


Figure 27 : Position of candidated primers on HBB gene

The simulate agarose gel was done , The suitable MW marker(1 kb plus DNA ladder) was selected and 2,5 % agarose was used to gives good resolution for small fragments .

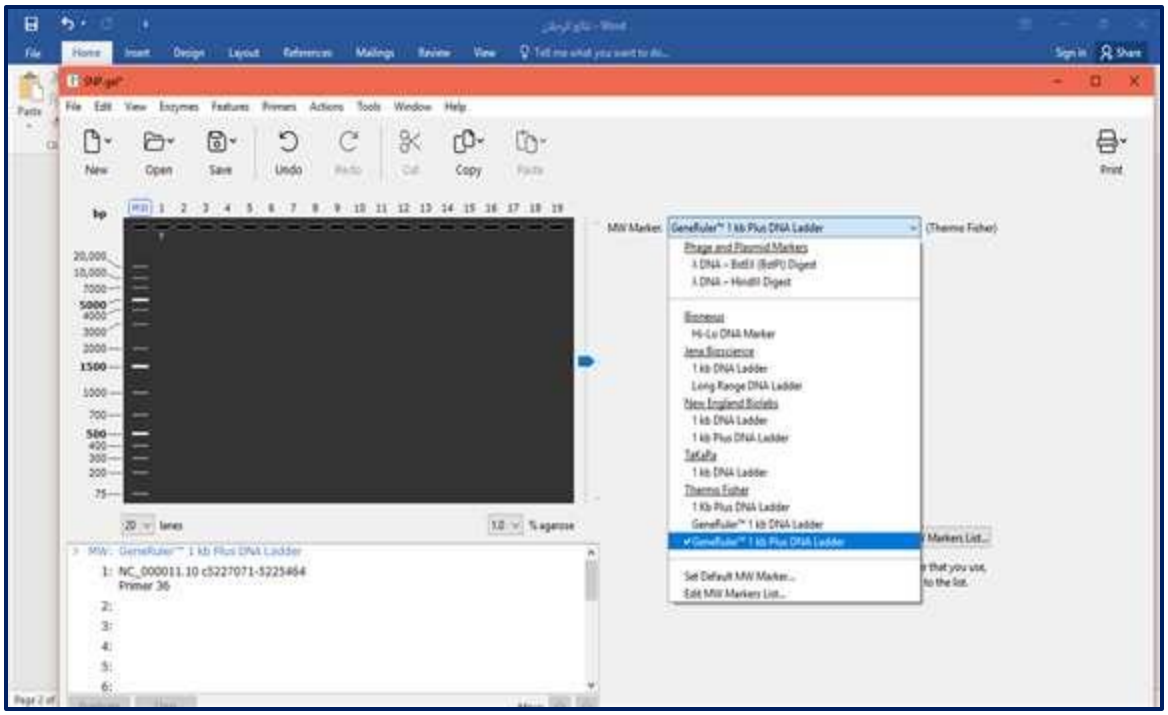


Figure 28 : Parameters for simulate agarose gel of SnapGene

3-4 : SOMPA

The protein sequences for HBB gene were retrieved from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/protein/NP_000509.1) and pasted it in SOMPA sever (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html) to determine proteins complex structures which is valuable to understand the biological process at the atomic level and thus develop therapeutic interventions or drugs targeting these interactions [39] .

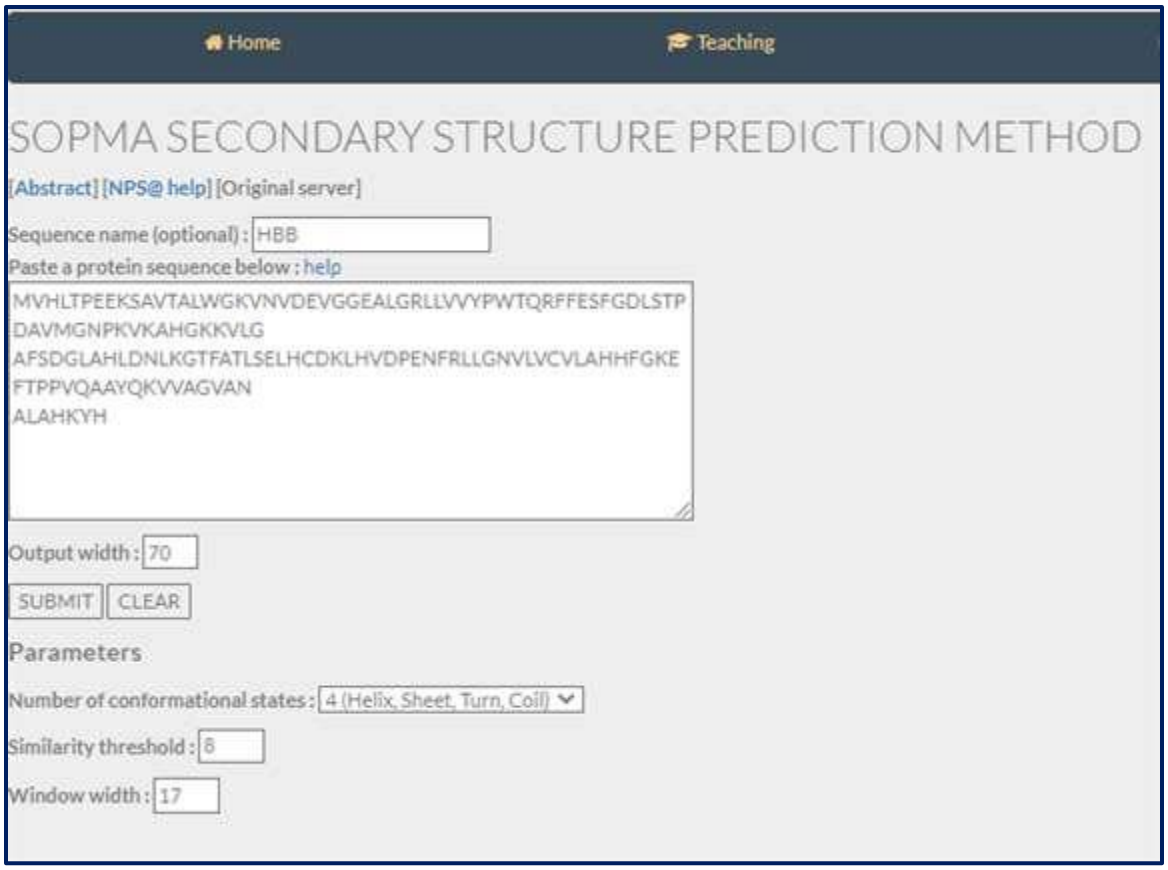


Figure 29 : Web interface of SOPMA

3-5 : SwissDOCK

SwissDOCK is web service to predict the molecular interactions that may occur between a target protein and a small molecule (http://www.swissdock.ch/)

A target protein structure (Hbb) was determined either by specifying its identifier from the Protein Data Bank (1DXT) or by uploading structure files .

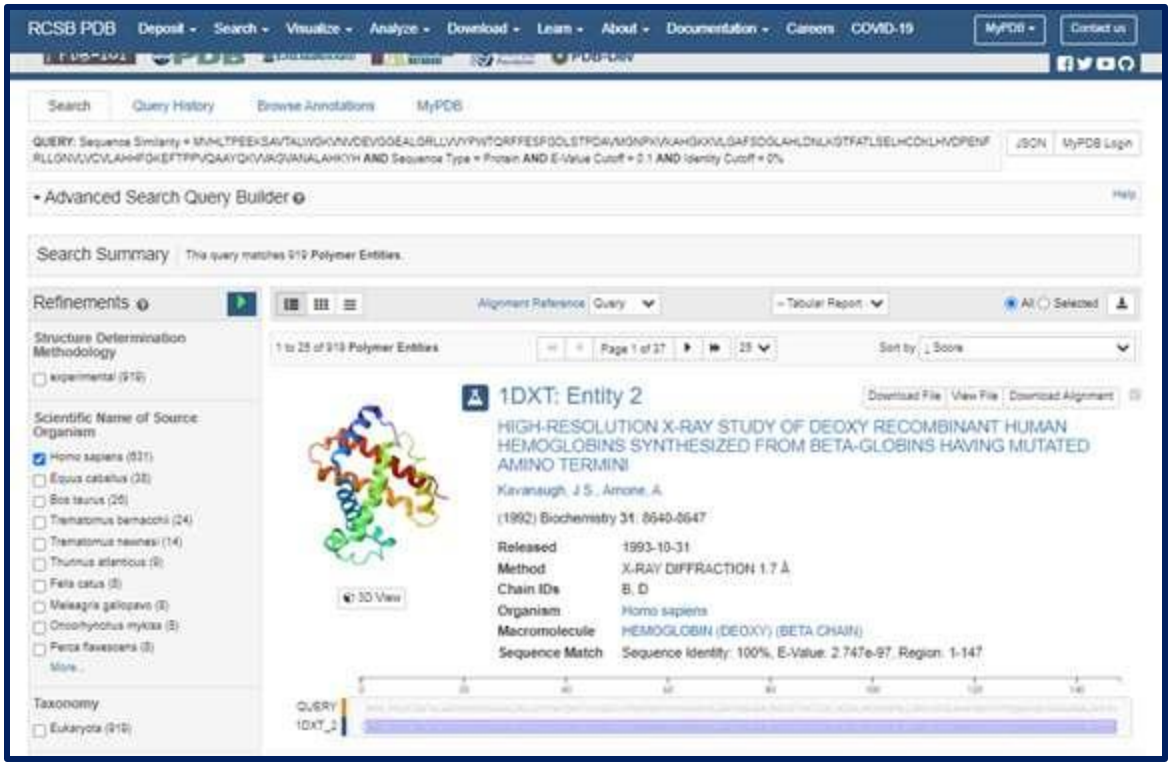


Figure 30 : Target protein structure 1DXT from PDB

Ligands selection (Indicaxanthin, Hydroxyurea , Berberine , Coptisine) were defined either by specifying its identifier from the ZINC database or by uploading structure files from Pupchem . The ligand is immediately set up after its definition and its prepared structure can be downloaded and reviewed prior to the docking assay, if needed. After a docking assay had been submitted, a link of the docking result web page will be sent once the docking was completed.

