Relationships of mRNA-protein secondary structures in the human β -globin gene *HBB* and four variants

LI YanFei^{1†}, YE DongHai^{1†}, ZHANG Wen², WANG ChuanMing¹,
LIU CiQuan^{1,3*} & CAO Huai^{1*}

¹ Modern Biological Research Center, Yunnan University, Kunming 650091, China;

² Department of Cell Biology and Genetics, Kunming Medical University, Kunming 650031, China;

³ Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

Received August 7, 2011; accepted December 6, 2011; published online March 8, 2012

Single nucleotide polymorphism is an interesting problem that can alter gene expression, recode amino acids and affect protein function. Protein structural changes have generally been attributed to amino acid replacements, and only a few research efforts have examined the effects of mRNA structural changes to the conformation of the corresponding protein coded by the mRNA. In the present study, the human β -globin *HBB* gene and four variants were examined. The mRNA secondary structures were constructed using the dynamic extended folding method and the encoded protein secondary structures were obtained from related databases. Comparisons were performed between these structures before and after mutations were introduced into the mature mRNAs and the proteins. We focused on the structural changes from mRNA to protein and found that regular protein conformations tend to match stable mRNA regions, whereas irregular protein conformations, such as β/γ turns and random coils, often match unstable mRNA regions. Mutations within unstable regions can alter the mRNA secondary structure and leave footprints in the protein structure. Comparison of the mRNA-protein secondary structure relationships represents a potential strategy to explore protein functional changes.

HBB gene, mutation, mRNA, protein, secondary structure relationships

Citation: Li Y F, Ye D H, Zhang W, et al. Relationships of mRNA-protein secondary structures in the human β -globin gene *HBB* and four variants. *Chin Sci Bull*, 2012, 57: 3898–3907, doi: 10.1007/s11434-012-4996-3

Hemoglobinopathies are among the most studied single gene disorders, and are often classified into abnormal hemoglobins and thalassemias. The former abnormality is due to a globin gene mutation resulting in structural and functional abnormalities in the globin proteins, whereas the latter disease comprises α - and β -thalassemias, caused by a decrease and imbalance of globin synthesis [1]. Many globin gene mutations have been identified as the causes of more than 100 types of hemoglobinopathies which have distinct gene deletions, mutations, and special traits and phenotypes; however, rare types are suitable for structural and functional studies [2].

Four important types of β -globin gene point mutants are Hemoglobin S (HbS), Hemoglobin E (HbE), Hemoglobin C (HbC) and Hemoglobin D (HbD). HbS is autosomal dominant that is mainly found in Africans and is rare in Chinese. This mutation causes the sickle cell disease [3]. In HbS the *HBB* point mutation GAG→GTG mutates the 6th glutamate into valine and forms an abnormal β -globin [4]. HbE is $\alpha_2\beta_2$ E26K (GAG→AAG), i.e., E26 is mutated into lysine in the β -globin. The mutation site locates at the boundary of exon 1 and intron 1, affects the efficiency of the pre-mRNA splicing and leads to a reduction in the expression of the β chain. Consequently, HbE is a β -thalassemia-like abnormal hemoglobinopathy [5]. HbE often combines hybrid β -thalassemia with clinical and homological traits that are similar to severe

[†]These authors contributed equally to this work.

*Corresponding authors (email: liucq@ynu.edu.cn; caohuai@ynu.edu.cn)

β -thalassemia, such as severe hemolytic anemia [6]. HbC is autosomal incompletely dominant, the majority of carriers are African and the global incidence is No. 3 (after HbS and HbE) [7]. The point mutation GAG $\rightarrow$ AAG changes the 6th glutamate to a lysine and this mutation forms abnormal β -globin. The variant was first detected in an African-American; however, this mutation is mainly found in the West African tropical region and the incidence is as high as 14% to 28% [8]. A few cases of HbC were also reported in Italy, Netherlands, England and Thailand [9]. HbD comprises several variants, the letter “D” represents any of the HbS-like basic-pH electrophoretic variants with normal solubility. HbD Punjab is $\alpha_2\beta_2$ E121Q (GAA $\rightarrow$ CAA), i.e., the 121st glutamate is mutated into glutamine in β -globin. HbD was first detected by Itano in 1951 [10] and is mainly found in India, Pakistan and Iran with incidence of 3% [11]. In China, HbD is mainly distributed in the Northern region and has been reported in Han, Meng, Hui, Uyghur and Qazaq Chinese from Inner Mongolia, Xinjiang, Beijing, Henan, Gansu, Sichuan and Heilongjiang [12].

