ORIGINAL ARTICLE

IN-SILICO ASSESSMENT OF COMMON β-GLOBIN GENE MUTATIONS FOUND IN KHYBER PAKHTUNKHWA, PAKISTAN

Tehmina Jalil, Ibrahim Rashid*, Inayat Shah, Ilyas Khan***, Yasar Mehmood Yousafzai****

Department of Pathology, Khyber Girls Medical College, Peshawar, *Atta ur Rahman School of Applied Biosciences, National University of Sciences and Technology, Rawalpindi, **Institute of Basic Medical Sciences, Khyber Medical University, ***Pak International Medical College, Peshawar, Pakistan

Background: β-thalassaemia manifests a spectrum of clinical phenotypes, ranging from mild subclinical disease to severe transfusion-dependent anaemia. This remarkable diversity of disease patterns is not completely explained, but various disease modifying factors have been identified and categorized in to primary, secondary and tertiary modifiers. Nearly 300 mutations in β-globin genes (HBB) have been reported so far. In our previous study we identified six mutations; (Cd 5 (-CT), FR 8-9(+G), FR 16(-C), FR 41-42(-TTCT), Cd 30(G>A) and Cd 15(G>A) to be the most frequent ones in Khyber Pakhtunkhwa province of Pakistan. It is aimed to observe if bioinformatics tools could be used to construct the structure of globin chains carrying these six common mutations. **Methodology:** Using a computational approach, the sequences of mutated HBB were hypothetically constructed, protein structures were formulated and analysed for homology, and post-translational modifications. In mutations where protein structure formation is halted *in vivo*, stop codons from the DNA sequence of each of the mutational variant were exclude to allow further analysis. **Results:** These mutants exhibited variable post-translational modification pattern with little effect on overall structure. Mutations at critical sequences in HBB that do not allow further translation of HBB *in vivo* and did not stop computer modelling from developing protein structure *in-silico*. **Conclusion:** Computational analysis for constructing mutant proteins does not take into account some of the critical checkpoints present in the cell. Studies using computational analysis should be followed by rigorous *in vivo* validation.

Keywords: Thalassaemia, β-globin genes, Post-translational Modification, Computational Genetic Polymorphic Analysis

Pak J Physiol 2018;14(3):68–73

INTRODUCTION

Thalassaemias are a group of congenital disorders characterized by genetic mutations in the globin chain genes.[1] The primary defect is quantitative, whereby the mutation(s) results in reduced production or absence of globin chains of haemoglobin molecule. Other less common forms include structural variants or unstable haemoglobins due to genetic mutations. Based on the globin chain gene involved, the disease is categorized into α-, β-, γ-, and δβ-thalassaemia.[2] Of these, β-thalassaemia is most prevalent form in Pakistan. With an autosomal recessive pattern, an estimated 10 million people are carriers and approximately 100,000 patients are currently affected by homozygous thalassaemia.[3,4] Like many other hereditary disorders, β- thalassaemics express a spectrum of clinical phenotypes —ranging from mild anaemia to severe transfusion-dependent anaemia. The pattern of diversity is not completely understood, but various modifying factors have been identified and categorized into primary, secondary and tertiary modifiers of disease.[5]

Nearly 300 mutations in β-globin genes have been reported so far —most of them are point mutations in the coding or regulatory regions of the gene (http://globin.cse.psu.edu). Many common mutations involve frameshift whereby the subsequent sequence is entirely altered, and the transcription process is halted. Secondary modifiers of the disease include loci that are involved in globin synthesis.[6] Family studies conducted on siblings with identical β-globin gene mutations have been reported to demonstrate variable clinical phenotype. The secondary modifiers include: co-inheritance of α-globin genes which reduces the cellular damage caused by excess α-chains, and polymorphisms in Xmn1 and BCL11 genes causing a dilutional effect on the α-chain excess. Tertiary modifiers include genetic variations associated with disease complications and treatment response. These include mutations in iron metabolism genes, and genes involved in immune system resulting in altered response to infections.[7]

As with most genetic diseases, the mutations in β-thalassaemia pool in each geographic region/ ethnic population presents with a different set of common mutations. In Pakistan, a number of common mutations have previously been reported.[8] In our own work (in press), we have reported that (Cd 5(-CT), FR 8-9(+G), FR 16(-C), FR 41-42(-TTCT), Cd 30(G>A) and Cd 15(G>A) are the six most common mutations in patients from Khyber Pakhtunkhwa. The first four of these mutations are frameshift mutations. The last two are splice, and nonsense mutation respectively.