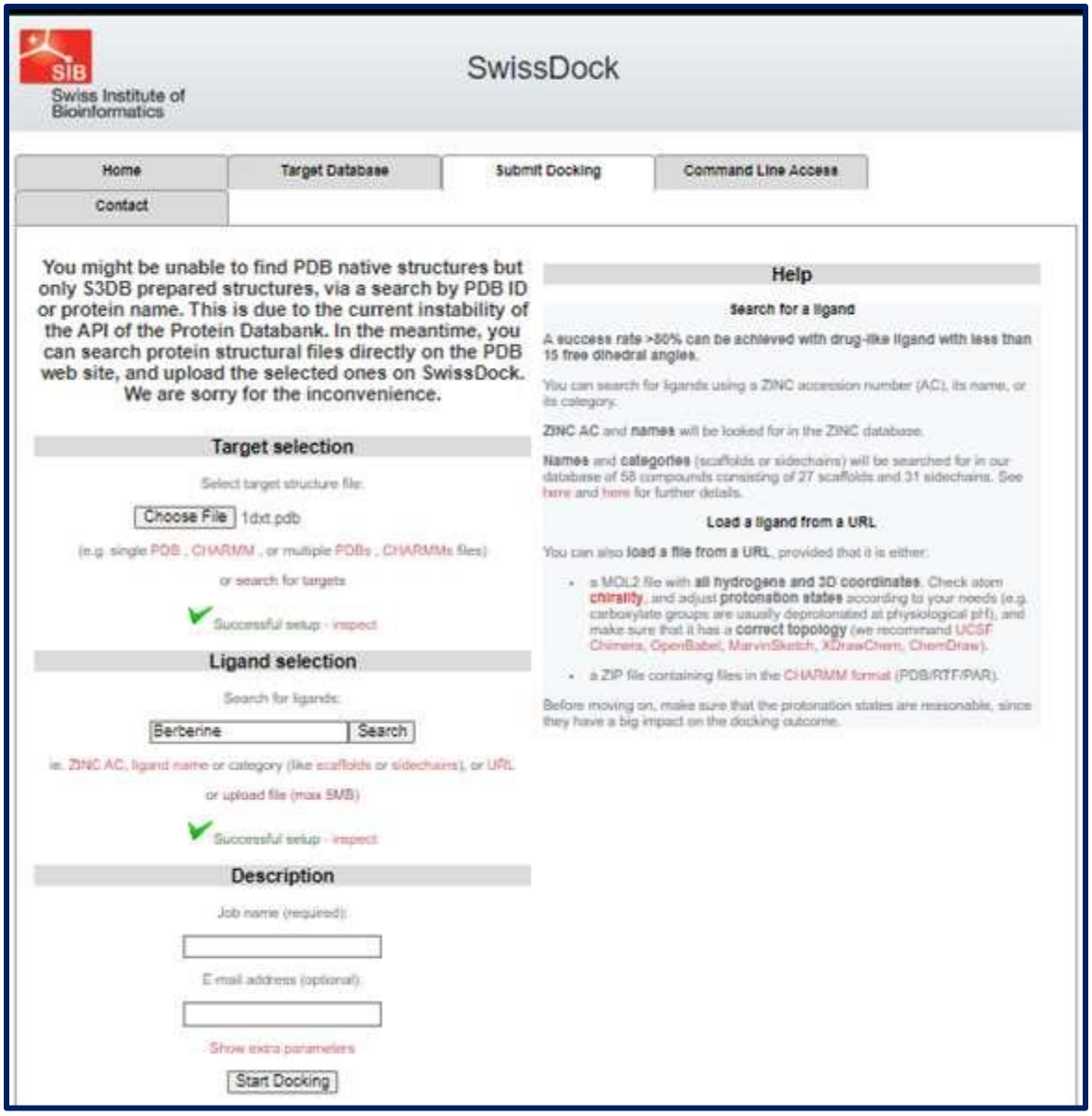


Figure 31 : Web interface of the SwissDOCK server

3-6 : Fuzzy inference system (fis)

MATLAB (a programming and numeric computing platform used by millions of engineers and scientists to analyze data, develop algorithms, and create models) was downloaded on PC from (https://matlab.mathworks.com/) then , mamdani fuzzy logic system was designed by next steps :

1- Problem specification and linguistic variables were definded , There were 3 input variables and 1 output variable.

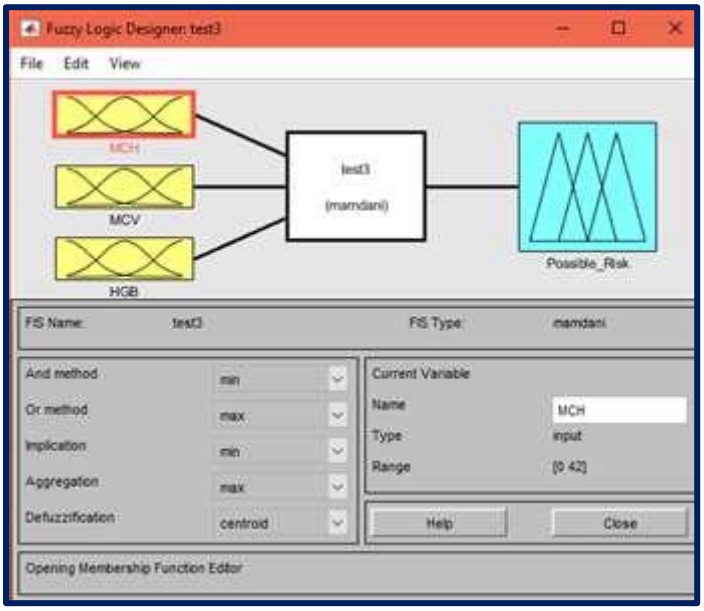


Figure 32 : Mamdani fuzzy inference system for thalassemia diagnosis

Table 8: Linguistic Variable for Input Variables

Input Variables	***Linguistic Variables***
Mean corpuscular hemoglobin (pg)	MCH
Mean Corpuscular Volume (fl)	MCV
Hemoglobin (g/dl)	HGB

Table 9 : Linguistic Variables for Output Variables

Output Variable	***Linguistic Variables***
Thalassemia Type	Possible_Risk

2- Ranges and values for input/output fields of the system were described .

Table 10 : Values for all Input Linguistic Variables

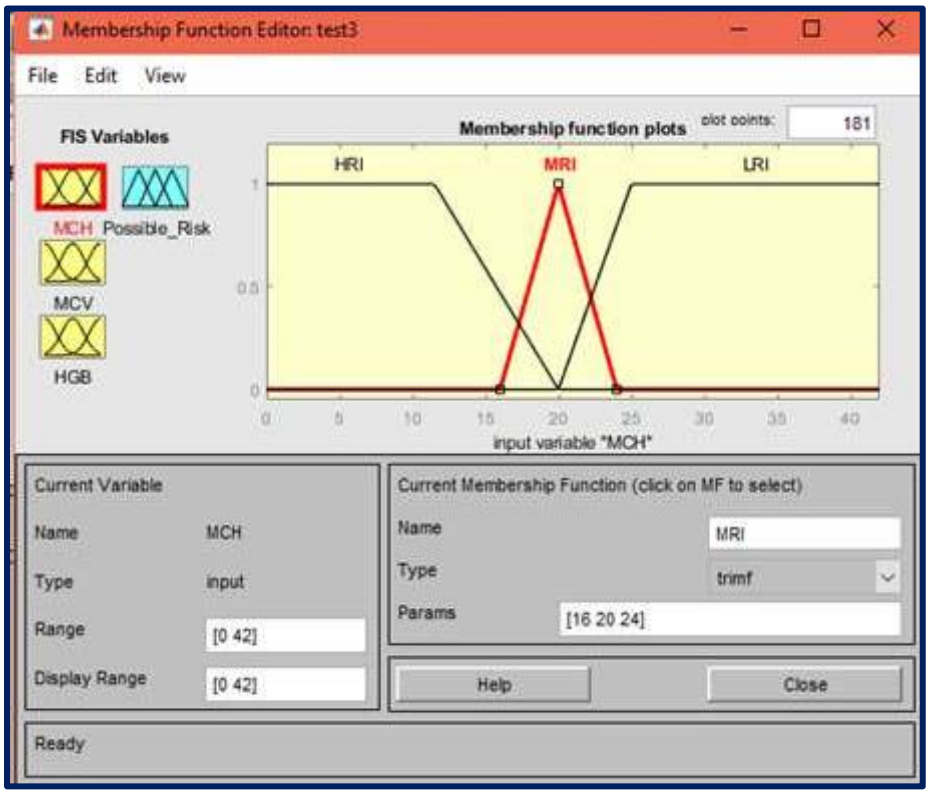
Linguistic Variables	***Ranges***	***Values***
MCH	<20 pg	HRI
	16-24 pg	MRI
	>20 pg	LRI
MCV	< 70 fl	HRI
	50-80 fl	MRI
	>60 fl	LRI
HGB	< 7 g/dl	HRI
	7 – 10 g/dl	MRI
	9 – 12 g/dl	LRI

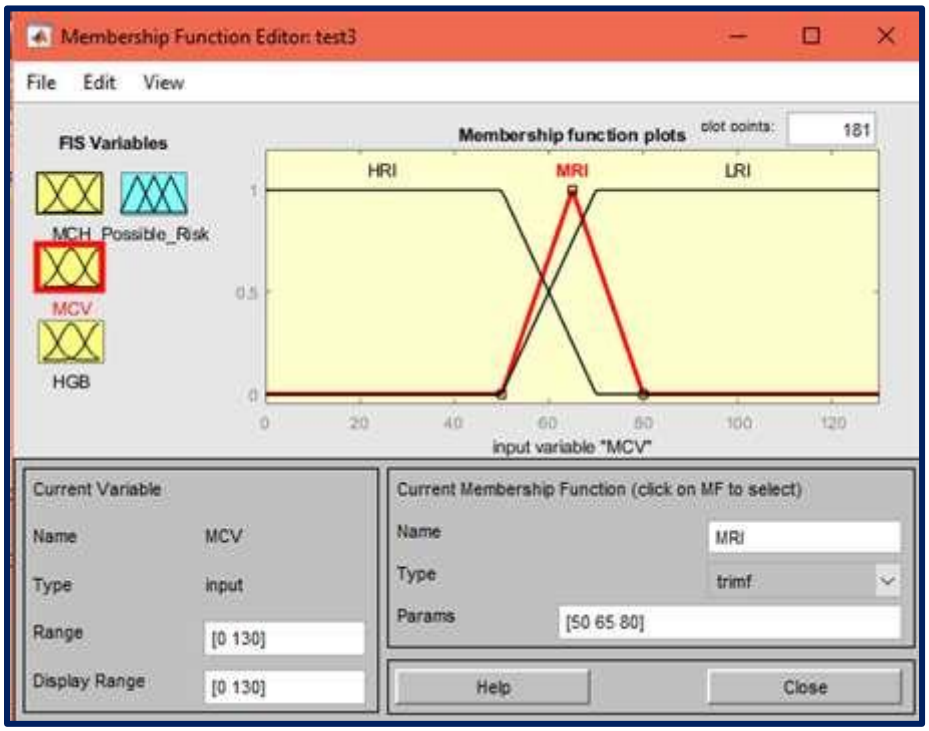
HRI=High Risk Interval, MRI=Moderate Risk Intervals, LRI=Low Risk Intervals.

Table 11 : Values for all Output Linguistic Variables

Linguistic Variables	***Ranges***	***Values***
Possible_Risk	HGB is 9 –12 g/dl	Thalassmia_Minor
	MCV is <80 fl	
	MCH is <27 pg	
	HGB is 7 –10 g/dl	Thalassemia_Intermedia
	MCV is 50 – 80 fl	
	MCH is 16 – 24 pg	
	HGB is < 7 g/dl	Thalassemia_Major
	MCV is >50 <70 fl	
	MCH is >12 <20 pg	

3- Membership functions of the variables were presented.