Although non-synonymous single nucleotide polymorphisms (nsSNPs) mutate the encoded amino acids and affect protein function [13,14], structure-function studies on the above four hemoglobinopathies have primarily focused on changes to the protein secondary structure caused by the variations in the amino acid sequence [15], and only a few reports have examined the changes in the mRNA secondary structures caused by the *HBB* gene mutations. According to a recent comprehensive review [16], the imbalance between the number of protein sequences and the much slower pace in determining their biological attributes has called for the development of more computational methods. The nucleic acid sequences and protein sequences of the *HBB* gene and its four variants are known. Remarkably, both mRNA and protein contain regular helical conformations, the spatial structures of the mRNA and protein are theoretically inter-related based on structural information. A simulated computational approach on their structures may help in finding this relationship. In this and previous work in our laboratory, we have determined that certain “deterministic” relationships may exist between the structure of the mRNA (i.e., the “frame structure” in dynamic homeostasis) and the protein structure (a type of “frame structure”) during the genetic information flow from mRNA to protein [17–19]. Many studies [20–25] have shown that sequence-based simulation approaches can generate useful data. These data, combined with structural bioinformatics [26], can provide useful insights for both basic research and drug development. In view of this, the present study investigated the relationships between mRNA-protein secondary structures in the human β -globin gene *HBB* and the four variants. We believe that our findings may be of use for basic research and drug development.

1 Materials and methods

The human hemoglobin pre-mRNAs span between 795 and 1800 nt, with three coding regions (cd1, cd2 and cd3) and four non-coding regions (5' UTR, in1, in2 and the 3' UTR). The split structure is represented as 5' UTR-cd1-in1-cd2-in2-cd3-3' UTR. The *HBB* gene sequence was obtained from the National Center for Biotechnology Information (NCBI) with the nucleotide database accession number of 3043. The sequence full-length is 1606 nt and the mature mRNA length is 626 nt. In the four variants, the mutation sites of *HbS*, *HbE* and *HbC* are located within the cd1 coding region, whereas the mutation site of *HbD* is located within the cd3 coding region.

Our laboratory has established and verified the “dynamic extended folding simulation” method, based on the fact that *in vivo* mRNAs change with co-transcriptional folding, splicing, processing and editing before maturation [27–29]. An mRNA sample was increased with the same step length. The first folding unit started at the first 30 nt, then the unit was increased step by step until the full-length was reached (such as 1–30 nt, 1–60 nt, 1–90 nt, ..., 1–full-length nt). The RNAstructure 4.4 program was used to predict the secondary structure of each folding unit [30]. The dynamic extended folding process was clearly represented with letter strings, which was also convenient for measuring the statistics of the frequencies of “hairpins” and “chain conformations” within the secondary structures. The prediction result of the last folding unit was the full-length pre-mRNA secondary structure. Based on the dynamic extended folding pre-mRNA structures, removing the introns, linking the exons, reserving the relatively conserved hairpins, and re-folding the fluctuating hairpins and the intron pairing single chains enabled us to obtain the mature mRNA secondary structure, which should closely represent the true secondary structure.