The current study aimed if bioinformatics tools could be used to construct the structure of globin chains carrying these 6 most common mutations. In a hypothetical scenario where most of the selected mutations did not incorporate early stop codons, it was aimed to assess the effect of these mutations on overall protein structure, post translational modifications and subsequent impairment of haemoglobin protein. To that end, the stop codons from the DNA sequence of each of the mutational variant were excluded. It is reported that mutations constructed through bioinformatics tools carry significant differences from HBB chains. The computational analysis failed to account for the molecular checkpoints that exist *in vivo* and halt further assembly of protein.

MATERIAL AND METHODS

Primary data (nucleotide and peptide sequences) were retrieved from GenBank and Uniprot data bases (www.ncbi.nlm.nih.gov/, and www.uniprot.org/). For each of the selected six mutations (Cd 5(-CT), FR 8-9(+G), FR 16(-C), FR 41-42(-TTCT), Cd 30(G>A) and Cd 15(G>A), the normal HBB gene sequences were manually mutated and six different sequences were developed. The mutant nucleotide sequences were translated computationally using online ExPASy-Translate tool (http://web.expasy.org/translate/). Sequences were submitted to the online MSA tool Clustal Omega (www.ebi.ac.uk/Tools/msa/clustalo/) with default parameters. The subsequent outcome of the MSA was subjected to Phylogenetic Tree development using the Tree development (Neighbour-joining) option in Clustal Omega tool.[9] Online servers provided by Center for Biological Sequence Analysis CBS (www.cbs.dtu.dk/services/) were employed for assessment of post-translational modifications. For observing glycosylation, mannosylation, phosphorylation sites in the normal and mutant sequences; NetCGlyc 1.0[10], NetCorona 1.0[11], NetGlycate 1.0[12], NetNGlyc 1.0[13], NetOGlyc 4.0[14], NetPhos 3.1[15] servers were used respectively.

For normal HBB, the crystal (experimentally determined) structure was searched and retrieved from Protein DataBank (www.rcsb.org/). This crystal structure was used as template for development of 3d structures for each of the mutant sequence using Swiss Model online protein prediction server (https://swissmodel.expasy.org/).[16] The resultant structures were analyzed by Rampage server for Ramachandran plot in order to evaluate the predicted structure quality.[17] Predicted protein structures were visualized via Biovia Discovery Studio version 4.5.[18] https://wishart.biology.ualberta.ca/SuperPose/ (The SuperPose server) was used for protein structure-structure alignment for observing effects of mutations on the protein 3d structure.