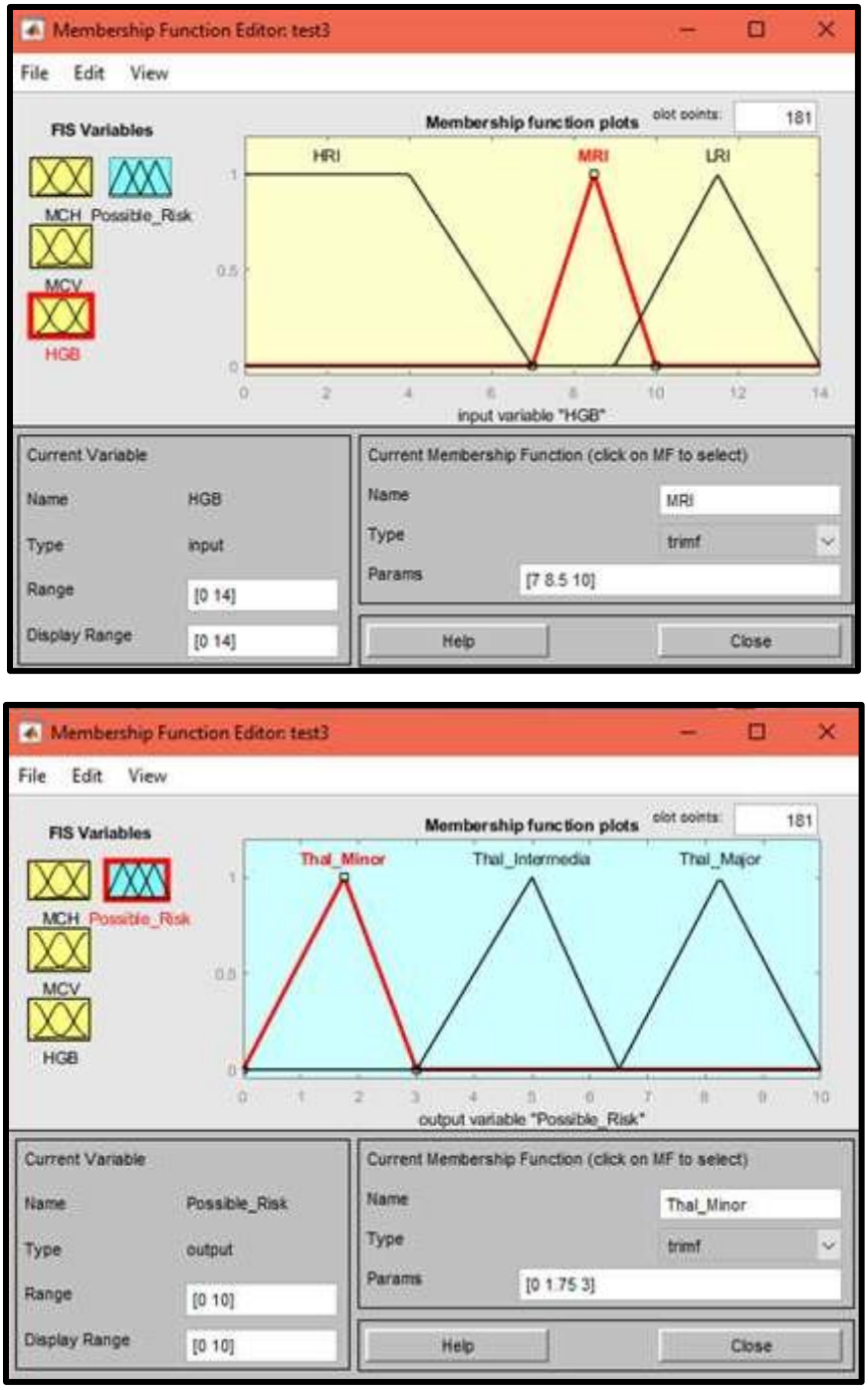


Figure 33 : Fuzzy Membership for all Input and Output Variables

The output will be a value within the range [0, 10]. The value 0 means that no Thalassemia problems exist as of yet. We have divided this range into smaller fuzzy sets to make a cluster of type of Thalassemia disease.

“Thal_Minor‟ (Thalassemia Minor) is given to those patients whose output value is in between 0 and 3.0 “Thal_Intermedia‟ (Thalassemia Intermedia) is given to the patients who gets a value between 3.0 and 6.5 also “Thal_Major‟ (Thalassemia Major) is given to the patients who gets a value between 6.5 and 10, The basic relationship is that the higher the severity of Thalassemia disease, the higher the output value.

4- Fuzzy Rules were Defined , in this section fuzzy inference rules generated; relevant inference rules can be determined by experience human operators well, as we have 26 rules conveniently are represented in IF- ELSE Form :

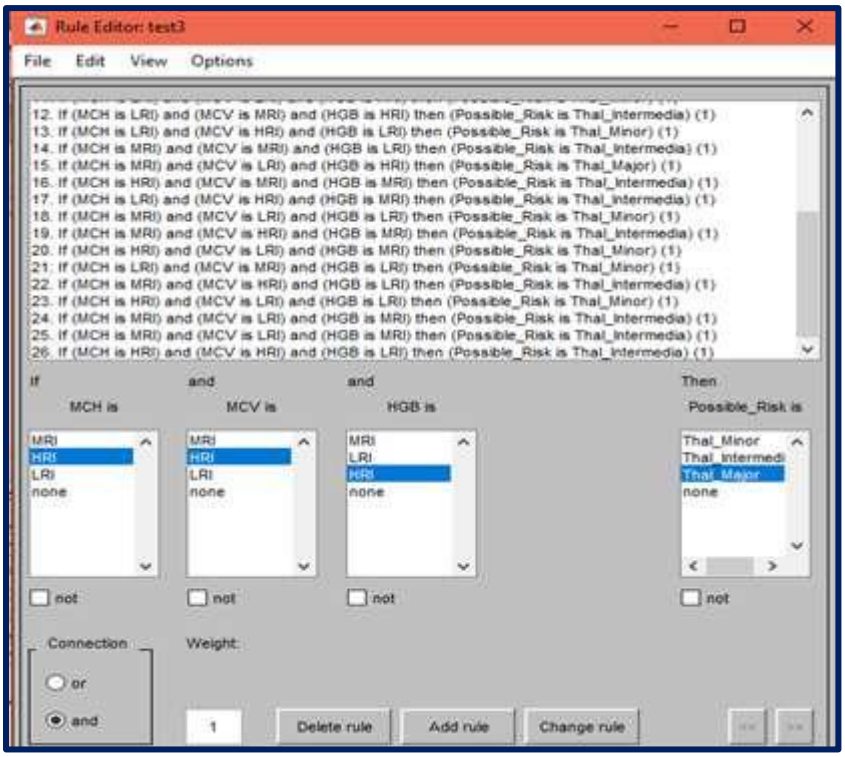


Figure 34 : Rule viewer for generated rules

The first three columns of plots (the yellow plots) show the membership functions referenced by the antecedent, or the if-part of each rule. The fourth column of plot (the blue plots) shows the membership functions referenced by the consequent, or the then-part of each rule. The plots which are blank in the if-part of any rule correspond to the characterization of none for the variable in the rule. The last plot in the fourth column of plots represents the aggregate weighted decision for the given inference system. This decision will depend on the input values for the system. The defuzzified output is displayed as a bold vertical line on this plot. The variables and their current values are displayed on top of the columns. In the lower left, there is a text field Input for entering specific input values.

To test the accuracy of our system , data was collected in 2022 from Hematology Center (Thalassemia) in Ibn Al-Baladi Hospital which was included 1073 patients with 12 features (https://data.mendeley.com/datasets/p8rv84hrbs) , Subsequently this data was checked for the presence of error in data entry including misspellings and missing data and filtered to 646 patients with 3 features (MCH,MCV,HGB). These cases were diagnosed by specialized local doctors .

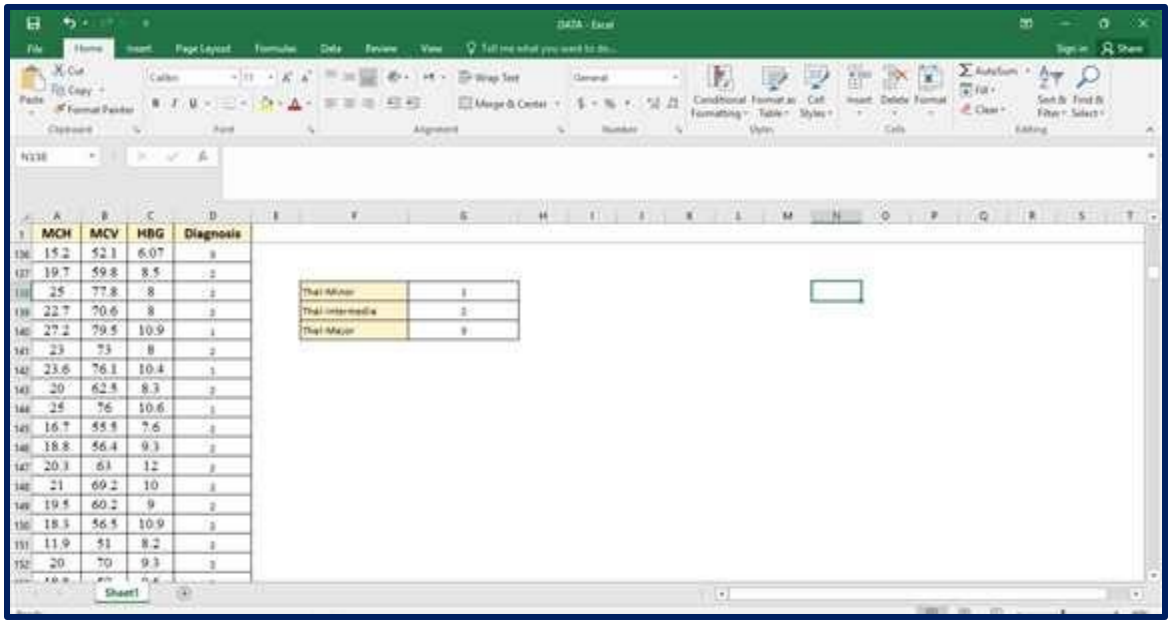


Figure 35 : Dataset for evaluating Fuzzy inference system

The fuzzy expert system was tested on dataset by recall code which had written by C languages in command window (Figure 36). Then, accurate classification presentage and number of error classification were calculated to compare our program results with doctor‘s diagnosis .

```
clc
clear
a=xlsread('DATA.xlsx');
input=a(:,1:3);
target=a(:,4);
A = readfis('BetaThalassemia');
Classf = evalfis(input, A) ;
for i=1:length(Classf)
    if Classf(i)<=3.2
        Classf(i)=1;
    else if Classf(i)>6.3
        Classf(i)=3;
        else
        Classf(i)=2;
        end
    end

end
Base =1:length(Classf);
subplot(2,1,1)
plot(Base,target,'b-',Base,Classf,'r-')
title('Classification of Beta Thalassemia')
xlabel('Case num')
ylabel('Possibel Risk')
legend('target','Class')
axis([0 650 -2 4])
grid
diff=(target-Classf)./target.*100;
subplot(2,1,2)
plot(diff,'k')
ylabel('Error')
xlabel('Case num')
legend('Error')
axis([0 650 -100 100])
grid
for i=1:length(Classf)
    if diff(i)>0 || diff(i) <0
        diff (i)=1;
    else
            diff(i)=0;
    end
end
diff;
Num_of_error_classification=sum(diff)
Accurate_classification_precentage=(1-
(Num_of_error_classification/length(Classf)))*100
```

Figure 36 : Code for evaluating fuzzy expert system results

Finally , These results were presented on confusion matrix , for this study the confusion matrix was a 3 x 3 owing to the three labels for the output class-risk of Thalassemia, namely: low, moderate and high risk.