We used RNAstudio for drawing the RNA secondary structures [31], and for marking the types of hairpins and chain conformations. The protein secondary structures corresponding to *HBB*, *HbS*, *HbE*, *HbC* and *HbD* were obtained from the PDBsum database (the PDB IDs were 2HHB, 1HBS, 3DUT, 1K1K and 2YRS). The protein secondary structure regions were marked at the corresponding regions of the mature mRNA secondary structures. The mRNA-protein secondary structure maps were used to compare the normal and mutated genes, as well as to compare the secondary structures of the mRNA and protein.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practical more useful models, simulated methods, or predictors [32], we shall make efforts in our future work to provide a web-server for the dynamic extended folding simulation method used in this paper.

2 Results and discussion

2.1 The framework structures of the native and the mutated *HBB* mRNAs

The initial 90 nt symbol string representation of the full-

length folding secondary structure of *HBB* pre-mRNA is shown in Figure 1. The full-length 1606 nt string of the five samples is available in Figure S1. The hairpin distributions are listed in Table 1.

Figure 2 shows the full sequence folding secondary

Figure 1 The initial 90 nt symbol string representation of the full-length folding secondary structure of *HBB* pre-mRNA. The first line is the number of the nucleotides. The second line is the nucleotide sequence whose UTR and exon are noted by yellow and red, respectively. In the following lines, 30, 60, 90, ..., 1606 show every EFU (elongated folding unit) in its symbol string of folding secondary structure. In the symbol strings, L denotes terminal loop, S denotes single strand and J denotes multi-branched loop. The nucleotides are defined as <, >, I, i, B and b for 5' to 3', helical stems, inner loops and bulge loops, respectively. For example, <<<<LLLLLL>>>> is a hairpin. The last lines are nucleotide sequence and numbers, the frequency of the segment conformation, the hairpin number and its frequency. Hairpins less than 5 EFUs are marked *, such as 66* *HBB*.

Table 1 The hairpin statistics of the pre-mRNA dynamic extended folding of *HBB* and the variants^{a)}

	5' UTR	cd1	in1	cd2	in2	cd3	3' UTR	Total
Length	50	92	130	223	850	129	132	1606
<i>HBB</i>								66
over-CH	1	0	1	0	5	1	0	8
CH	0	1	0	0	2	0	0	3
sub-CH	0	1	1	3	7	0	0	12
fluc-CH	1	4	5	10	22	0	1	43
<i>HbS</i>								68
over-CH	1	0	1	0	5	1	0	8
CH	0	1	0	0	2	0	0	3
sub-CH	0	0	1	3	8	0	0	12
fluc-CH	1	4	6	9	23	0	2	45
<i>HbE</i>								65
over-CH	1	0	1	0	5	1	0	8
CH	0	0	0	0	2	0	0	2
sub-CH	0	2	1	3	9	0	0	15
fluc-CH	1	4	6	5	23	0	1	40
<i>HbC</i>								68
over-CH	1	0	1	0	5	1	0	8
CH	0	1	0	0	2	0	0	3
sub-CH	0	2	1	3	6	0	0	12
fluc-CH	1	3	6	11	23	0	1	45
<i>HbD</i>								66
over-CH	1	0	1	0	5	1	0	8
CH	0	1	0	0	2	0	0	3
sub-CH	0	1	1	3	7	0	0	12
fluc-CH	1	4	5	10	22	0	1	43

a) Over-CH, over-conserved hairpin; CH, conserved hairpin; sub-CH, sub-conserved hairpin; fluc-CH, fluctuating hairpin. Black-gray indicates a changed hairpin (number, and/or position, frequency) in *HbS*, *HbE* and *HbC*. *HbD* is identical to *HBB* with respect to hairpin position, frequency and number.

structure of *HBB* pre-mRNA. It also represents the probable hairpin and chain conformation structures at the end of RNA synthesis. In the figure, the three coding regions and the 5'-terminal hairpins are marked with conservations, which can be used to estimate the probability of the presence of a hairpin during the real co-transcriptional folding and pre-mRNA splicing into mature mRNA processes. The distributional statistics of the conservations of hairpins and chain conformations is the basis of frame structure construction. Since the two introns will be removed during the splicing process, the intron conformational types were ignored, but the hairpin numbers and conserved regions were marked. The conserved hairpins and chain conformations are stable structures under the dynamic extended folding model, which simulates the stable structures during the process of eukaryotic mRNA synthesis.