RESULTS

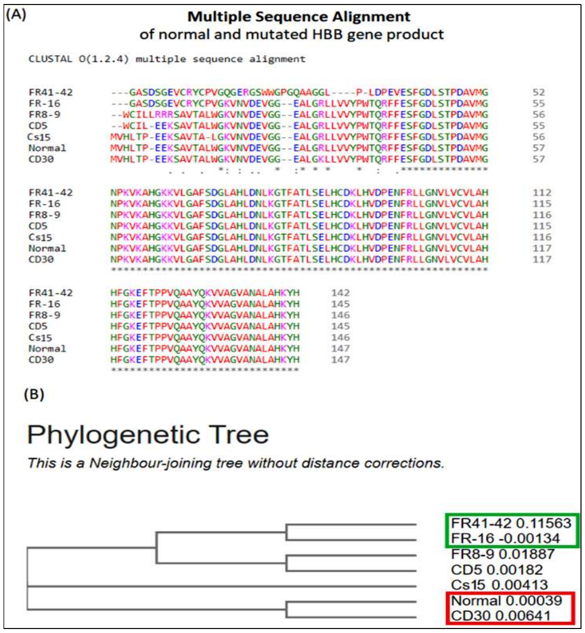
The retrieved HBB sequences were DNA (NCBI id NC_000011.10) and protein (UniProt id P68871). The variant sequences and subsequent translated products are shown in Table-1. Homology and phylogenetic analysis revealed that Cd30 was most homologous to the normal HBB in terms of sequences similarity while variants with frameshift mutations, i.e., FR 8-9, FR 16, and FR 41-42 were least homologous. The MSA revealed the Fr 41-42 to be most distant in terms of sequence identity with a unique gap in the alignment. Two distinct gap patterns were observed in the FR mutations (FR 8-9(+G), FR 16(-C), and FR 41-42(-TTCT)) and CD mutations (CD5 (-CT), CD30(G>A), and CD 15(G>A)). Interestingly CD5(-CT) exhibited the both patterns (Figure-1 A). This sequential arrangement resulted in Cd5 (-CT) to at the center of the Phylogenetic Tree (Figure-1 B). These gaps were the early stop codons which we deleted in the transcript and hence appeared as *deletions* or *gaps* in the translated sequences. CBS servers observed the submitted sequences and provided a comprehensive overview of the post-translational modifications for HBB normal and mutant sequences. The default cut-off was 0.5 for each of the prediction tool, with greater values suggesting greater probability. No C-mannosylation sites were observed in the sequences using NetCGlyc 1.0 server in neither the protein nor the variants. NetGlycate 1.0 server computed about 3–5 similar glycation sites while 0 N-linked Glycosylation sites in all the submitted sequences while only Cd15(G>A) variant sequence harbored the only predicted O-linked Glycosylation site, computed by NetNGlyc1.0 and NetOGlyc 4.0 servers respectively. About 7-11 similar phosphorylation sites were estimated by NetPhos 3.1 server. The 6 sites observed in the normal sequence were observed in all of the variants with exception of phosphorylation site at position 5 which was absent in Cd5 (-CT), FR 8-9 (+G), FR 16 (-C), and FR 41-42 (-TTCT). Interestingly FR 8-9 (+G) and FR 41-42 (-TTCT) both exhibited two unique phosphorylation sites at position 9 and 12 and 21 and 40 respectively. All PTM sites are shown in Table-1. Crystal structure of human HBB (PDB ID 1A00) was retrieved from PDB database. The superimposed variants modelled using Swiss Model are shown in Figure-2. Each of the variant structure is shown in red while the normal HBB structure is depicted in yellow. Overall no drastic change was observed in the variant structures while when Ramachandran plots were observed in comparative manner minor effects were observed in placement and strain on amino acids within the structures.

Table-1: List of HBB mutations and respective nucleotide and peptide sequences

Mutations	Nucleotide sequence	Translated product
>CD5	ATGGTGCATCTGACTCGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAAC GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCC AGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCT AAGGTGAAGGCTCATGGCAAGAAAGTGCTC GGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACAC TGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGG CAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTG CAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATC ACTAA	WCILEEKSAVTALWGKVNV DEVGGEALGRLLVVYPWTQ RFFESFGDLSTPDAVMGNPK VKAHGKKVLGAFSDGLAHL DNLKGTFATLSELHCDKLH VDPENFRLLGNVLVCVLAH HFGKEFTPPVQAAYQKVVA GVANALAHKYH
>FR 8-9	ATGGTGCATCTGACTCCTGAGGAGAAGGTCTGCCGTTACTGCCCTGTGGGGCAAGGTG AACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGA CCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAAC CCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTC ACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCT GCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCC CATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGG CTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA	WCILLRRRSAVTALWGKVN VDEVGGEALGRLLVVYPWT QRFFESFGDLSTPDAVMGNP KVKAHGKKVLGAFSDGLA HLDNLKGTFATLSELHCDK LHVDPENFRLLGNVLVCVL AHHFGKEFTPPVQAAYQKV VAGVANALAHKYH
>FR 16	ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGAAGGTGAA CGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACC CAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCC TAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCAC CTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGC ACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCT GGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA	GASDSGEVCRYCPVGKVNV DEVGGEALGRLLVVYPWTQ RFFESFGDLSTPDAVMGNPK VKAHGKKVLGAFSDGLAHL DNLKGTFATLSELHCDKLH VDPENFRLLGNVLVCVLAH HFGKEFTPPVQAAYQKVVA GVANALAHKYH
>FR 41 42	ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGA ACGTGGATGAAG TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTGAG TCCTTTGG GGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAG AAAGTGCTCGGT GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGA GTGAGCTGCACT GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTG TGTGCTGGCCCA TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCT GGTGTGGCTAAT GCCCTGGCCCACAAGTATCACTAA ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGA ACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGAC CCAGAGGTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTA AGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCT GGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCAC GTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATC ACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGG TGTGGCTAATGCCCTGGCCCACAAGTATCACTAA	GASDSGEVCRYCPVGQGER GSWWGPGQAAGGLPLDPE VESFGDLSTPDAVMGNPKV KAHGKKVLGAFSDGLAHLD NLKGTFATLSELHCDKLHV DPENFRLLGNVLVCVLAHH FGKEFTPPVQAAYQKVVAG VANALAHKYH
>CD30	ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGA ACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAAGCTGCTGGTGGTCTACCCTTGGAC CCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACC CTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCA CCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTG CACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCC ATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGC TGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA	MVHLTPEEKSAVTALWGK VNVDEVGGEALGKLLVVYP WTQRFFESFGDLSTPDAVM GNPKVKAHGKKVLGAFSD GLAHLDNLKGTFATLSELH CDKLHVDPENFRLLGNVLV CVLAHHFGKEFTPPVQAAY QKVVAGVANALAHKYH
>Cs 15	ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGAGGCAAGGTGA ACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGAC CCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACC CTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCA CCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTG CACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCC ATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGC TGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA	MVHLTPEEKSAVTALGKVN VDEVGGEALGRLLVVYPWT QRFFESFGDLSTPDAVMGNP KVKAHGKKVLGAFSDGLA HLDNLKGTFATLSELHCDK LHVDPENFRLLGNVLVCVL AHHFGKEFTPPVQAAYQKV VAGVANALAHKYH