Table 12 : The diagram of the confusion matrix that was used for evaluating the performance Fuzzy inference system

Confusion Matrix		Predicted		
		Minor	Intermedia	Major
Expected	Minor	A	B	C
	Intermedia	D	E	F
	Major	G	H	I

Accuracy, precision, recall and f-score were calculated from The formula :

Accuracy : is the total number of correct classifications (positive and negative). Accuracy : A + E + I / All
Error rate : 1- Accuracy

Precision : is the proportion of the predicted cases that were correctly predicted. Precision $_{minor}$: A / A+D+G , Precision $_{Intermedia}$: E / B+E+H , Precision $_{Major}$: I / C+F+I

Recall / Sensitivity : is the proportion of actual cases that were correctly predicted. TP $_{minor}$: A / A+B+C , TP $_{Intermedia}$: E / D+E+F , TP $_{Major}$: I / G+H+I

f-score : It combines precision and recall into a single score . f-score : (2 * Precision* Recall) / (Precision + Recall)

4- Results

4-1 : BatchPrimer3 and In-silico PCR :

The result of Allele flanking primers , Allele Specific primers and PCR amplification were as following :

Table 13 : Primers for both alleles of each SNP (hetero / homozygosity identification)

	Allele flanking primers	Allele Specific primers	Amplicon Location
N: rs34563000 c.1A>C /G P: 5227021/ B^0	**F:** ACTCCTAAGCCAGTGCCAGA TM : 60.5°C	**R (mutant):** TCAGGAGTCAGATGCACCA**C** TM : 58.3°C	11:5227002-5227186 **Length :** 205 bp
	F: GACAGGTACGGCTGTCATCA TM : 59.7°C	**R (wild):** CTCAGGAGTCAGATGCACCA**T** TM : 59.9°C	
	R: CAGCCTAAGGGTGGGAAAAT TM : 58.3°C	**F (wild):** GCAACCTCAAACAGACACC**A** TM : 59.0°C	11:5226796-5227021 **Length :** 246 bp
N: rs33941849 c.2T>G /C P: 5227020/ B^0	**F:** ACTCCTAAGCCAGTGCCAGA TM : 60.5°C	**R(mutant):** TCAGGAGTCAGATGCACC**G** TM : 56.6°C **R(mutant):** TCAGGAGTCAGATGCACC**C** TM : 57.1°C **R(wild):** CTCAGGAGTCAGATGCACC**A** TM: 59.0°C	11:5227002-5227215 **Length :** 233 bp
	R: AACGGCAGACTTCTCCTCAG TM : 59.7°C	**F(wild):** GCAACCTCAAACAGACACCA**T** TM : 59.0°C	11:5226986-5227020 **Length :** 55 bp
N: rs63750783 c.47G>A P: 5226975/ B^0	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** CACGTTCACCTTGCCC**T** TM : 69.2°C	11:5226948-5227078 **Length :** 150 bp
	F: TCAGGGCAGAGCCATCTATT TM : 58.1°C	**R(wild):** CACGTTCACCTTGCCC**C** TM: 58.7°C	11:5226959-5227076 **Length :** 137 bp
	R: CAGCCTAAGGGTGGGAAAAT TM : 58.3°C	**F(wild):** GCCGTTACTGCCCTGT**G** TM: 58.7°C	11:5226796-5226975 **Length :** 196 bp
N: rs34716011 c.48G>A P: 5226974/ B^0	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** ATCCACGTTCACCTTGCC**T** TM : 68.1°C	11:5226947-5227078 **Length :** 151 bp
		R(wild): CCACGTTCACCTTGCC**C** TM : 59.1°C	11:5226958-5227078 **Length :** 140 bp
	R: CAGATCCCCAAAGGACTCAA TM : 57.0°C	**F(wild):** CGTTACTGCCCTGTG**G** TM : 54.6°C	11:5226745-5226974 **Length :** 245 bp
N: rs33986703 c.52A>T P: 5226970/ B^0	**F:** GTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** CAACTTCATCCACGTTCACCT**A** TM : 68.6°C	11:5226943-5227078 **Length :** 155 bp
	R: CAGATCCCCAAAGGACTCAA TM : 57.0°C	**F(mutant):** TACTGCCCTGTGGGGC**T** TM : 54.9°C	11:5226745-5226971 **Length :** 254 bp
N: rs11549407 c.118C>T P: 5226774/ B^0	**F:** GGCACTGACTCTCTCTGCCT TM : 62.0°C	**R(mutant):** CCAAAGGACTCAAAGAACCTCT**A** TM : 65.5°C	11:5226747-5226824 **Length :** 98 bp
	R: CTTTCTTGCCATGAGCCTTC TM: 57.9°C	**F(mutant):** GGTGGTCTACCCTTGGACC**T** TM : 57.9°C	11:5226690-5226775 **Length :** 113 bp
N: rs33953406	**F:** GGTTGGGATAAGGCTGGATT TM : 57.8°C	**R(mutant):** CCACACCAGCCACCACTT**A** TM : 62.5°C	11:5225627-5225765 **Length :** 158 bp

c.397A>T P: 5225645/ B^0	**F:** GAGTCCAAGCTAGGCCCTTT TM : 59.3°C	**R(wild):** CCACACCAGCCACCACTT**T** TM : 61.3°C	
	R: AGTGATACTTGTGGGCCAGG TM : 59.7°C	**F(wild):** GTGCAGGCTGCCTATCAG**A** TM : 59.9°C	11:5225600-5225645 **Length :** 64 bp
N: rs35256489 c.332T>C P: 5225710/ B^+	**F:** TATGGTTGGGATAAGGCTGG TM : 56.9°C	**R (mutant):** GCCAGCACACAGACC**G** TM : 56.9°C	11:5225695-5225793 **Length:** 118 bp
	F: GAGTCCAAGCTAGGCCCTTT TM : 59.3°C	**R(wild):** GGCCAGCACACAGACC**A** TM : 59.4°C	11:5225694-5225765 **Length:** 91 bp
	R: GATGCTCAAGGCCCTTCATA TM : 57.6°C	**F (wild):** TCCTGGGCAACGTGC**T** TM : 58.2°C	11:5225503-5225710 **Length:** 223 bp

Table 14 : Primers for only one allele of each SNP (hetero / homozygosity identification)

	Allele flanking primers	**Allele Specific primers**	**Amplicon Location**
N: rs33930702 c.3G> C / A / T P: 5227019/ B^0	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** CCTCAGGAGTCAGATGCAC**G** TM : 58.5°C	11:5227000-5227078 **Length :** 99 bp
		R(mutant): CCTCAGGAGTCAGATGCAC**A** TM : 59.0°C	
	R: CAGCCTAAGGGTGGGAAAAT TM : 58.3°C	**F(mutant):** GCAACCTCAAACAGACACCAT**T** TM : 51.1°C	11:5226796-5227020 **Length :** 246 bp
		F(mutant): GCAACCTCAAACAGACACCAT**A** TM : 51.4°C	
N: rs33930165 c.19G>T P: 5227003/ B^0	**R:** AGTGATACTTGTGGGCCAGG TM : 59.7°C	**F(mutant):** CAAAGAATTCACCCCACCAG**T** TM : 58.3°C	11:5225600-5225662 **Length :** 83 bp
N: Rs334 c.20A>T P: 5227002/ B^0	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** CAGTAACGGCAGACTTCTCC**A** TM : 59.9°C	11:5226982-5227078 **Length :** 116 bp
N: rs33959855 c.67G>T P: 5226955/ B^0	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** GGGCCTCACCACCAACTT**A** TM : 62.1°C	11:5226937-5227078 **Length :** 161 bp
	F: TCAGGGCAGAGCCATCTATT TM : 58.1°C	**R(wild):** GGCCTCACCACCAACTT**C** TM : 58.5°C	11:5226938-5227076 **Length :** 158 bp
N: rs35684407 c.91A>G P: 5226931/ B^0	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R(mutant):** GTCTTGTAACCTTGATACCAACC**C** TM : 58.6°C	11:5226908-5227078 **Length :** 190 bp
N: rs33982568 c.108C>A/ G P: 5226784/ B^0	**R:** CAGCATCAGGAGTGGACAGA TM : 59.0°C	**F(wild):** CTTAGGCTGCTGGTGGTCTA**C** TM : 60.6°C	11:5226729-5226784 **Length :** 76 bp