The *HBB* pre-mRNA folding secondary structure contained 66 hairpins, among which were eight highly con-

served, three conserved, 12 sub-conserved, and 43 fluctuating hairpins. There were 43 hairpins in the intron regions. During the splicing, hairpin 1, located completely within the 5'-UTR, is highly conserved, so it should remain after splicing. Hairpin 4 links the multi-branched loop 1, and the 0.50 frequency was the critical point of sub-conserved and fluctuating. During the dynamic extended folding, hairpin 3 also appeared in the same region with a frequency of 0.66. Since hairpins 3 and 4 overlapped, they were taken as one folding unit in the spliced mRNA structure. Hairpin 7 had a conserved frequency of 0.72 and was located completely in exon 1, thus its structure was stable and should be present after splicing. The 5'-terminal end of exon 2 formed a stem helix with nucleotides located in intron 1. After splicing, this region became a free single chain and linked to the 3'-terminal end of exon 1. Such an event should lead to re-folding. The RNAstructure 4.4 prediction result of this part showed a new local structure with three new hairpins 67, 68

Figure 2 The secondary structures of *HBB* pre-mRNA. The dynamic extended folding simulation of the last folding unit of *HBB* pre-mRNA calculated using RNAstructure 4.4 and drawn using RNAsudio. The point-highlighted terminal loops show the over-conserved (purple), conserved (blue), sub-conserved (green) hairpins and some of the fluctuating hairpins (brown) in the cd1, cd2 and cd3 coding regions. The conservations of the chain conformations are marked in base colors and some in frequencies.

and 69, marked in pink (Figure 1). Hairpin 23 was centrally located in exon 2 with a relatively conserved frequency of 0.59. It should be maintained during and follow the splicing event. After intron 2 was spliced and the 5'-terminal of exon 3 was linked to the 3'-terminal of exon 2, new base pairs were formed. Hairpin 28 refolded and the stem pairs came

from the 5'-terminal end of exon 3 and the 3'-terminal end of exon 2; however, the terminal loop was entirely from exon 2 with a frequency of 0.61. Hairpin 65 is a unique, highly conserved hairpin (the frequency was 1) in the exon regions. It is located completely in exon 3, so it should be kept following splicing. A large part of exon 3 paired with

intron 2. After intron 2 removal, this part was observed to refold into a new structure with the formation of hairpins 70 and 71. The 3'-UTR also refolded, with hairpin 66 being unaffected following splicing; however, the formation of new hairpins 72, 73 and 74 was observed. Hairpin 66 emerged only in three end folding units, so it was noted as

66*. The above processes of splicing, re-pairing and refolding led to the secondary structure of *HBB* mature mRNA (Figure 3).

Using the same approach as the *HBB* mature mRNA construction, we also obtained the mature mRNA secondary structures of the four variants. The protein secondary structures

Figure 3 The secondary structures of *HBB* mature mRNA (marked as in Figure 2).

were marked on the five maps of the mature mRNA secondary structures, as shown in Figure S2.

2.2 Comparison of the mature mRNA secondary structures of *HBB* and its four variants

HbS has a single nucleotide mutation, GAG→GTG, at the 6th codon of the *HBB* gene. The mutation site is marked with a blue arrow in Figure S2(b). Compared with Figure 3, there were changes in the hairpins and chain conformations in the mature mRNA secondary structure of the variant *HbS*. Strikingly, hairpins 3 and 4 (frequencies of 0.66 and 0.50, respectively) of normal *HBB* became fluctuating hairpins with lower stabilities in the variant, or, hairpins 3 and 4 in the normal *HBB* disappeared and became a non-structured single chain region. Consequently, when the *HbS* mature mRNA secondary structure was constructed, this part refolded and formed the new hairpin 69. The frequency of hairpin 6 changed from 0.82 to 0.88. The frequency of hairpin 27 changed from 0.61 to 0.71. The two hairpins became more conserved following the mutation. The hairpin 73 in *HbS* was a new hairpin after refolding.