Table-2: List of predicted post-translational sites

Sequences	Post-Translational Sites							
	netglycate-1.0 prediction results			# netphos-3.1b prediction results				
	position	Score	Site nature	Position	Serine/ Threonine	Context	Score	Kinase
Normal	9	0.89	glycate	5	T	MVHLTPEEK	0.557	p38MAPK
	18	0.778	glycate	5	T	MVHLTPEEK	0.929	unsp
	67	0.842	glycate	39	T	VYPWTQRFF	0.633	PKC
	121	0.825	glycate	45	S	RFFESFGDL	0.571	CKI
	145	0.779	glycate	45	S	RFFESFGDL	0.62	unsp
				50	S	FGDLSTPDA	0.987	unsp
				51	T	GDLSTPDAV	0.529	p38MAPK
				88	T	GTFATLSEL	0.624	PKA
				88	T	GTFATLSEL	0.586	unsp
CD5	7	0.796	glycate	37	T	VYPWTQRFF	0.633	PKC
	16	0.781	glycate	43	S	RFFESFGDL	0.571	CKI
	65	0.841	glycate	43	S	RFFESFGDL	0.62	unsp
	119	0.825	glycate	48	S	FGDLSTPDA	0.987	unsp
	143	0.779	glycate	49	T	GDLSTPDAV	0.529	p38MAPK
				86	T	GTFATLSEL	0.624	PKA
				86	T	GTFATLSEL	0.586	unsp
Fr 8-9	17	0.93	glycate	9	S	LRRRSAVTA	0.951	unsp
	66	0.841	glycate	9	S	LRRRSAVTA	0.856	PKA
	120	0.825	glycate	9	S	LRRRSAVTA	0.646	PKG
	144	0.779	glycate	12	T	RSAVTALWG	0.561	PKC
				38	T	VYPWTQRFF	0.633	PKC
				44	S	RFFESFGDL	0.62	unsp
				44	S	RFFESFGDL	0.571	CKI
				49	S	FGDLSTPDA	0.987	unsp
				50	T	GDLSTPDAV	0.529	p38MAPK
				87	T	GTFATLSEL	0.624	PKA
				87	T	GTFATLSEL	0.586	unsp
Fr 16	16	0.719	glycate	37	T	VYPWTQRFF	0.633	PKC
	65	0.841	glycate	43	S	RFFESFGDL	0.62	unsp
	119	0.825	glycate	43	S	RFFESFGDL	0.571	CKI
	143	0.779	glycate	48	S	FGDLSTPDA	0.987	unsp
				49	T	GDLSTPDAV	0.529	p38MAPK
				86	T	GTFATLSEL	0.624	PKA
				86	T	GTFATLSEL	0.586	unsp
Fr 41-42	62	0.841	glycate	21	S	GERGSWWGP	0.659	PKA
	116	0.825	glycate	21	S	GERGSWWGP	0.557	unsp
	140	0.779	glycate	40	S	PEVESFGDL	0.98	unsp
				40	S	PEVESFGDL	0.521	CKI
				45	S	FGDLSTPDA	0.987	unsp
				46	T	GDLSTPDAV	0.529	p38MAPK
				83	T	GTFATLSEL	0.624	PKA
				83	T	GTFATLSEL	0.586	unsp
cd30	9	0.89	glycate	5	T	MVHLTPEEK	0.929	unsp
	18	0.778	glycate	5	T	MVHLTPEEK	0.557	p38MAPK
	67	0.842	glycate	39	T	VYPWTQRFF	0.633	PKC
	121	0.825	glycate	45	S	RFFESFGDL	0.62	unsp
	145	0.779	glycate	45	S	RFFESFGDL	0.571	CKI
				50	S	FGDLSTPDA	0.987	unsp
				51	T	GDLSTPDAV	0.529	p38MAPK
				88	T	GTFATLSEL	0.624	PKA
				88	T	GTFATLSEL	0.586	unsp
cs15	9	0.834	glycate	5	T	MVHLTPEEK	0.929	unsp
	66	0.841	glycate	5	T	MVHLTPEEK	0.557	p38MAPK
	120	0.825	glycate	38	T	VYPWTQRFF	0.633	PKC
	144	0.779	glycate	44	S	RFFESFGDL	0.62	unsp
				44	S	RFFESFGDL	0.571	CKI
				49	S	FGDLSTPDA	0.987	unsp
				50	T	GDLSTPDAV	0.529	p38MAPK
				87	T	GTFATLSEL	0.624	PKA
				87	T	GTFATLSEL	0.586	unsp


