International Journal of Engineering & Technology, 8 (1.11) (2019) 1-8

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Diagnose Mutations Causes B-Thalassemia: Biomining Method Using an Optimal Neural Learning Algorithm

Ayad Ghany Ismaeel[1*]

[1]Technical Engineering College, Al-kitab University, Kirkuk-IRAQ
**E-mail: dr.ayad.ghany.ismaeel@gmail.com*

Abstract

The problems in genome and proteome classification of mutations causing a thalassemia are synthesis, e.g. which thalassemia's database will choose? and then the technique that used in biomining to classify mutations causing thalassemia who can say is effective/optimal. This paper proposed genomics classification for β-thalassemia's mutations in ITHALNET-IthaGenes database [1] (which is a modern and more comprehensive comparing to other thalassemia databases about 63% of thalassemia's mutations) using data biomining method based on multiple neural network learning algorithms (Conjugate Gradient Descent, quick propagation, online backpropagation BP and batch BP algorithm). The experimental results based on architecture of BP [457-228-1] with (1000) iteration shows conjugate gradient descent is optimal biomining technique comparing to other techniques of diagnosis mutation of B-thalassemia, which shows in training stage with error improvement= 5.20E-08 and testing stage Correlate= 0.999601 & R-Squared= 0.9992, in quick propagation gives error improvement= 5.20E-08, Correlate= 0.997086 & R-Squared= 0.994173, in Batch BP reveals error improvement = 0.257249, Correlate= 0.975762 & R-Squared= 0.931719, finally the online propagation error improvement= 0.000013 and testing stage Correlate= 0.975277 & R-Squared= 0.900057).

Keywords: *Biomining; Backpropagation; Batch Backpropagation; β-thalassemia; Conjugate Gradient Descent; Genome; ITHALNET-IthaGenes database; Mutation; Neural Learning Algorithm; Proteome; Quick Propagation.*

1. Introduction

Thalassemia was caused by a genetic mutation in the DNA of the cells forming the hemoglobin, and this mutation is passed genetically from parents to children. Genetic mutations disrupt the production of normal hemoglobin, so low levels of hemoglobin and high red blood cell damage (which occurs in thalassemia patients) lead to anemia., Thalassemia involves the absence of or errors in genes (β, δ, α, γ, etc.) responsible for production of hemoglobin, thalassemia is spread in the Mediterranean region, the diagnosis can be done using laboratory test called Electrophoresis.

Thalassemia disease depends on the mutations in (β, δ, α1, α2, Aγ, Gγ, and there is more.) genes, e.g. genes on chromosome 16 are responsible for alpha subunits, while genes on chromosome (11) control the production of beta subunits. A lack of a particular subunit determines the type of thalassemia (a loss of alpha subunits results in alpha-thalassemia). The loss of subunits thus corresponds to errors in the genes on the related chromosomes.

The defect is either alpha (α-thalassemia) or beta (β-thalassemia) and is the most common inherited single-gene disorders in the world with the highest prevalence in areas where malaria was or still is endemic, e.g. in Iran an estimated that about 8,000 pregnancies are at risk each year. In the Mediterranean region some long-established control programs have achieved 80-100% prevention of newly affected births.

The new technique of diagnosis thalassemia based on genetic test , the risk of this disease without diagnosis in advance or early will a raise this risk. The idea of thalassemia trait is inherited, and the risk which comes from parents in β-thalassemia must be define the following terminologies, a healthy gene with (β) and an abnormal gene with (βo) will be denoted as a minor β-thalassemia (ββo), major β-thalassemia (βoβo) and the proper person (ββ). There are multiple cases of β-thalassemia the details related of each case explained as follow:

1. Marriage between a healthy/ proper person (ßß) and another person with a minor β-thalassemia (ββo) has the carrier thalassemia (i.e. the probability of the fetus being infected with a minor beta thalassemia is 50%) as shown in Table 1:

Table 1: Reveals Minor Beta Thalassemia Is 50%

	ß	ß^{o}
ß	ßß	ßß^{o}
ß	ßß	ßß^{o}

2. Marriage between a person who is pregnant with the disease (ββo) of the another person with the same case (ββo) pregnant with the disease are the possibilities Infection in the fetus with minor β-thalassemia is 50% (pregnant), 25% by major β-thalassemia and 25% by normal/proper, as illustrate in Table 2:

Table 2: Shows Minor β-Thalassemia Is 50% (Pregnant), 25% By Major B- Thalassemia and 25% By Normal/Proper

	β	$β^{o}$
β	ββ	$ββ^{o}$
$β^{o}$	$ββ^{o}$	$β^{o}β^{o}$

3. Marriage between an infected of major β-thalassemia (βoβo) from a healthy/proper person (ββ) all children are carriers of the disease minor β-thalassemia (ββo) as appears in Table 3 :

Table 3: Appears All Children are Carriers of The Disease Minor β-Thalassemia (ββo)

	$β^{o}$	$β^{o}$
β	$ββ^{o}$	$ββ^{o}$
β	$ββ^{o}$	$ββ^{o}$

4. Marriage between an infected major β-thalassemia (βoβo) from a person with a disease (ββo), the possibilities Infection of the fetus in pregnancy with major beta thalassemia is 50% and 50% by minor thalassemia, as shown in Table 4:

Table 4: Shows Major Beta Thalassemia Is 50% and 50% by Minor Thalassemia

	β°	β°
β	ββ°	ββ°
β°	β°β°	β° β°

5. If an infected person major β-thalassemia (βoβo) marries another infected person major β-thalassemia (βoβo), all children have a major β-thalassemia (βoβo).

This paper suggested new biomining classifying/diagnosis technique of mutations cause β-thalassemia using data mining based on optimal neural network technique, which is first important step for gene therapy, i.e. modification of gene therapy, is the focus of new research direction on thalassemia.

2. Related Works

DOMINGOS, Ana L. B. and et al. [2010] Variations in the phenotypic expression of heterozygous beta thalassemia reflect the formation of different populations. To better understand the profile of heterozygous beta thalassemia of the Brazilian population, the paper aimed at establishing parameters to direct the diagnosis of carriers and calculate the frequency from information stored in an electronic database. Using a data mining tool, evaluated information on 10,960 blood samples deposited in a relational database, over the years, improved diagnostic technology has facilitated the elucidation of suspected beta thalassemia heterozygote cases with an average frequency of 3.5% of referred cases. Also found the Brazilian beta thalassemia trait has classic increases of Hb A2 and Hb F(60%), mainly caused by mutations in beta zero thalassemia, especially in the southeast of the country[8].

Altug Akay, Andrei Dragomir and el at. [April 2009] proposed study analyzed β-thalassemia's socioeconomic geography and how it affects the afflicted population, processed survey data and performed data mining using self-organizing maps to identify underlying data structure. Hypothesized in this study that certain variables mark subgroups within the affected population and aimed at identifying these subgroups and used a correlation based measure to assess the variable's importance to the subgroup's distinction, The population's education level was one of the major factors that divided it into different subgroups. This study appears that recurring patterns of specific variables separated the affected population into disparate subgroups based on their response to questionnaires [6]..

Ou XB, Zhang L and el at and el at. [Jan. 2005] suggested study had aim was to explore the application value of the diagnostic genechips in determining thalassemia, this method focus on subjects group 62 children had (α-thalassemia) and 93 children with (β-thalassemia) 60 with thalassemia trait, 33 with thalassemia major) from Guangdong province were tested from July 2002 to July 2003; 115 were males and 40 were females, the age ranged from 1 day to 11 years. DNA was extracted from ACD coagulated blood with Invisorb DNA extraction kit. After preparation, the alpha and beta globin gene organization and structure of sample was analyzed by genechips technology. using genechip in identifying thalassemia mutations has the advantages of simplicity, economy and shorter time. This technique does not use radioisotope and could also detect alpha and beta thalassemia mutations simultaneously. (2) The occurrence of alpha and beta thalassemia dual heterozygotes is frequent in Guangdong province and the genechip technology is important in genetic counseling and prenatal diagnosis of thalassemia in this area [7].