The HbS disease is also called sickle cell anemia because the protein structure changes the red cell from a flat round pie shape into a sickle shape. We know that the one-dimensional nucleotide sequence determines the protein amino acid sequence and then affects the protein high-level structure. The above analysis indicated that an important reason for the formation of a HbS abnormal structure would be the changes to the mature mRNA secondary structure.

HbE is the GAG→AAG mutation at the 26th codon of the *HBB* gene (Figure S2(c)). Compared to Figure 3, there were changes in the mature mRNA secondary structure of the variant *HbE*. Hairpin 3 changed frequency from 0.66 to 0.58 and the stability decreased. Hairpin 4 changed frequency from 0.50 to 0.27 and did not remain. Hairpin 7 changed frequency from conserved (0.82) to sub-conserved (0.71). The hairpins 22 and 24 (frequencies 0.56 and 0.55, respectively) in the variant corresponded to hairpins 23 and 28 at same positions in the native sequence (frequencies 0.59 and 0.61, respectively) with only small changes. Hairpin 64 and the chain conformation were highly conserved before and after the mutation. Hairpin 65 was located within the 3'-end noncoding region and noted as 65*. The hairpins 66 to 73 were refolded hairpins.

HbC is the mutation at the 6th codon as observed for *HbS*, but the GAG becomes AAG (Figure S2(d)). Compared with Figure 3, the mature mRNA secondary structure of the variant *HbC* was similar to native with only minor changes in the hairpins and chain conformations. An obvious change was that hairpin 4 changed from a fluctuating frequency of 0.50 into a sub-conserved frequency of 0.77 with elevated stability. Minor changes were observed in hairpins 7, 24 and 30 within the same conserved regions. The frequency of hairpin 67 was still highly conserved at 1.00. The refolded

secondary structure units were numbered 69 to 76.

HbD is the mutation at the 121st codon GAA→CAA (Figure S2(e)). Comparing with Figure 3, the mature mRNA secondary structure of the variant *HbD* is the same as that of the normal with same frequencies in hairpins and chain conformations. The *HbD* mutation site was located at a cd3 highly conserved hairpin region. This secondary structure unit was highly conserved, keeping frequency of 1.00 in the normal and the four variant pre-mRNAs. It could be postulated that mutations within this region can hardly affect the secondary structure.

2.3 Comparison of the protein secondary structures and the mature mRNA secondary structures

HbS, *HbE*, *HbC* and *HbD* are all products of a single nucleotide mutation of the *HBB* gene. They encode similar globin sequences and secondary structures, but differences are present. The structural similarities and differences during the process from pre-mRNA to mature mRNA and subsequently to protein showed that the protein structure is not simply determined by the one-dimensional sequence; it is also affected by the spatial structural information. Table 2 presents the counts of the secondary structure units in the five proteins.

As shown in Figure S2(a), hairpins 4 and 7 of mature *HBB* mRNA correlate to the protein H1 helix, hairpin 67 to the H2 helix, hairpin 69 to the H3 helix, hairpin 23 to the H4 helix, hairpins 28 and 65 to the H6 helix and hairpins 70 and 71 to the H7 helix. The H5 helix was not encoded by a hairpin but by a highly conserved paired base helix region. The G1, G2, G3 and G4 helices are mainly correlated to *HBB* secondary structure regions of multi-branched loops, inner loops or hairpin roots. Analysis from the map marking both *HBB* secondary structures of the encoded protein and the mature mRNA indicated the correlations.