(A)
Multiple Sequence Alignment
of normal and mutated HBB gene product
(B)
Phylogenetic Tree
This is a Neighbour-joining tree without distance corrections.
FR41-42 0.11563
FR-16 -0.00134
FR8-9 0.01887
CD5 0.00182
Cs15 0.00413
Normal 0.00039
CD30 0.00641

```
CLUSTAL O(1.2.4) multiple sequence alignment

FR41-42         ---GASDSGEVCRYCPVGQGERGSWWGPGQAAGGL----P-LDPEVESFGDLSTPDAVMG     52
FR-16           ---GASDSGEVCRYCPVGKVNVDEVGG--EALGRLLVVYPWTQRFFESFGDLSTPDAVMG     55
FR8-9           --WCILLRRRSAVTALWGKVNVDEVGG--EALGRLLVVYPWTQRFFESFGDLSTPDAVMG     56
CD5             --WCIL-EEKSAVTALWGKVNVDEVGG--EALGRLLVVYPWTQRFFESFGDLSTPDAVMG     55
Cs15            MVHLTP-EEKSAVTA-LGKVNVDEVGG--EALGRLLVVYPWTQRFFESFGDLSTPDAVMG     56
Normal          MVHLTP-EEKSAVTALWGKVNVDEVGG--EALGRLLVVYPWTQRFFESFGDLSTPDAVMG     57
CD30            MVHLTP-EEKSAVTALWGKVNVDEVGG--EALGKLLVVYPWTQRFFESFGDLSTPDAVMG     57
                         . .  .  *: : ..  *  :* * *    *  :  .**************

FR41-42         NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     112
FR-16           NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     115
FR8-9           NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     116
CD5             NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     115
Cs15            NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     116
Normal          NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     117
CD30            NPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH     117
                ************************************************************

FR41-42         HFGKEFTPPVQAAYQKVVAGVANALAHKYH     142
FR-16           HFGKEFTPPVQAAYQKVVAGVANALAHKYH     145
FR8-9           HFGKEFTPPVQAAYQKVVAGVANALAHKYH     146
CD5             HFGKEFTPPVQAAYQKVVAGVANALAHKYH     145
Cs15            HFGKEFTPPVQAAYQKVVAGVANALAHKYH     146
Normal          HFGKEFTPPVQAAYQKVVAGVANALAHKYH     147
CD30            HFGKEFTPPVQAAYQKVVAGVANALAHKYH     147
                ******************************
```

Figure-1: HBB Homology and Phylogenetic Analysis. (A) Multiple sequence alignment (B) Phylogenetic tree. Both analyses revealed Cd30 to be more identical and homologous to normal HBB while Fr41-42 was the least identical variant