N: rs33991059 c.113G>A P: 5226779/ B^0	**F:** GGCACTGACTCTCTCTGCCT TM : 61.6°C	**R(wild):** ACTCAAAGAACCTCTGGGTC**C** TM : 60.4°C	11:5226759-5226824 **Length :** 85 bp
N: rs33922842 c.130G>T P: 5226762/ B^0	**F:** GGCACTGACTCTCTCTGCCT TM : 61.6°C	**R(mutant):** GGACAGATCCCCAAAGGACT**A** TM : 61.0°C **R(wild):** GACAGATCCCCAAAGGACT**C** TM : 57.9°C	11:5226743-5226824 **Length :** 101 bp
N: rs33969400 c.178A>T P: 5226714/ B^0	**R:** GAAAACATCAAGCGTCCCAT TM : 57.6°C	**F(wild):** TGCTGTTATGGGCAACCCT**A** TM : 58.9°C	11:5226549-5226715 **Length :** 185 bp
N: rs33995148 c.184A>T / G P: 5226708/ B^0	**F:** GGCACTGACTCTCTCTGCCT TM : 61.6°C	**R(wild):** CACTTTCTTGCCATGAGCT**T** TM : 56.4°C	11:5226688-5226824 **Length :** 156 bp
N: rs33933298 c.295G>A P: 5226597/ B^0	**R**: GAAAACATCAAGCGTCCCAT TM : 57.6°C	**F(wild)**: CACTGTGACAAGCTGCAC**G** TM : 60.9°C	11:5226549-5226597 **Length :** 67 bp
N: rs33930977 c.339T>A P: 5225703/ B^0	**R:** GATGCTCAAGGCCCTTCATA TM : 57.6°C	**F(wild)**: GGCAACGTGCTGGTCTG**T** TM : 60.5°C	11:5225503-5225703 **Length :** 218 bp
N: rs33946267 c.364G>T / C P: 5225678/ B^0	**F:** GAGTCCAAGCTAGGCCCTTT TM : 59.3°C	**R(mutant):** GCACTGGTGGGGTGAATT**A** TM : 60.0°C	11:5225660-5225765 **Length :** 125 bp
	F: TATCATGCCTCTTTGCACCA TM : 57.1°C	**R(mutant):** CACTGGTGGGGTGAATT**G** TM : 56.9°C	11:5225661-5225960 **Length :** 319 bp
N: rs34407387 c.419A>C P: 5225623/ B^0	**R:** GACCTCCCACATTCCCTTTT TM 57.3°C	**F(wild)**: GGTGGCTGGTGTGGCTA**A** TM : 60.4°C	11:5225410-5225623 **Length :** 231 bp
N: rs35799536 c.33C>A P: 522698/ B^+	**F:** TCAGGGCAGAGCCATCTATT TM : 58.1°C	**R (wild):** CCCACAGGGCAGTAAC**G** TM : 57.6°C	11:5226973-5227076 **Length :** 123 bp
N: rs33972047 c.59A>G P: 5226963/ B^+	**F:** TCAGGGCAGAGCCATCTATT TM : 58.1°C	**R (mutant):** CACCAACTTCATCCACG**C** TM : 54.7°C	11:5226946-5227076 **Length :** 150 bp
N:rs33951465 c.75T>A P: 5226947/ B+	**F:** TCAGGGCAGAGCCATCTATT TM: 58.1°C	**R (mutant):** GCCCAGGGCCTCACC**T** TM: 61.6°C	11:5226932-5227076 **Length :** 164 bp
N: rs33916412 c.86T>G P: 5226936/ B^+	**F:** AGTCAGGGCAGAGCCATCTA TM : 60.0°C	**R (mutant):** CTTGATACCAACCTGCCC**C** TM : 57.2°C	11:5226918-5227078 **Length:** 180 bp
		R(wild): CTTGATACCAACCTGCCC**A** TM : 57.5°C	11:5226918-5227078 **Length:** 180 bp
N: rs33931779 c.182T>A P: 5226710/ B^+	**F:** TTGGACCCAGAGGTTCTTTG TM : 57.3°C	**R (mutant):** TTTCTTGCCATGAGCCTTC**T** TM : 57.2°C	11:5226691-5226762 **Length:** 91 bp
	F: GGCACTGACTCTCTCTGCCT TM : 61.6°C	**R(wild):** TTCTTGCCATGAGCCTTC**A** TM : 56.9°C	11:5226692-5226824 **Length:** 152 bp

N: rs35485099 c.347C>A P: 5225695/ B^+	**R:** AGTGATACTTGTGGGCCAGG TM : 59.7°C	**F (wild):** GCTGGTCTGTGTGCTGG**C** TM : 62.1°C	11:5225600-5225695 **Length:** 113 bp
N:rs33925391 c.380T>G P: 5225662 /B^+	**R:** GATGCTCAAGGCCCTTCATA TM : 57.6°C	**F (wild):** CAAAGAATTCACCCCACCAG**T** TM : 58.3°C	11:5225503-5225662 **Length:** 180 bp

4-2 : SnapGene

Figure 37 : Electrophoresis simulation for 18 primer pairs via SnapGene

4-3 : SOPMA

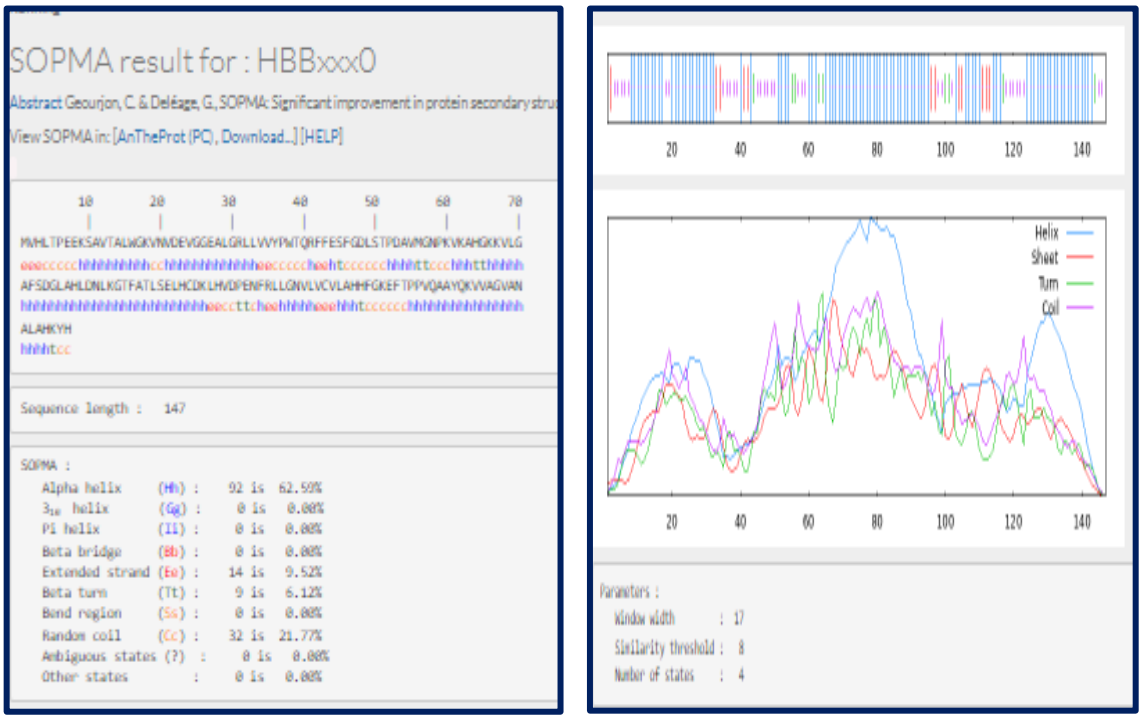


Figure 38 : Secondary structure prediction of HBB gene via SOPMA

4-4: SwissDOCK

The results page provides a list of the docked complexes in a tabular manner on the right hand side of the page and the 3D structure of the docked complex on the left hand side of the page . In the table, the first column namely ‘Show’ contains the radio buttons, by clicking on any one of those radio buttons, it’ll display the conformation for the complex against which it’ll present. Then in the next two columns, it provides the number of conformations (in the ‘Elements’ column) against each cluster number (in the ‘Cluster’ column). Each cluster represents various conformations of a ligand at the given location on the protein and each cluster represents a pocket on the target protein. The different clusters could be viewed in the SwissDock web browser through a Jmol applet . Then in the next column it provides the values for the ‘FullFitness’ and then in the next column, namely ‘Estimated ΔG’ it provides the estimated binding free energy of the complex.