Analyses of the figures in Figure S2 and Table 2 revealed that the major differences between the normal and the varied proteins HbS, HbE, HbC and HbD were in the helical conformations and the β -turns. In comparison to the native protein, HbS changed in the first helix, with the starting lagged three amino acids and the ending lagged one amino acid. Comparing the two mature mRNAs (Figure S2(a) and (b)), hairpin 4 has a lower frequency in *HbS*, which correlates with the absence of the three amino acids in the first helix. The three 3_{10} G-helices 4, 7 and 10 in the native protein were replaced by nine β -turns in the same region of HbS. The H8 in the native protein was divided into two helices, H6 and H7, in HbS. Comparison of HbE and the native protein showed that the two interrupt 3_{10} G-helices in the native protein became a continuous helical structure and a new β -turn in the HbE structure. The G-helix 7 was replaced by two β -turns. The amino acid 1–3 (Val1-Leu3) became a new γ -turn. HbC is composed of 11 helices, and this is the same number of helices found in the native protein.

Table 2 The statistics of protein secondary structure units^{a)}

Hemo-globin	Helix number	Helix type	Start amino acid	End amino acid	Count	β -turn	Start-end amino acid	γ -turn	Start-end amino acid
Normal	1	H	Pro5	Trp15	11	β 1	Ala142-Tyr145	γ 1	Glu121-Thr123
	2	H	Val20	Val34	15				
	3	G	Pro36	Phe41	6				
	4	G	Glu43	Phe45	3				
	5	H	Pro51	Gly56	6				
	6	H	Pro58	Leu75	18				
	7	G	Ala76	Asn80	5				
	8	H	Leu81	Asp94	14				
	9	H	Glu101	Phe118	18				
	10	G	Gly119	Glu121	3				
	11	H	Pro124	Ala142	19				
HbS	1	H	Lys8	Gly16	9	β 1	Phe42-Phe45	γ 1	Glu121-Thr123
	2	H	Val20	Val34	15	β 2	Glu43-Gly46		
	3	G	Thr38	Phe41	4	β 3	Leu48-Pro51		
	4	H	Pro51	Gly56	6	β 4	Leu75-Leu78		
	5	H	Pro58	Ser72	15	β 5	His77-Asn80		
	6	H	Leu81	Thr84	4	β 6	Leu78-Leu81		
	7	H	Ala86	Asp94	9	β 7	Phe118-Glu121		
	8	H	Glu101	Phe118	18	β 8	Gly119-Phe122		
	9	H	Pro124	Leu141	18	β 9	Lys120-Thr123		
						β 10	Ala142-Tyr145		
HbE	1	H	Pro5	Trp15	11				
	2	H	Val20	Val34	15	β 1	Phe45-Leu48	γ 1	Val1-Leu3
	3	G	Pro36	Phe45	10	β 2	His77-Asn80	γ 2	Glu121-Thr123
	4	H	Pro51	Gly56	6	β 3	Leu78-Leu81		
	5	H	Pro58	Leu75	18	β 4	Ala142-Tyr145		
	6	H	Leu81	Asp94	14				
	7	H	Glu101	Phe118	18				
	8	G	Gly119	Glu121	3				
	9	H	Pro124	Ala142	19				
HbC	1	H	Pro5	Gly16	12	β 1	Asn19-Glu22	γ 1	Glu121-Thr123
	2	H	Val23	Val34	12	β 2	Val20-Val23		
	3	G	Pro36	Phe41	6	β 3	Ala142-Tyr145		
	4	G	Glu43	Phe45	3	β 4	His143-His146		
	5	H	Pro51	Met55	5				
	6	H	Pro58	Ala76	19				
	7	G	Leu78	Asn80	3				
	8	H	Leu81	Asp94	14				
	9	H	Glu101	Phe118	18				
	10	G	Gly119	Glu121	3				
	11	H	Pro124	Ala142	19				
HbD	1	H	Pro5	Trp15	11	β 1	Phe42-Phe45	γ 1	Gln121-Thr123
	2	H	Val20	Val34	15	β 2	Glu43-Gly46		
	3	G	Pro36	Phe41	6	β 3	His77-Asn80		
	4	H	Pro51	Gly56	6	β 4	Leu78-Leu81		
	5	H	Pro58	Leu75	18	β 5	Asp94-His97		
	6	H	Leu81	Thr84	4	β 6	Ala142-Tyr145		
	7	H	Ala86	Asp94	9				
	8	H	Glu101	Phe118	18				
	9	G	Gly119	Gln121	3				
	10	H	Pro124	Ala142	19				

a) H, α -helix; G, 3_{10} helix. The amino acids shown in shade represent regions differing from the native sequence.