In general there isn't diagnosis of thalassemia based on mutations in genes (via genome test which is more effective test than traditional or using sequence analysis) caused thalassemia, all previous techniques either not based on genome databases related nor working on DNA sequence analysis, i.e. focus on mutations of thalassemia's genes. Addition to there isn't dependence on using more than one technique to find the optimal technique can use in data mining. The motivation overcomes all drawbacks of previous techniques by suggested a genomic classification/diagnosis for β thalassemia via its mutations based on common and modern genome database like ITHALNET using more than one backpropagation techniques to reach the optimal of them.

3. Proposed Biomining Method for Diagnosis Mutation of B-Thalassemia

This proposed approach contains two important stages as follow:

1. Select optimal NN algorithm for biomining of mutations caused disease of β-thalassemia:

Backpropagation BP algorithm is the most popular algorithm for training of multilayer perceptron and is often used by researchers. First must be selected the genome database of β-thalassemia as known there are multiple database related like HbVar generated in 90s of previous century [9], iTHANET: IthaGenes which is created in period (2006-2008) and updated till now, it is under supervision/sponsor by European Union [1], etc. In this paper will select iTHANET: IthaGenes which is modern than HbVar, and HbVar has 483 mutations while iTHANET: IthaGenes has 615 mutations 63% related to β-thalassemia. Fig. 1 shows the main tasks in finding the optimal BP algorithm, and then used in diagnosis stage:

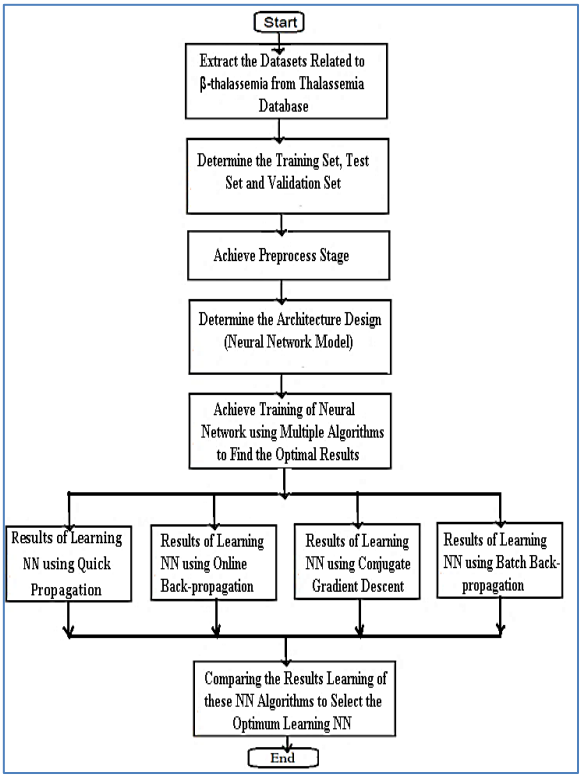

Fig. 1: Flowchart to determining the optimal neural learning algorithm.

There is no single best training algorithm for neural networks, needed to choose a training algorithm based on the characteristics of the problem, but the general-purpose training algorithms of choice as follow:

A. Quick Propagation Algorithm or Quick Prop [5]:

Is an iterative method for determining the minimum of the loss function of an artificial neural network, Fig. 2 shows it's an algorithm.

B. Conjugate Gradient Descent Algorithm:

This learn algorithm contains steps shown in Fig. 3.

C. Online Back-propagation Algorithm:

Uses the error measured on the validation set instead of the training set to dynamically adjust the global learning rate. Fig. 4: Shows the steps of learning this algorithm [2].