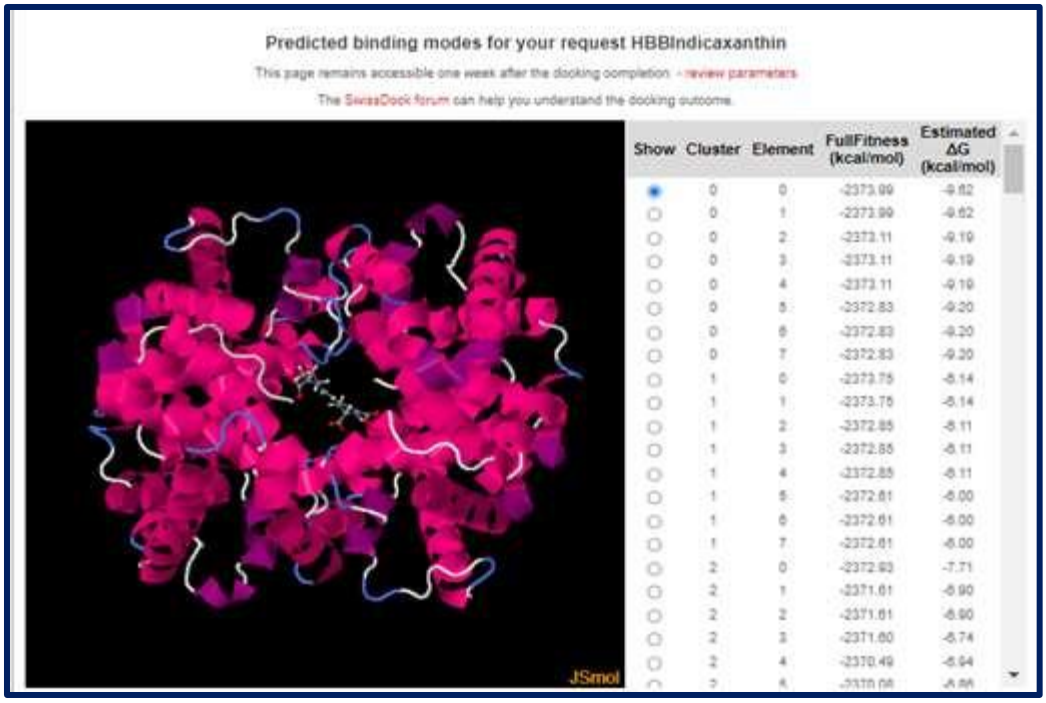


Figure 39 : Prediction binding modes for HBB with indicaxanthin via SwissDock

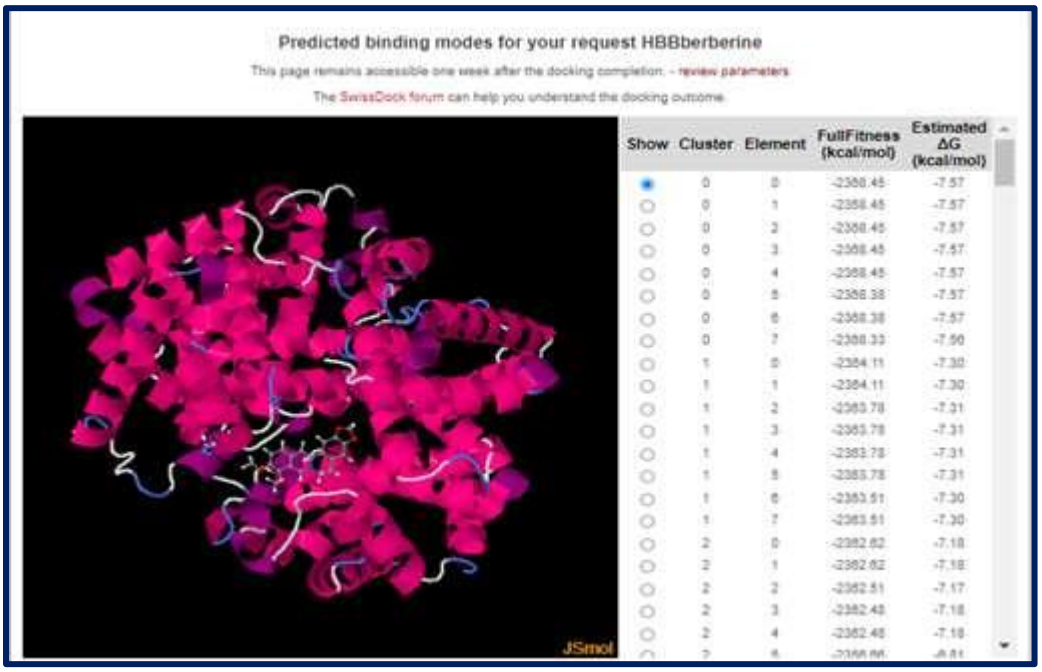


Figure 40 : Prediction binding modes for HBB with berberine via SwissDock

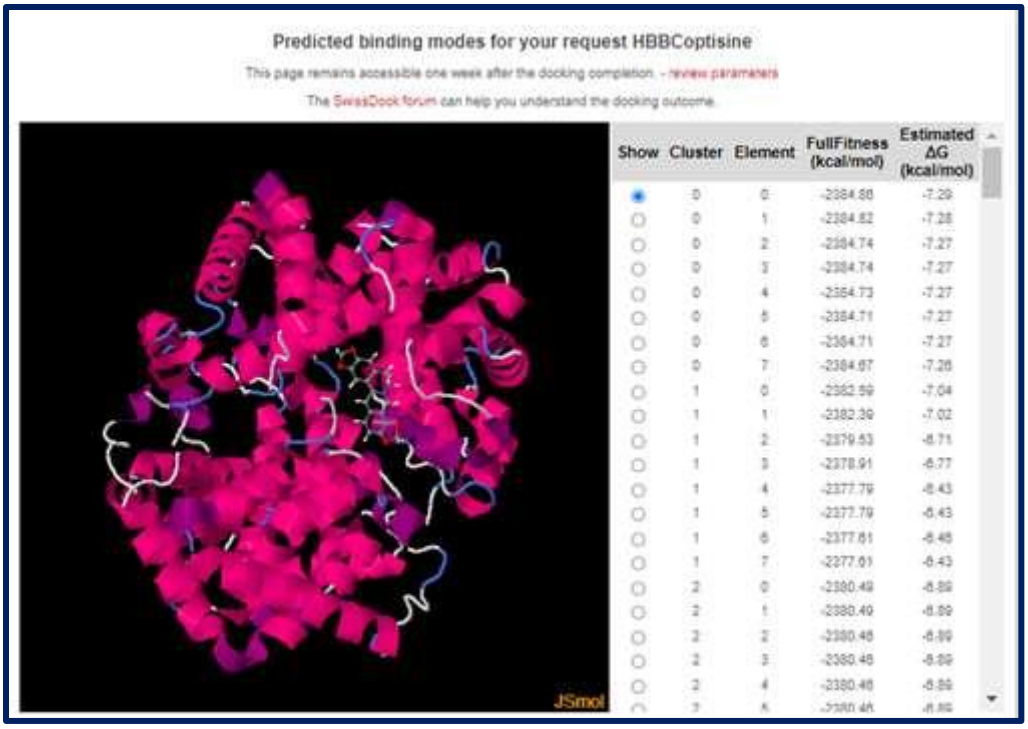


Figure 41: Prediction binding modes for HBB with coptisine via SwissDock

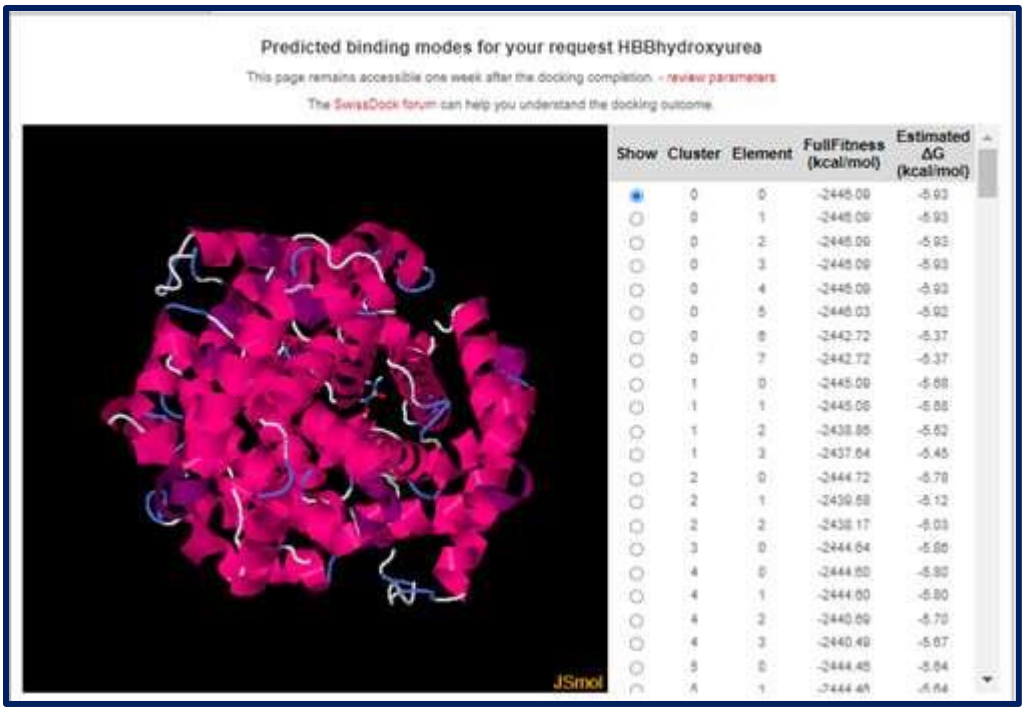


Figure 42 : Prediction binding modes for HBB with hydroxyurea via SwissDock

The following results are a comparison between current study of the SwissDOCK tool with a previous study of the AutoDOCK4 program .

Tabel 15 : Binding energy of available drugs with hemoglobin beta chain.

Drugs	*Binding energy from SwissDOCK*	*Binding energy from AutoDOCK4*
Indicaxanthin	-9.62 Kcal/mol	-5.13 Kcal/mol
Hydroxyurea	-5.93 Kcal/mol	-2.82 Kcal/mol

Table 16 : Binding energy of bioactives from coptidisrhizome with hemoglobin beta chain.

Bioactives	*Binding energy from SwissDOCK*	*Binding energy from AutoDOCK4*
Berberine	-7.57 Kcal/mol	- 6.0 Kcal/mol
Coptisine	-7.29 Kcal/mol	-5.49 Kcal/mol

4-3 : Fuzzy inference system (fis)