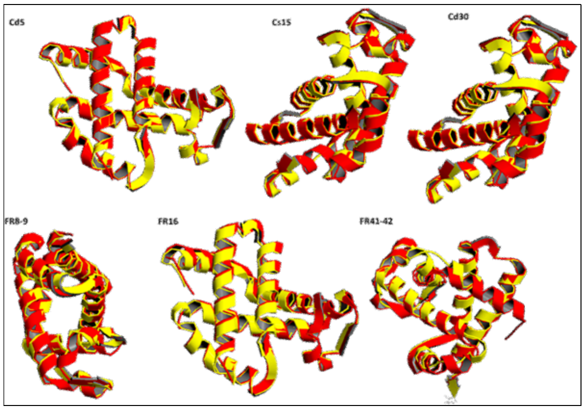


Figure-2: Built 3-D structures of the HBB variants superimposed on the normal structure. Each variant is shown in red while normal HBB is depicted in yellow

DISCUSSION

In this study we have focused on the effects of HBB mutations on sequence, post-translational modification, and structure using *in-silico* computational approaches. Hypothetically all the nonsense mutations were assumed to be missense. Moreover, early stop codons from the translated sequences were removed. This strategy relied on prediction tools available in public domain.

Post-translational modifications in general, have been reported to enhance haemoglobin yield without effecting the gene expression levels.[19] Hence, studying such, in our opinion, was a critical factor in better understanding of haemoglobin disorders. Our selection of HBB mutations was influenced by our previous study (in press) in which prominent HBB genetic polymorphism among various ethnic groups within Khyber Pakhtunkhwa, Pakistan population were reported. The same study observed Fr 8-9 (+G) to be the most frequent or the only mutation in 13 ethnic groups followed by CD 5 (-CT). When considered in reference to geographic distribution, Fr 8-9 (+G) mutation was most common in central regions of Khyber Pakhtunkhwa, i.e., Kurram, Khyber, Peshawar, Charsadah, Mardan, Hangu, Swabi. While distant regions harbored different mutations, i.e., Cd-15 (G>A) in North Waziristan, ISV 1-5/Cap+1 (A-C) in Karak, and Fr 41-42 (-TTCT) in Swat. These reported mutations are in accordance with local and regional reported data.[20–23]

A unique O linked glycosylation site was observed in CD15 (G>A) variant unlike in the other mutants or the normal protein itself. Generally, glycation increases the overall stability of a glycoprotein.[24] Yet for haemoglobin glycation is considered as diagnostic marker for diabetes.[25] Another key post-translational modification was the varying number of Phosphorylation sites. Four such unique sites (two in each) in proteins FR8-9(+G) and FR41-42(-TTCT) make them target for unnecessary victim of PTMs. Unlike the other variants where the sites were shifted up or down stream due to *indel* mutations, these four sites were different in protein sequence. Phosphorylation of eukaryotic initiation factor 2α (eIF2α) is reported to enhance fetal haemoglobin production in thalassemia patients.[19] However, the implication of new phosphorylation sites needs to be studied in more details before deducing any implications. In a surprising manner the built variant structures did not exhibit any prominent structural variation in comparison to the normal structure, yet when Ramachandran plots were studied, the effects of the variations were evident in terms of slight strains on the amino acids as observed due to varying sequence.

The current results underscore the utility of computational modelling in hypothesizing the pathogenesis of genetic disorders. Despite its potential, ease of access and low costs, this approach carries a number of weaknesses. Firstly, human body has a complex ecosystem and processes such as haemoglobin function are modulated by numerous cellular and molecular mechanisms, not taken into consideration during *in-silico* analysis. It is remarkable how protein structures which are halted from production in the body were produced hypothetically. Secondly, lack of understanding of computational biology by clinical scientists makes it difficult to use this in clinical research. Nonetheless, this study highlighted several critical dissimilarities of HBB molecules from mutated sequences which may help in understanding the pathogenesis of reduced production of HBB in thalassaemia. There is a need for greater collaboration between computational biologists and clinical and lab scientists to utilize the potential of computational biology in understanding, preventing and treating disease.

CONCLUSION

HBB genetic polymorphic variants exhibit a varying pattern of post-translational modification sites, i.e., glycation and phosphorylation sites. At least one O linked glycosylation site in CD 15(-CT) and two phosphorylation sites in FR 8-9(+G) and FR 41-42(-TTCT) were predicted. This study is proof-of-principle that thalassaemia genetic mutations not only reduce the amount of globin chain synthesis, but also produce differentially structured proteins, adding to the complexity of the genotype-phenotype relationship. Integration of computational biology and clinical sciences will be required to fully utilize the potential of *in-silico* studies.