Moreover, the types and positions of the helices are also basically identical, with minor amino acid differences at the starts and the ends of the helices. HbC helix 2 lagged by three amino acids at the start, which were replaced by $\beta 2$. The $\beta 3$ (Ala142-Tyr145) was unchanged before and after the mutation. The two β -turns, $\beta 1$ and $\beta 4$, had formed after the mutation. Both protein structures of HbC and the native protein matched each other with their mRNA secondary structures. The two mature mRNA secondary structures were unchanged in the frame structure; however, a few hairpins changed their conservations. Hairpins 3 and 7 of mature HbC mRNA are correlated to the protein helix 1. The two hairpins in *HBB* were conserved and corresponded to the native protein helix 1 with a single amino acid increase.

Among the four variants of the *HBB* gene, the protein secondary structure changes of HbS, HbE and HbC could be traced from the changes found in their mRNA secondary structures; however, HbD did not present any mRNA structural changes. The changes in the HbD protein included the loss of the two 3_{10} G-helices 4 and 7 that were replaced by four β -turns. The mutation site in the mRNA was located in the highly conserved region. The changes to the protein secondary structures may also depend on amino acid preferences.

3 Conclusions

The mRNA frame structures of the *HBB* gene and its four variants were constructed and compared with their encoding proteins. We found that the protein secondary structures showed good correlations to the mRNA secondary structures, thereby providing a link between protein structure and mRNA conformations. Generally, long helical conformations in the protein are correlated to the relatively stable conserved hairpins and chain conformations in the mRNA frame structure, especially the terminal loops of highly conserved hairpins. The protein β - and γ -turns are correlated to less conserved regions of the mRNA frame structure, including major multi-branched loops, inner loops or fluctuating helical pair regions. When mutations were located in unstable regions, corresponding protein secondary structures such as 3_{10} G-helices were found to transform to β -turns. These observations suggest that regular protein conformations, i.e., the relatively stable helical structures, correlate to the conserved mRNA conformations. In contrast, regions of the protein with less stable structural elements, i.e., β - and γ -turns or irregular coil structures, correlate to the less-conserved sequence regions of the mRNA.

This work was supported by the National Natural Science Foundation of China (30971454 and 90208018).