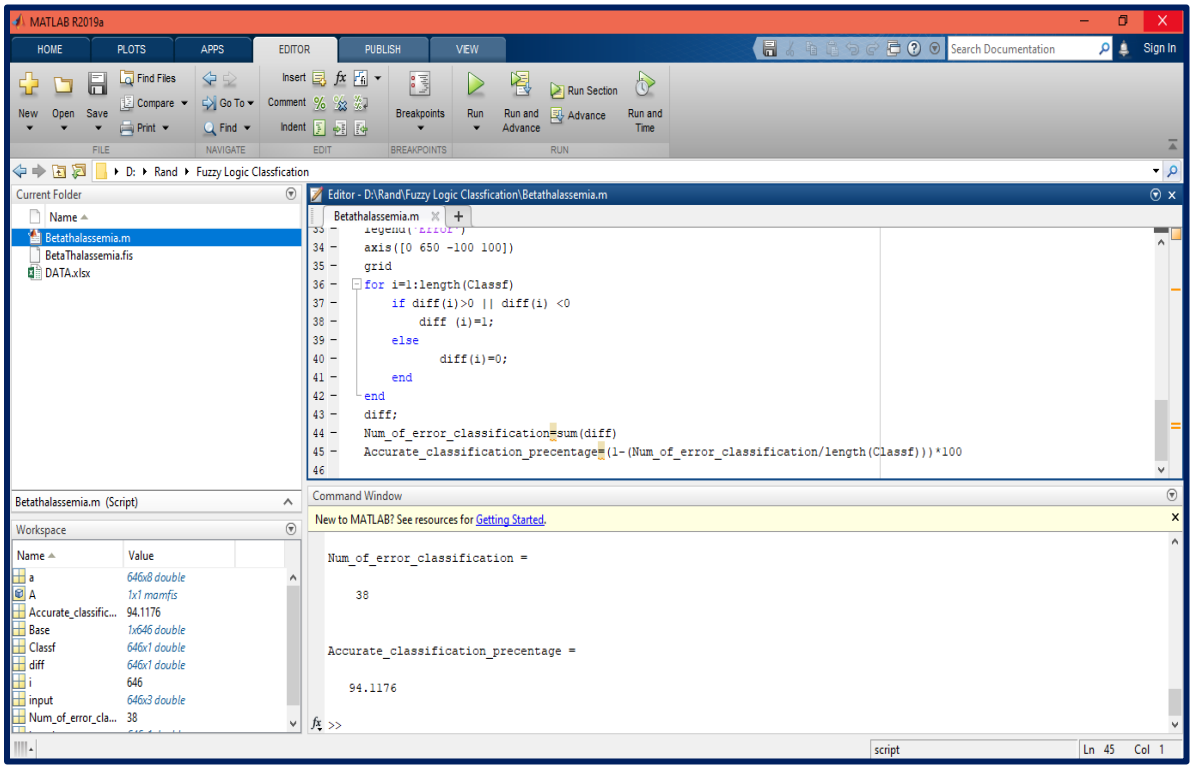


Figure 43 : Accuracy result for Fuzzy inference system

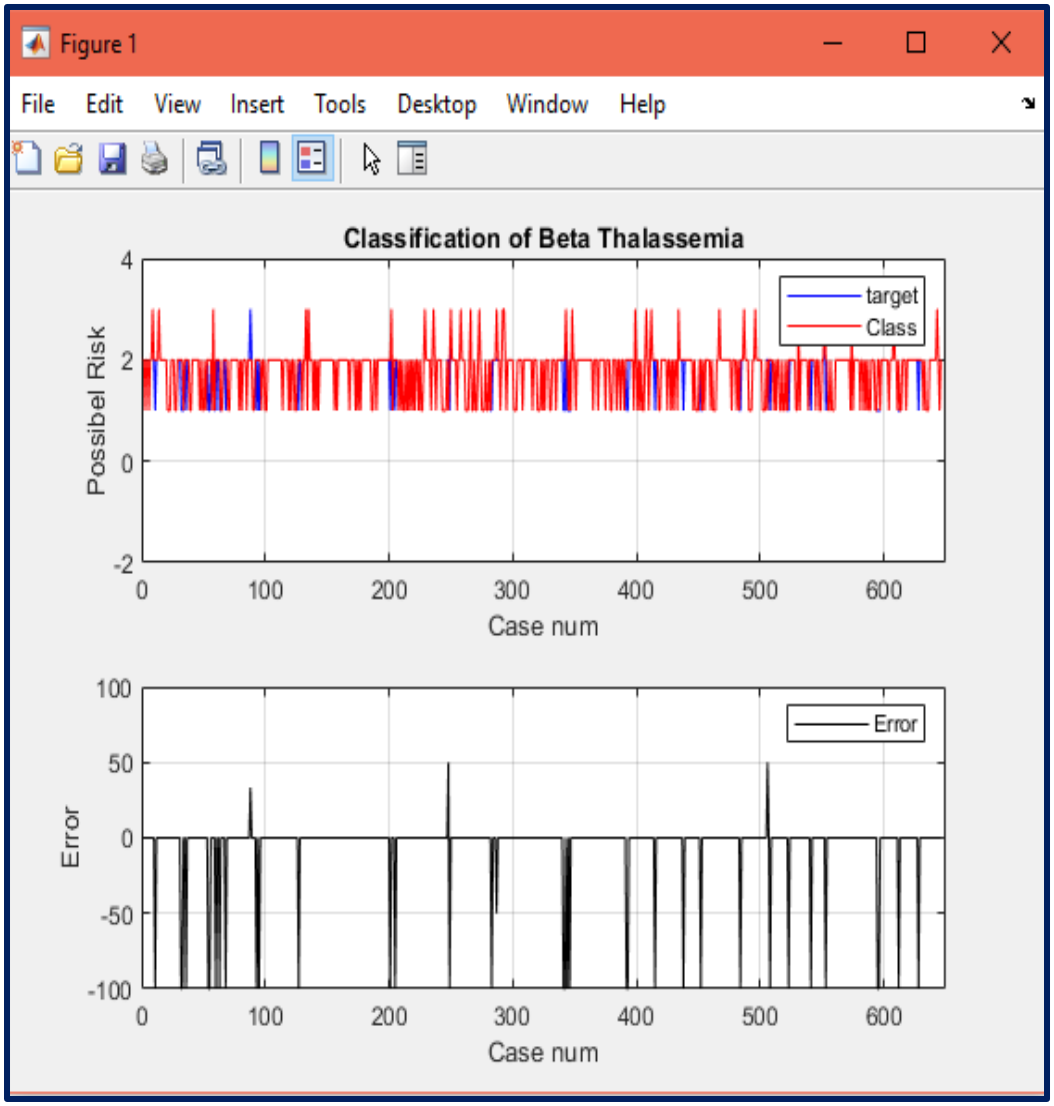


Figure 44 : Possible risk and error rate for Fuzzy inference system

Table 17 : Confusion matrix results

Confusion Matrix		Predicted			
		Minor	Intermedia	Major	Total
Expected	Minor	175	34	0	209
	Intermedia	2	405	1	408
	Major	0	1	28	29

Table 18 : Performance evaluation of the results of fuzzy inference system

	Accuracy	Error rate	Precision	Recall / Sensitivity	f-score
Th-Minor	0.941	0.059	0.98	0.83	0.89
Th-Intermedia			0.92	0.99	0.95
Th-Major			0.96	0.96	0.96

5- Discussion

Beta-thalassemia is one of most common autosomal recessive disorders worldwide . Thalassemia and hemoglobinopathy genotype interpretation and phenotype determination as well as their clinical symptoms and electrophoresis are the one of the most complex subjects of Hematology. The application of a combined molecular approach with clinical data and efficient bioinformatics tools will enable a guideline for functional studies and prenatal diagnosis to be developed as basis for future studies. Bioinformatics has become an essential tool not only for basic research but also for applied research in biotechnology and biomedical sciences. Diagnostic methods based on single nucleotide polymorphism (SNP) biomarkers are essential for the real adoption of personalized medicine. The Allele-specific PCR method was developed for allele analysis of clinically significant mutations to facilitate reliable discrimination between two alleles highly .Optimal primer sequence and appropriate primer concentration are essential for maximal specificity and efficiency of PCR. The key to the PCR lies in the design of the two oligonucleotide primers. It is essential that care is taken in the design of primers for PCR. Several parameters including the length of the primer, %GC content and the 3' sequence need to be optimized for successful PCR. A number or software packages such as BatchPrimer3 has allowed the process of primer design to be less troublesome , Silica confirm primer quality and SnapGene separate DNA fragments according to their size by simulate gel agarose and compare with Silica results .

In this study , Allele-specific primers corresponding to 37 kinds of SNP in β thalassemia were designed according to different combinations between mismatch base and mismatch site . Optimization of melting temperature, primer length and amplified products length were achieved using primer program BatchPrimer3 v1.0 (https://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi). Primers were designed according to the genome sequences near these SNP sites. The program output was a list of the primer sets (wild-type FP/ RP, mutant-FP / RP, common FP /RP) selected for each targeted region. Single assays were performed for the specific amplification of each target allele using the tested primers , and all amplicons were between 100 - 300 bp fragments in which containing the corresponding SNPs by Silica tool (https://www.gear-genomics.com/silica/). PCR products were detected on 2.5% agarose gel by insilico electrophoresis via SnapGene tool (https://www.snapgene.com/free-trial)

The result in (table 13) showed set of primers selected that have a unique sequence within the template DNA, optimal melting temperatures , appropriate primer length, suitable GC content and Specificity for allele pairs for each SNPs . These primers result refers to 8 SNPs in studied dataset which are rs34563000 , rs33941849 , rs63750783 , rs34716011 , rs33986703 , rs11549407 , rs33953406 , rs35256489 . The result in (table 14) showed set of primers selected that have a unique sequence within the template DNA, optimal melting temperatures , appropriate primer length, suitable GC content and Specificity for only one allele for each SNPs . These primers result refers to 21 SNPs in studied dataset which are rs33930702 , rs33930165 , Rs334 , rs33959855 , rs35684407 , rs33982568 , rs33991059 , rs33922842 , rs33969400 , rs33995148 , rs33933298 , rs33930977 , rs33946267 , rs34407387, rs35799536 , rs33972047 , rs33951465 , rs33916412 , rs33931779 , rs35485099 , rs33925391 . Other SNPs in studied dataset like rs33950507 , rs33960103 , rs33974936 , rs33913712 , rs33910569, rs33950507 , rs35424040 , rs35578002 were ruled out due to getting a non specificity amplification or the melting temperature of primers were smaller than 50 c^0 or the melting temperature differences between

primers were bigger than 7 c^0 . All these reasons form hairpin or primer dimer and lead to misleading results.

For the simulated results of electrophoresis on agarose gel , All 18 bands that showed in (figure 37) were identical with Silica results for rs34563000 , rs33941849 , rs63750783 , rs34716011 , rs33953406 (x2) , rs35256489 (x2) , rs33982568 , rs33991059 , rs33922842 , rs33969400 , rs33995148 , rs33933298 , rs33930977 , rs33931779 , rs35485099 , rs33925391 respectively . These results helped us to Verify the validity of the amplification results under particular gel conditions.

For Secondary structures of HBB protein predicted by SOPMA tool , (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html) , The sequence length was 147 amino acids ,a considerable prediction was observed in mainly alpha helix classes of protein by 62.59% and less prediction for beta sheet (extended strand) by 9.52% , in addition random coils and beta turn were found by 21.77% , 6.12 % respectively (figure 38). These accurate predictions corroborate that this method are efficient because It will be anticipated that the combination of protein secondary structure prediction methods with additional protein structure features has found to provide more accurate results. Therefore, it could be concluded that synchronized employment of secondary structure prediction methods could enhance the accuracy of insilico analysis. Based on the observed results, we can compare ssecondary structure data of mutant proteins obtained through experimental methods with that of predictions made by SOPMA to determine the best treatment of thalassemia in the future compared to the already available drugs available .

In SwissDOCK tool results (http://www.swissdock.ch/) , A better docking score corresponds to Low Binding Energy. The binding energy parameter is used to determine which ligand has a stable complex interaction with protein. The more negative value or lower binding affinity, the more stable ligand-receptor is achieved , so in our study , Among the drugs , the hit molecule of beta hemoglobin was Indicaxanthin with binding energy of -9.62 Kcal/mol (table 15). Among the bioactives derived from Coptidis Rhizome,

the hit molecule of beta hemoglobin was Berberine with binding energy of -7.57 Kcal/mol (table 16) . These results corresponded with many works have been previously done for the treatment of thalassemia with the help of bioactive compounds from Coptidis Rhizome by AutoDOCK4 program (Chowdhury, A. and Sruthi, V.S. (2021) that revealed berberine is a safe bioactive compound without any toxicity and electrolyte imbalance when administered with the herbal concoction of Coptidis Rhizome for 1252 days. Based on traditional dosage and indication, this herb is safe for oral concoction .

For fuzzy inference system results, It was found that our program matched the doctor's diagnosis in 608 cases perfectly from 646 cases . The other 38 were marginally off. This results with an accuracy of about 94.11 % (figure 43-44) which is more stable than the results obtained from the similar contents in (Thakur et al., 2016) with accuracy 83% because we defined output value between 0 and 3.0 to make a cluster of "Thal_Minor" (Thalassemia Minor) of Thalassemia disease instead 0 -3.5 which was more matching between predicted and actual results for Thalassemia Minor and Thalassemia intermedia .

Based on the results presented in the confusion matrix , it was discovered that there were 608 correct classifications (175 for low, 405 for moderate and 28 for high risk-along the diagonal) as shown in (table 17).Also, the results showed that the TP rate which gave a description of the proportion of actual cases that was correctly predicted was TP_{minor} : 0.83, $TP_{Intermedia}$: 0.99, TP_{Major}: 0.96 , while the precision which gave a description of the proportion of predictions that were correctly classified was $Precision_{minor}$: 0.98 , $Precision_{Intermedia}$: 0.92 , $Precision_{Major}$: 0.96 ,while the f-score which combines precision and recall into a single score was $f\text{-}score_{minor}$: 0.89, $f\text{-}score_{Intermedia}$: 0.95 , $f\text{-}score_{Major}$: 0.96 . From the viewpoint of an end-user, the results of this work can facilitate laboratory work by reducing the time and cost.

6- Conclusion and Recommendation :

- Bioinformatics has become an essential tool not only for basic research but also for applied research in biotechnology and biomedical sciences. In particular, low-cost genotyping tools are absolutely necessary for effective personalized medicine. Single nucleotide polymorphisms (SNPs) have been proposed as the next generation of markers to identify loci associated with complex diseases and their therapeutic treatment .
- Constructing a relationship between the genotype and phenotype experimentally is an important aspect of research , but it can prove to be highly difficult ,so the in silico analysis provides a solution here which it helps researchers analyze enormous amounts of data in biology to narrow down the positive leads that can be further analyzed experimentally for validation. This saves an extensive amount of labor, time, and costs like AS-PCR methods which are quick, excellent and inexpensive strategies and require minimal instruments that are found in most laboratories to be developed for massive implementation into clinical laboratories .

As a future perspective, augmentations could be brought about in the protein secondary structure prediction methods and molecular docking which computationally predicts the complex from individual structures to further enhance the prediction accuracy where the discovery of a natural inducer in an existing licensed medicine based on the findings a significant step forward in the development of drugs to treat beta thalassemia.

we hope that finally, in addition to providing detailed information to help promote enhanced β-thalassemia pre-natal screening, achieving these objectives will also help to identify the mechanisms responsible for fetal hemoglobin control, since reactivation of fetal hemoglobin can provide major therapeutic benefits to people affected by β-hemoglobinopathies .

By performing fuzzy inference system we hope the result can help the medical sector to classify and detect whether the patient has thalassemia or not, so the patient can receive the right treatment to increase their life expectancy and reduce the risk of thalassemia to the next generation.