REFERENCES

1. Taher AT, Weatherall DJ, Cappellini MD. Thalassaemia. Lancet 2018;391(10116):155–67.
2. Thein SL. Molecular basis of beta thalassemia and potential therapeutic targets. Blood Cells Mol Dis. 2018;70:54-65.
3. Shamsi TS. Beta-thalassaemia —a major health problem in Pakistan. J Pak Med Assoc 2004;54(10):498.
4. Shamsi T, Ansari S. Medical management of beta-thalassaemia without blood transfusion: a myth or a reality? J Pak Med Assoc 2013;63(3):304–5.
5. Weatherall DJ. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. Nat Rev Genet 2001;2(4):245–55.
6. Thein SL. The molecular basis of beta-thalassemia. Cold Spring Harb Perspect Med 2013;3(5):a011700.
7. Mettananda S, Higgs DR. Molecular Basis and Genetic Modifiers of Thalassemia. Hematol Oncol Clin North Am 2018;32(2):177–91.
8. Yasmeen H, Toma S, Killeen N, Hasnain S, Foroni L. The molecular characterization of Beta globin gene in thalassemia patients reveals rare and a novel mutations in Pakistani population. Eur J Med Genet 2016;59(8):355–62.
9. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, *et al.* Analysis tool web services from the EMBL-EBI. Nucleic Acids Res 2013;41(W1):W597–W600.
10. Julenius K. NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. Glycobiology 2007;17(8):868–76.
11. Kiemer L, Lund O, Brunak S, Blom N. Coronavirus 3CL pro proteinase cleavage sites: Possible relevance to SARS virus pathology. BMC Bioinformatics 2004;5(1):72.
12. Johansen MB, Kiemer L, Brunak S. Analysis and prediction of mammalian protein glycation. Glycobiology 2006;16(9):844–53.
13. Gupta R, Jung E, Brunak S. Prediction of N-glycosylation sites in human proteins. NetNGlyc 1.0.
14. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Katrine T, *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. The EMBO Journal 2013;32(10):1478–88.
15. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 2004;4(6):1633–49.
16. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. Electrophoresis 1997;18(15):2714–23.
17. Lovell SC, Davis IW, Arendall III WB, De Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins: Structure, Function, and Bioinformatics 2003;50(3):437–50.
18. BIOVIA DS. Discovery Studio Modeling Environment, Release 4.5. Dassault Systèmes, San Diego. 2015.
19. Hahn CK, Lowrey CH. Eukaryotic initiation factor 2α phosphorylation mediates fetal hemoglobin induction through a post-transcriptional mechanism. Blood. 2013:blood-2013-03-491043.
20. Baig S, Azhar A, Hassan H, Baig J, Kiyani A, Hameed U, *et al.* Spectrum of beta-thalassemia mutations in various regions of Punjab and Islamabad, Pakistan: establishment of prenatal diagnosis. Haematologica 2006;91(3):ELT02.
21. Black M, Sinha S, Agarwal S, Colah R, Das R, Bellgard M, *et al.* A descriptive profile of β-thalassaemia mutations in India, Pakistan and Sri Lanka. J Community Genet2010;1(3):149–5s7.
22. Rahim F, Abromand M. Spectrum of ß-Thalassemia mutations in various Ethnic Regions of Iran. Pak J Med Sci 2008;24(3):410.
23. Ayub MI, Moosa MM, Sarwardi G, Khan W, Khan H, Yasmin S. Mutation Analysis of the HBB Gene in Selected Bangladeshi β-Thalassemic Individuals: Presence of Rare Mutations. Genetic Testing and Molecular Biomarkers 2010;14(3):299–302.
24. Solá RJ, Griebenow K. Effects of glycosylation on the stability of protein pharmaceuticals. J Pharmaceut Sci 2009;98(4):1223–45.
25. Selvin E, Steffes MW, Zhu H, Matsushita K, Wagenknecht L, Pankow J, *et al.* Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. New Eng J Med 2010;362(9):800–11.

Address for Correspondence:

Dr Tehmina Jalil, Khyber Girls Medical College, Peshawar, Khyber Pakhtunkhwa, Pakistan. **Cell:** +92-345-9116990
Email: tehminajalil3@gmail.com

Received: 4 Jan 2018 Reviewed: 26 Apr 2018 Accepted: 21 May 2018