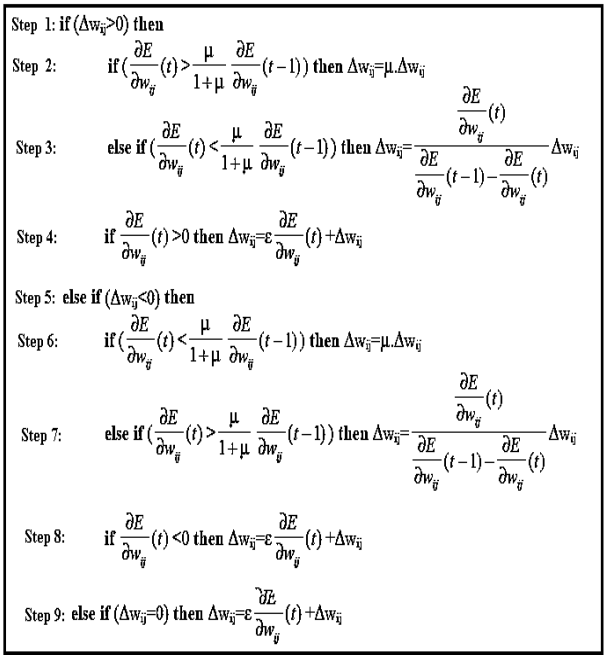

Fig. 2: Shows main steps related to Quick Propagation.

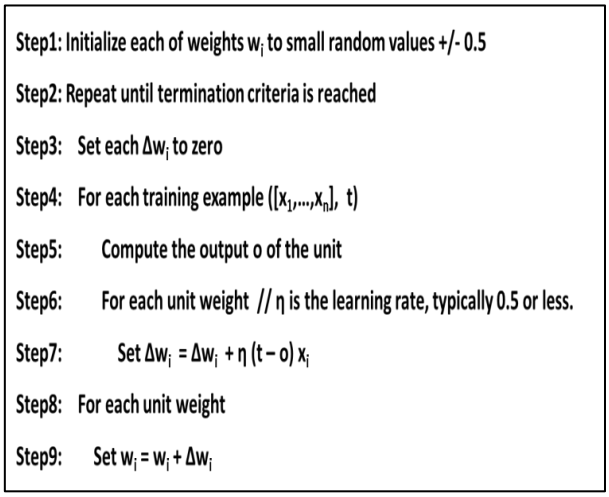

Fig. 3: Conjugate Gradient Descent Neural Learning Algorithm [10]

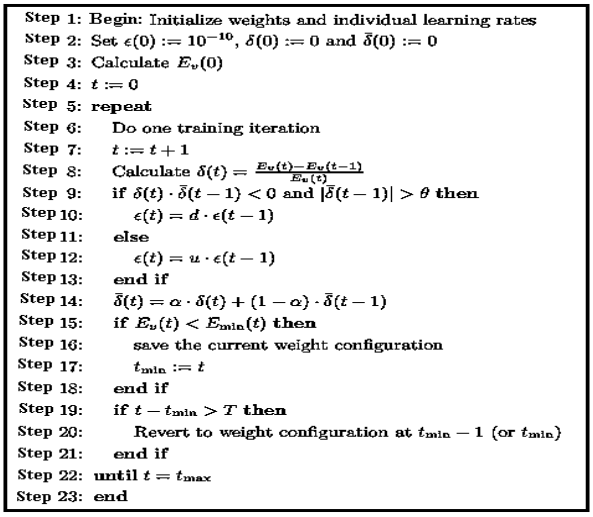

Fig. 4: Shows the steps of online BP learning algorithm.

D. Batch Back-propagation Algorithm:

Same to work on the whole dataset to perform learning. Fig. 5 Shows the main steps of learning this algorithm [3]:

Step 1 **begin initialize** network topology (# hidden units), w, criterion $\theta, \eta, r \leftarrow 0$
Step 2 **do** $r \leftarrow r + 1$ (increment epoch)
Step 3 $m \leftarrow 0$; $\Delta w_{ij} \leftarrow 0$; $\Delta w_{jk} \leftarrow 0$
Step 4 **do** $m \leftarrow m + 1$
Step 5 $\mathbf{x}^m \leftarrow$ select pattern
Step 6 $\Delta w_{ij} \leftarrow \Delta w_{ij} + \eta \delta_j x_i$; $\Delta w_{jk} \leftarrow \Delta w_{jk} + \eta \delta_k y_j$
Step 7 **until** $m = n$
Step 8 $w_{ij} \leftarrow w_{ij} + \Delta w_{ij}$; $w_{jk} \leftarrow w_{jk} + \Delta w_{jk}$
Step 9 **until** $\nabla J(\mathbf{w}) < \theta$
Step 10 **return** w
Step 11 **end**

Fig. 5: Batch Backpropagation learning algorithm [3].