7- References

1- Galanello, R., & Origa, R. (2010). **Beta-thalassemia**. Orphanet journal of rare diseases, 5, 1-15. https://doi.org/10.1186/1750-1172-5-11

2- Cao, A., & Galanello, R. (2010). **Beta-thalassemia. Genetics in medicine**, 12(2), 61-76. https://doi.org/10.1097/GIM.0b013e3181cd68ed

3- Rund, D., & Rachmilewitz, E. (2005). **β-Thalassemia.** New England Journal of Medicine, 353(11), 1135-1146. DOI: 10.1056/NEJMra050436

4- Danjou, F., Anni, F., & Galanello, R. (2011). **Beta-thalassemia: from genotype to phenotype**. haematologica, 96(11), 1573. doi: 10.3324/haematol.2011.055962

5- Origa, R. (2017). **β-Thalassemia. Genetics in Medicine**, 19(6), 609-619. https://doi.org/10.1038/gim.2016.173

6- Needs, T., Gonzalez-Mosquera, L. F., & Lynch, D. T. (2018). **Beta thalassemia**. PMID: 30285376

7- Badens, C., Joly, P., Agouti, I., Thuret, I., Gonnet, K., Fattoum, S., ... & Pissard, S. (2011). **Variants in genetic modifiers of β-thalassemia can help to predict the major or intermedia type of the disease. haematologica**, 96(11), 1712-1714. doi: 10.3324/haematol.2011.046748

8- Thein, S. L. (2013). **The molecular basis of β-thalassemia**. Cold Spring Harbor perspectives in medicine, 3(5), a011700. doi: 10.1101/cshperspect.a011700

9-labpedia.net/anemia-part-4-thalassemia-alpha-thalassemia-beta-thalassemia-discussion-and-workup/

10- Lee, J. S., Rhee, T. M., Jeon, K., Cho, Y., Lee, S. W., Han, K. D., ... & Lee, Y. K. (2022). **Epidemiologic Trends of Thalassemia, 2006–2018: A Nationwide Population-Based Study**. Journal of Clinical Medicine, 11(9), 2289. https://doi.org/10.3390/jcm11092289

11- Upendraa, R. S., & MananChamariab, L**. Single nucleotide polymorphisms (SNPs) in causing β-thalassemia**. http://www.ijpsr.info/docs/IJPSR19-10-01-001.pdf

12- Thalassaemia.org.cy/haemoglobin-disorders/thalassaemia/

13- Tripathi, P. (2022). **Genetics of Thalassemia**. DOI: 10.5772/intechopen.106748

14- Pooja Advani .(2022) **. Beta Thalassemia Treatment & Management** . Medscape

15- Razmjooee, S. (2017). **Scalar Scoring in Thalassemia Genotype: As a New Overview**. International Journal of Applied, 4(2), 9. DOI: 10.21767/2394-9988.100060

16- Taher, A. T., Musallam, K. M., & Cappellini, M. D. (2021). **β-Thalassemias**. New England Journal of Medicine, 384(8), 727-743. DOI: 10.1056/NEJMra2021838

17- El-Kamah, G. Y., & Amr, K. S. (2015). **Thalassemia—from genotype to phenotype.** Inherited hemoglobin disorders, 13. DOI: 10.5772/61433

18- Carlice-dos-Reis, T., Viana, J., Moreira, F. C., Cardoso, G. D. L., Guerreiro, J., Santos, S., & Ribeiro-dos-Santos, A. (2017). **Investigation of mutations in the HBB gene using the 1,000 genomes database**. PLoS One, 12(4), e0174637. https://doi.org/10.1371/journal.pone.0174637

19- Liu, J., Huang, S., Sun, M. *et al***. An improved allele-specific PCR primer design method for SNP marker analysis and its application**. *Plant Methods* **8**, 34 (2012). https://doi.org/10.1186/1746-4811-8-34

20- Tortajada-Genaro, L. A., Puchades, R., & Maquieira, Á. (2017). **Primer design for SNP genotyping based on allele-specific amplification—Application to organ transplantation pharmacogenomics**. Journal of Pharmaceutical and Biomedical Analysis, 136, 14-21. http://dx.doi.org/10.1016/j.jpba.2016.12.030.

21- Delghandi, M., Delghandi, M. P., & Goddard, S. (2022**). The significance of PCR primer design in genetic diversity studies: exemplified by recent research into the genetic structure of marine species**. PCR Primer Design, 3-15. DOI: 10.1007/978-1-0716-1799-1_1

22- Smith, K., Vatman, D., & Ponce, J. (2002). **Genetic polymorphism and SNPs. Genotyping**, haplotype assembly problem, haplotype map, functional genomic and proteomics.

23- Chowdhury, A. and Sruthi, V.S. (2021). **A Bioinformatics Approach for the Treatment of Thalassemia using Molecular Docking**. Biological Forum – An International Journal, 13(3): 332-338.

24- Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011**). Molecular docking: a powerful approach for structure-based drug discovery.** Current computer-aided drug design, 7(2), 146-157. doi: 10.2174/157340911795677602

25- Pagadala, N.S., Syed, K. & Tuszynski, J. **Software for molecular docking: a review.** Biophys Rev 9, 91–102 (2017). https://doi.org/10.1007/s12551-016-0247-1

26- Chatterjee, T., Gupta, S. K., & Patel, A. **Molecular Docking and Molecular Dynamic Simulation of Andrographoilde and HDAC2 Inhibitor an Approach to Manage for Beta Thalassemia.**

27- Thakur, S. A. P. N. A., Raw, S. N., & Sharma, R. (2016). **Design of a fuzzy model for thalassemia disease diagnosis: Using mamdani type fuzzy inference system (FIS)**. International Journal of Pharmacy and Pharmaceutical Sciences, 8(4), 356-361.

28- Thakur, S., Raw, S. N., Sharma, R., & Mishra, P. (2016). **Detection of type of thalassemia disease in patients: A fuzzy logic approach**. International Journal of Applied Pharmaceutical Sciences and Research, 1(02), 88-95. DOI:10.21477/ijapsr.v1i2.10944

29- Carlice-dos-Reis, T., Viana, J., Moreira, F. C., Cardoso, G. D. L., Guerreiro, J., Santos, S., & Ribeiro-dos-Santos, A. (2017). **Investigation of mutations in the HBB gene using the 1,000 genomes database.** PLoS One, 12(4), e0174637.. https://doi.org/10.1371/journal.pone.0174637

30- You, F.M., Huo, N., Gu, Y.Q. *et al.* **BatchPrimer3: A high throughput web application for PCR and sequencing primer design**. *BMC Bioinformatics* 9, 253 (2008). https://doi.org/10.1186/1471-2105-9-253

31- Sabri, N., Aljunid, S. A., Salim, M. S., Badlishah, R. B., Kamaruddin, R., & Malek, M. A. (2013). **Fuzzy inference system: Short review and design**. Int. Rev. Autom. Control, 6(4), 441-449. https://www.researchgate.net/publication/280739444

32 - Singla, J., Grover, D., & Bhandari, A. (2014). **Medical expert systems for diagnosis of various diseases**. International Journal of Computer Applications, 93(7).

33- Wangkumhang, P., Chaichoompu, K., Ngamphiw, C. *et al.* **WASP: a Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations**. *BMC Genomics* **8**, 275 (2007). https://doi.org/10.1186/1471-2164-8-275

34 - Qadah, T., & Jamal, M. S. (2019). **Computational analysis of protein structure changes as a result of nondeletion insertion mutations in human β-globin gene suggests possible cause of β-thalassemia**. BioMed Research International, 2019. doi: 10.1155/2019/9210841

35- Abd-Elsalam, K. A. (2003). **Bioinformatic tools and guideline for PCR primer design.** african Journal of biotechnology, 2(5), 91-95.

36 - Geourjon, C., & Deleage, G. (1995). SOPMA: **significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments**. Bioinformatics, 11(6), 681-684.

37- Goller, C. C., Srougi, M. C., Chen, S. H., Schenkman, L. R., & Kelly, R. M. (2021, July). **Integrating bioinformatics tools into inquiry-based molecular biology laboratory education modules**. In Frontiers in education (Vol. 6, p. 711403). Frontiers Media SA. https://doi.org/10.3389/feduc.2021.711403

38- Angamuthu, K., & Piramanayagam, S. (2017). **Evaluation of in silico protein secondary structure prediction methods by employing statistical techniques**. Biomedical and Biotechnology Research Journal (BBRJ), 1(1), 29. DOI: 10.4103/bbrj.bbrj_28_17

39- Yan, Y., Zhang, D., Zhou, P., Li, B., & Huang, S. Y. (2017). **HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy.** Nucleic acids research, 45(W1), W365-W373. doi: 10.1093/nar/gkx407

40- Grosdidier, A., Zoete, V., & Michielin, O. (2011). **SwissDock, a protein-small molecule docking web service based on EADock DSS**. Nucleic acids research, 39(suppl_2), W270-W277. doi: 10.1093/nar/gkr366

41- Rehman, I., Farooq, M., & Botelho, S. (2021). **Biochemistry, secondary protein structure**. In StatPearls [Internet]. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK470235/

42- Aszhari, F. R., Rustam, Z., Subroto, F., & Semendawai, A. S. (2020, March). **Classification of thalassemia data using random forest algorithm**. In Journal of Physics: Conference Series (Vol. 1490, No. 1, p. 012050). IOP Publishing. DOI :10.1088/1742-6596/1490/1/012050

43- Ngozi Chidozie Egejuru, Sekoni Olayinka Olusanya, Adanze Onyenonachi Asinobi, Omotayo Joseph Adeyemi, Victor Oluwatimilehin Adebayo, Peter Adebayo Idowu. **Using Data Mining Algorithms for Thalassemia Risk Prediction**. International Journal of Biomedical Science and Engineering. Vol. 7, No. 2, 2019, pp. 33-44. doi: 10.11648/j.ijbse.20190702.12

44- Mahdieh, N., & Rabbani, B. (2016**). Beta thalassemia in 31,734 cases with HBB gene mutations: pathogenic and structural analysis of the common mutations; Iran as the crossroads of the Middle East**. Blood reviews, 30(6), 493-508. , http://dx.doi.org/10.1016/j.blre.2016.07.001

45- Yadav, A. K. (2010). **Comparative analysis of protein structure of common Hb Q variants**. Indian Journal of Pathology and Microbiology, 53(4), 696.

46- Munkongdee, T., Tongsima, S., Ngamphiw, C., Wangkumhang, P., Peerapittayamongkol, C., Hashim, H. B., ... & Svasti, S. (2021). **Predictive SNPs for β0-thalassemia/HbE disease severity**. Scientific Reports, 11(1), 10352. https://doi.org/10.1038/s41598-021-89641-2

47- Zhang, J., Li, P., Yang, Y., Yan, Y., Zeng, X., Li, D., ... & Zhu, B. (2019). **Molecular epidemiology, pathogenicity, and structural analysis of haemoglobin variants in the Yunnan province population of Southwestern China**. Scientific Reports, 9(1), 8264. https://doi.org/10.1038/s41598-019-44793-0