1 Old J. Haemoglobinopathies and thalassaemias. In: Rimoin D L,

- Connor J M, Pyeritz R E, eds. Emery and Rimoin's Principles and Practice of Medical Genetics. 5th ed. Churchill Livingstone: Elsevier, 2007. 1638–1674
- 2 Wiwanitkit V. Analysis of functional aberration of some important beta hemoglobinopathies (hemoglobin C, D, E and S) from nanostructures. *Nanomedicine*, 2005, 1: 213–215
- 3 Pauling L, Itano H, Singer S J, et al. Sickle cell anemia, a molecular disease. *Science*, 1949, 110: 543–548
- 4 Schnog J B, Duits A J, Muskiet F A, et al. Sickle cell disease: A general overview. *Neth J Med*, 2004, 62: 364–374
- 5 Orkin S H, Kazazian Jr H H, Antonarakis S E, et al. Abnormal RNA processing due to exon mutation of the β^E -globin gene. *Nature*, 1982, 300: 768–769
- 6 Rees D C, Styles L, Vichinsky E P, et al. The hemoglobin E syndromes. *Ann N Y Acad Sci*, 1998, 850: 334–343
- 7 Schwab J G, Abelson H T, Fairhurst R M, et al. Hemoglobin C disease. *New Engl J Med*, 2004, 351: 1577
- 8 Kwiatkowski D P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*, 2005, 77: 171–192
- 9 Modiano D, Luoni G, Sirima B S, et al. Hemoglobin C protects against clinical plasmodium falciparum malaria. *Nature*, 2001, 414: 305–308
- 10 Itano H A. Third abnormal hemoglobin associated with hereditary hemolytic anemia. *Proc Natl Acad Sci USA*, 1951, 37: 775–784
- 11 Lehmann H, Huntaman R G. Man's Haemoglobins. 2nd ed. Amsterdam: North-Holland, 1974. 310
- 12 Yu W Z, Li H J, Li L, et al. Characteristic on genetic epidemiology of Chinese silk road region. *Acta Anthropol Sin*, 2001, 20: 69–75
- 13 Chasman D, Adams R M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. *J Mol Biol*, 2001, 307: 683–706
- 14 Huang T, Wang P, Ye Z Q, et al. Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLoS ONE*, 2010, 5: e11900
- 15 George C, Rao S. Impact of single nucleotide polymorphisms in HBB gene causing haemoglobinopathies: *In silico* analysis. *New Biotechnol*, 2009, 25: 214–219
- 16 Chou K C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J Theor Biol*, 2011, 273: 236–247
- 17 Zhang J, Liu C Q. Is there are genetics relationship between mRNA and encoded protein at three-dimensional level? *Prog Nat Sci*, 1999, 9: 75–77
- 18 Liu C Q, Liu S Q. mRNA may carry spatial genetic information—The topological configuration of mRNA affect the initial structure of nascent polypeptides. *Amer J Human Genet*, 2001, 69(Suppl.1): 362
- 19 Liu C Q. The correlation between the chain conformations of unfolded mrna and corresponding nascent polypeptid. *Chin J Nat*, 2006, 28: 31–36
- 20 He Z, Zhang J, Shi X H, et al. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, 2010, 5: e9603
- 21 Huang T, Niu S, Xu Z, et al. Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. *PLoS ONE*, 2011, 6: e22940
- 22 Huang T, Shi X. H, Wang P, et al. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. *PLoS ONE*, 2010, 5: e10972
- 23 Wang P, Xiao X, Chou K C. NR-2L: A two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS ONE*, 2011, 6: e23505
- 24 Chou K C, Shen H B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem Biophys Res Comm*, 2008, 376: 321–325
- 25 Wang P, Hu L, Liu G, et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS ONE*, 2011, 6: e18476

- 26 Chou K C. Review: Structural bioinformatics and its impact to biomedical science. *Curr Med Chem*, 2004, 11: 2105–2134
- 27 Zhang S X, Wang X, Chen X F, et al. Secondary structural analysis of the mRNA regions encoding the hemagglutinin cleavage site basic amino acids of the avian influenza virus H5N1 subtype samples. *Chin Sci Bull*, 2008, 53: 377–383
- 28 Cao H, Xie H Z, Zhang W, et al. Dynamic extended folding: Modeling the RNA secondary structures during co-transcriptional folding. *J Theor Biol*, 2009, 261: 93–99
- 29 Zhang X, Guo C, Zhang W, et al. A folding “framework structure” of tetrahymena group I intron. *J Theor Biol*, 2010, 267: 495–501
- 30 Mathews D H, Disney M D, Childs J L, et al. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA*, 2004, 101: 287–292
- 31 Zhang H, Liu C Q. RNASTudio, a full-featured object-oriented program for visualizing RNA secondary structure. *J Mol Graphics Model*, 2002, 21: 1–2
- 32 Chou K C, Shen H B. Review: Recent advances in developing web-servers for predicting protein attributes. *Nat Sci*, 2009, 2: 63–92

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Supporting Information

Figure S1 The symbol string representations of the pre-mRNA dynamic extended folding of *HBB* and the variants.

Figure S2 The secondary structures of the mature mRNAs marked the coded protein structures. *HBB* (a) and its four variants *HbS* (b), *HbE* (c), *HbC* (d) and *HbD* (e).

The supporting information is available online at csb.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.