2. Diagnosis/Classify the mutation in person's gene is caused β-thalassemia:

In this stage will use the optimum learning algorithm in stage (1) above to classify/diagnosis the mutation in Person gene is caused β-thalassemia [11, 12, 13]. Fig. 6 shows the steps of algorithm needed to this task.

4. Experimental Results

The proposed biomining Technique for diagnosis mutations of β-thalassemia is simulated using Alyuda NeuroIntelligence ANI ver. 2.1 on Laptop Intel® Core i5 processors. The dataset related to β-thalassemia extracted about (384) records from common database in ITHALNET-IthaGenes http://www.ithanet.eu/db/ithagenes [1], where these dataset not enough for proposed technique, so constructed sub-database related to β-thalassemia using proper gene (NM_000518 vs Genomic) via HbVar DBcan obtained via HbVar database http://globin.bx.psu.edu/cgi-bin/hbvar/query_vars3 [9], as shown in Fig. 7, also will need the Refs-HBB DNASeq Exons and Introns can obtain using the URL:
http://genatlas.medecine.univ-paris5.fr/11/html/HBB_1.html, as shown in Fig. 8.

Input: Proper/Healthy gene & Person (or patient) gene sequence.
Output: Diagnose there is malignant mutation in Person gene sequence or Not.

BEGIN
Step1: Use ClustalW (via BioEdit/online) to check similarity Between proper gene and Person Sequences
Step2: If there is matching
Person's gene is proper/healthy
Step3: Else
Convert Person's DNA sequence to Protein sequence
Step4: Apply ClustalW to check similarity between proteins
Step5: If there is matching
There isn't risk, i.e. Person's gene is proper/healthy
Step6: Else
There is mutation, but can't say causes thalassemia
Step7: Call the optimal training Back-propagation neural network to classify/diagnosis the mutation is malignant or not.
Step8: If there is matching
There is risk, i.e. Person's gene will cause thalassemia as referring in Table (1, 2, 3 or 4).
Step9: Else
If there is symptoms of thalassemia
Step10: Recognize new mutation (Unknown)
Step11: Else
There isn't risk, however recognition mutation.
END

Fig. 6: Shows the algorithm of classify/diagnose the mutation in Person gene caused β-thalassemia

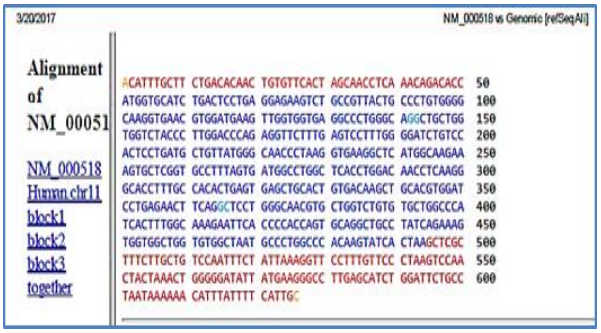

Fig. 7: Shows the proper gene (NM_000518 vs Genomic)

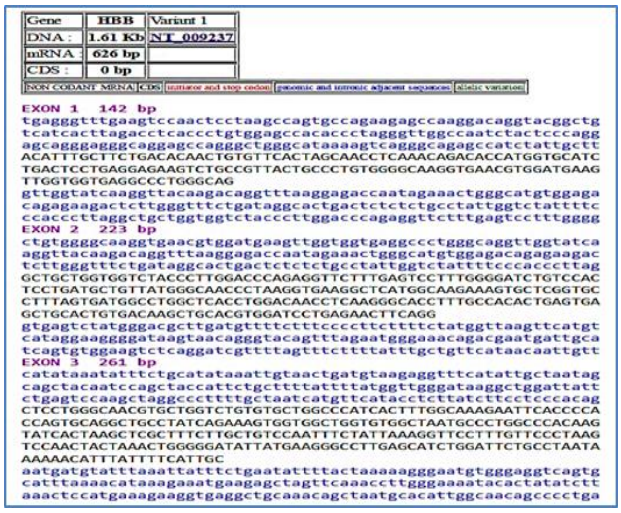

Fig. 8: Shows the Refs-HBB DNASeq Exons and Introns

Sample of this sub-database shown in Fig 9, i.e. an effective fields (IThaID, Position, Normal Base, Mutated Base and Phenotype) selected from dataset and the target field was (Phenotype). ANI can select an ideal train set TRN 68%, validation set VLD and test set TST each of them 16% of all records to train/learning to classify/diagnosis β-thalassemia.

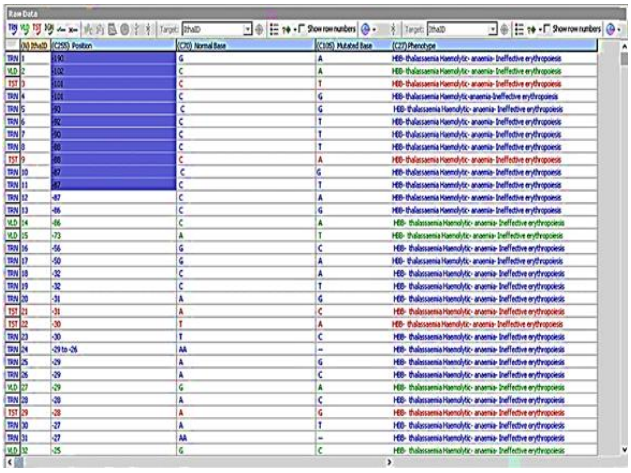

Fig. 9: Shows sample of β-thalassemia dataset extracted from ITHALNET-IthaGenes for NN training algorithm.

ANI gives ideal topology (457-228-1) for training with (1000) iteration, and implementing of the proposed biomining technique for diagnosis/classify β-thalassemia as follow:

1) Determine the optimal NN algorithm for biomining technique of mutations caused disease of β-thalassemia:

The implement and simulating the dataset of β-thalassemia using the proposed biomining technique to find the optimal learning algorithm shows the results as in Table 5.

Table 5: Results of Training and Testing Stages of Proposed Biomining Method for Diagnosis Mutation Caused B-Thalassemia

Training Algorithm	*Absolute Error/CCR, %*		*Network Error*		*Error improvement*	*Training Speed, Iter./Sec.*	*Architecture*	*Testing Summary*	
	TRN	VLD	TRN	VLD				*Correlate*	*R-Squared*
Quick Propagation	7.813026	121.0246	0.00099	0	5.20E-08	1.26448	[457-228-1]	0.997086	0.994173
Online Back-propagation	226.04137	276.20123	0.000621	0	0.000013	0.795716	[457-228-1]	0.975277	0.900057
Batch Back-propagation	-100	-100	85.364176	0	0.257249	0.871789	[457-228-1]	0.975762	0.931719
Conjugate Gradient Descent	8.854733	80.350636	0.000083	0	1.15E-08	0.275514	[457-228-1]	0.999601	0.9992

Based the results reached in test stage the optimum result of this case study of diagnosis/classify the mutation in gene β-thalassemia causes disease or not can obtained using conjugate gradient descent algorithm which gives highest correlation (0.999601) and R-Squared (0.9992) comparing to other NN algorithms (quick propagation, online backpropagation and batch backpropagation) as shown in Fig.10.

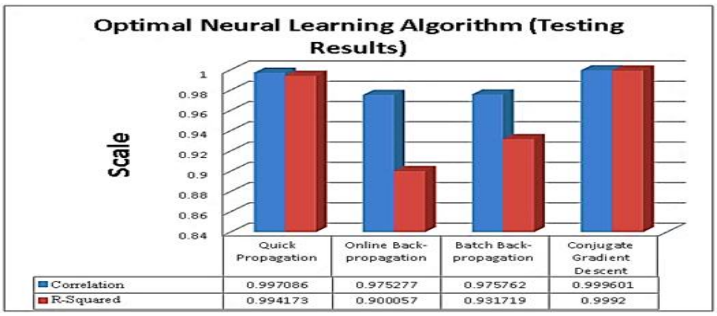

Fig. 10: Shows conjugate gradient descent algorithm highest correlation (0.999601) and R-Squared (0.9992) comparing to other NN algorithms.

2. Diagnosis/Classify the mutation in person's gene is caused β-thalassemia:
As referring in proposed algorithm steps for this stage as shown if Fig. 6 above, the first task needed a classifying mutation by check similarity between proper β-thalassemia gene sequence as shown in Fig. 7 above and Person gene sequence (which can obtain via genetic testing) using CLASTALW. When classifying/diagnosis shows there is different between the proper gene and Person gene sequences that means there is mutation otherwise there isn't. In case of classifying there is mutation needed checkup the proteins sequences of proper β-thalassemia and Person's protein of β-thalassemia, if there is different in these proteins sequences that means there is mutation in gene and protein of Person [11, 12, 13] otherwise there isn't mutation or no risk. The next task needed to determine this mutation which classified will cause β-thalassemia (i.e. one of mutations with dataset/sub-database of β-thalassemia) or not. The proposed biomining technique using optimal NN algorithm, which is conjugate gradient descent algorithm as referring in subsection 4; 1. This second diagnosis/classifying achieved via ANI by implemented the feature of multiple queries using conjugate gradient descent algorithm, one of them shown in Fig. 11.

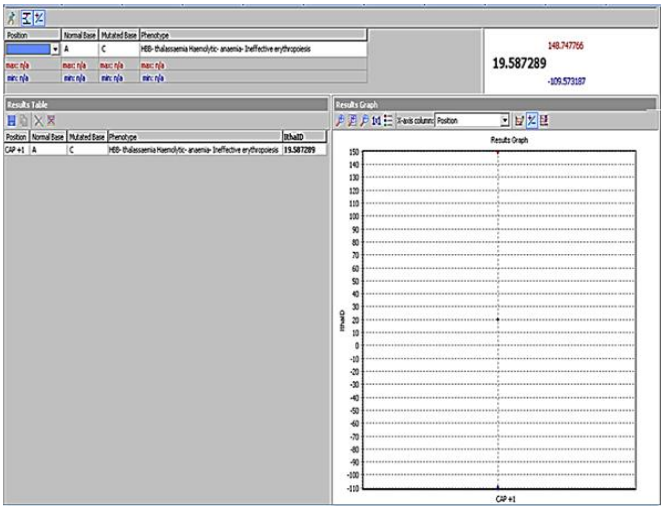

Fig. 11: Using the query of conjugate gradient descent algorithm to determine the person holder or patient of β-thalassemia.

3. Discussion the Results:
Table 6 shows comparing the results of proposed biomining technique with other techniques.

5. Conclusions and future works

The implement and applied the proposed biomining technique of diagnosis mutation caused β-thalassemia shows the following conclusions:
A. The proposed technique works on whole dataset related of β-thalassemia gene as referring to in section 4.
B. This biomining technique is effective in diagnosis and classify mutations of β-thalassemia using optimal learning algorithm (conjugate gradient descent learning algorithm), as shown in Table 4 and Fig. 10.
C. The suggested biomining technique allows confirming diagnosis/classify via mutations β-thalassemia in two sequences the genome and proteome.

Table 6: Reveals Comparison of Proposed Biomining Technique With Other Techniques

The Proposed Biomining Technique	DOMINGOS, Ana L. B. and et al. [8]	Altug Akay, Andrei Dragomir and el at. [6]	Ou XB, Zhang L and el at. [7]
Using multiple learning NN algorithm to find an optimal one	Traditional data mining tools	Traditional data mining	Genechips technology in classification
Using an effective fields (5 only) of whole dataset (384 records) related	Blood samples in a relational database, over the years	β-thalassemia's socioeconomic geography	Limited sample in period (July 2002 to July 2003) and size (115 male and 40 female)
Using genome dataset based genetic test to classify mutations of β-thalassemia	Traditional dataset as results of classic Lab.	Self-organizing maps to identify underlying data structure	DNA was extracted from ACD coagulated blood with Invisorb DNA extraction kit
Genome & Proteome sequences used to classify mutations of β-thalassemia.	There isn't	There isn't	Based on genome sequence only

Results Obtained more effective, since classified based on optimal algorithm	There isn't learning NN algorithm	There isn't learning NN algorithm	There isn't learning NN algorithm
Using comprehensive European (ITHALNET-IthaGenes) database	Locally (Brazilian citizen)	Locally	Locally

D. The future works are:

- Implement and applied the proposed biomining technique comprehensive genes (δ, α, γ, etc.) caused thalassemia not β-thalassemia only.
- Using another package/programming language like MATLAB/Java in implement and applied the biomining techniques to comparing the results and determine the better tool for this proposed technique.
- The confirmed results which obtained in this proposed biomining technique can use for therapy like replacement or correction the mutations caused β-thalassemia which classified or diagnosed.

References

[1] "IthaGenes: Homepage", supported by European Union, created at period (2006-2008). URL: http://www.ithanet.eu/db/ithagenes

[2] Stefan Duffner, Christophe Garcia , "An Online Backpropagation Algorithm with Validation Error-Based Adaptive Learning Rate", Springer Link, International Conference on Artificial Neural Networks ICANN 2007, DOI: 10.1007/978-3-540-74690-4_26, pp 249-258. URL:https://link.springer.com/chapter/10.1007/978-3-540-746904_26

[3] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification, Chapter 6: Multi-layer feedforward neural networks, ISBN: 978-0-471-05669-0, 2nd Edition WILEY, November 2000. URL: http://eu.wiley.com/WileyCDA/WileyTitle/productCd0471056693.html

[4] João Carlos Negrão Ventura, "NEURAL NETWORKS IMPLEMENTATION IN PARALLEL DISTRIBUTED PROCESSING SYSTEMS", Departamento de Informática Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa. URL: http://venturas.org/sites/venturas.org/files/mydmbp.pdf

[5] Clemens-Alexander Brust and el. at. , "Evaluation of QuickProp for Learning Deep Neural Networks -- A Critical Review", June 2016. URL: https://www.researchgate.net/publication/303969910

[6] Altug Akay, Andrei Dragomir and el at, "A Data Mining Approach for Investigating Social and Economic Geographical Dynamics of Thalassemia's Spread ", Published in: IEEE Transactions on Information Technology in Biomedicine (Volume: 13, Issue: 5, Sept. 2009), DOI: 10.1109/TITB.2009.2020062. Pages 774-780. http://ieeexplore.ieee.org/document/4814673/

[7] Ou XB, Zhang L and el at, "Diagnosis of thalassemia by using genechips", Zhonghua Er Ke Za Zhi. 2005 Jan;43(1):31-4. https://www.ncbi.nlm.nih.gov/pubmed/15796804

[8] DOMINGOS, Ana L. B. et al. "The profile of beta thalassemia obtained by data mining analysis in a database", Revista Brasileira de Hematologia e Hemoterapia ABHH, vol.32, n1,P 78-79. ISSN 15168484. (2010). http://www.scielo.br/pdf/rbhh/v32n1/a18v32n1.pdf

[9] Prof. Titus H.J. Huisman, Mrs. Marianne F.H. Carver and el at, "HbVar: A database of Human Hemoglobin Variants and Thalassemias", 1990s. URL: http://globin.bx.psu.edu/cgi-bin/hbvar/query_vars3

[10] Pierre de Lacaze, "Reinforcement Learning and Artificial Neural Nets", September 20th, 2016. rpl@lispnyc.org , pierre@shareablee.com . URL: https://www.slideshare.net/delaray/reinforcement-learning-and-artificial-neural-nets

[11] A. Gh. Ismaeel, and R. Zuhair Yousif. " Novel Mining of Cancer via Mutation in Tumor Protein P53 using Quick Propagation Network", International Journal of Computer Science and Electronics Engineering (IJCSEE),vol.3, issue 2, pp.121-126, (2015).

[12] A. Gh. Ismaeel, and A. A. Ablahad, "Novel Method for Mutational Disease Prediction using Bioinformatics Techniques and Backpropopagation Algorithm", IRACST- Engineering Science and Technology: An International Journal Vol. 3, pp. 150-156, 2013, (online).

[13] A. Gh. Ismaeel, and A. A. Ablahad, "Enhancement of a Novel Method for Mutational Disease Prediction using Bioinformatics Techniques and Backpropagation Algorithm", international journal of scientific & engineering research, vol. 4, issue 6, 2013.