Mutation Pattern of Amino-Acid Pairs in Human Hemoglobin β-Chain[*]

人血红蛋白β链氨基酸对的变异模式

YAN Shao-min[1], WU Guang[1,2**]
严少敏[1],吴　光[1,2**]

(1. State Key Laboratory of Non-food Biomass Enzyme Technology, National Engineering Research Center for Non-food Biorefinery, Guangxi Key Laboratory of Biorefinery, Guangxi Academy of Sciences, Nanning, Guangxi, 530007, China; 2. DreamSciTech Consulting, Shenzhen, Guangdong, 518054, China)
(1.广西科学院非粮生物质酶解国家重点实验室、国家非粮生物质能源工程技术研究中心、广西生物炼制重点实验室,广西南宁　530007; 2.深圳市追梦科技咨询有限公司,广东深圳 518054)

Abstract: 244 mutations in human hemoglobin β-chain were analyzed in terms of amino-acid pair predictability in order to determine which amino-acid pairs were more sensitive to mutations. The results revealed four mutation patterns, characterized by (1) 85.66% of mutations occurred at unpredictable amino-acid pairs, (2) the vast majority of substituted amino-acid pairs had one or both pairs whose actual frequency larger than predicted one, (3) 79.51% of mutations generated one or both substituting amino-acid pairs that were absent in the normal human-hemoglobin and (4) most mutations narrowed the difference between actual and predicted frequencies in amino-acid pairs. Thus, the mutations generally led the human hemoglobin β-chain variants to be constructed more randomly and stably.

Key words: amino-acid pair predictability, hemoglobin β-chain, mutation, pattern

摘要:根据氨基酸对可预测性分析人血红蛋白β链的244个变异,以确定哪些氨基酸对对变异更敏感。结果发现4种变异模式:(1)85.66%的变异发生在不可预测的氨基酸对,(2)绝大多数被替换掉的氨基酸对中包含一个或两个氨基酸对它们的实际频率大于预测频率,(3)79.51%的变异中被替换出的氨基酸对包含一个或两个在正常人β血红蛋白中不存在的氨基酸对,(4)大多数变异缩小了氨基酸对实际频率与预测频率之差。说明变异通常导致人血红蛋白β链的构成更随机、更稳定。

关键词:氨基酸对可预测性　血红蛋白β链　变异　模式
中图法分类号:R556.6$^+$1　**文献标识码**:A　**文章编号**:1002-7378(2011)01-0029-05

The human hemoglobin (Hb) is subject to variants and mutations, which lead to various disorders[1~4]. The root cause of sickle cell disease, the first molecular disease, is a single β-globin gene mutation coding for the sickle β-hemoglobin chain. Sickle hemoglobin tetramers polymerize when deoxygenated, damaging the sickle erythrocyte[5]. Thus an important question raised is whether the mutation has any pattern to follow. If so, the humans may take some measures to prevent mutations from occurring.

A useful way to determine the mutation pattern is to find the so-called "hotspot" sites in a protein, where the amino acids are subject to endogenous and exogenous mutagens[6~8]. This approach provides us the mutation pattern with respect to the physicochemical property of amino

收稿日期:2010-10-18

作者简介:严少敏(1958-),女,博士,研究员,主要从事定量诊断病理学和计算变异学研究。

[*] This study was partly supported by Guangxi Science Foundation (10-046-06, 2010GXNSFA013046 and 2010GXNSFF013003 and Guangxi Academy of Sciences (08 YJ6SW06).

[**] 通讯作者。

acid in the hotspot. On the other hand, it is also important and interesting to find the mutation pattern with respect to the numeric property of amino acid in a protein.

The public-accessible database gives us the possibility to find out the mutation patterns with respect to the numeric property of amino acid in a protein. In this study, 244 mutations in human hemoglobin β-chain were analyzed in terms of amino-acid pair predictability in order to determine their mutation pattern with respect to the numeric property of amino-acid pairs.

1 Materials and methods

1.1 Data

The amino-acid sequence of the human hemoglobin β-chain and its 244 mutations are obtained from the UniProtKB/Swiss-Prot (access number P02023)[9].

1.2 Numeric property of amino-acid pair predictability

The amino-acid pair predictability[10~13] is used to represent the numeric property of amino-acid pairs along the human hemoglobin β-chain. For example, there are 15 alanines (A) and 18 valines (V) in the human hemoglobin β-chain: if the permutation can predict the frequency of appearance of amino-acid pair AV, then it would appear twice ($15/146 \times 18/145 \times 145 = 1.849$); actually there are indeed two AVs in the human hemoglobin β-chain, so the frequency of appearance of AV is predictable.

Similarly, there are 9 histidines (H) in human hemoglobin β-chain: the frequency of amino-acid pair AH would be one ($15/146 \times 9/145 \times 145 = 0.925$), while AH appears four times, so the frequency of appearance of AH is randomly unpredictable. This is the case that the actual frequency is larger than the predicted frequency. There are also the cases where the actual frequency is smaller than the predicted frequency, for instance, there are 13 glycines (G) in human hemoglobin β-chain and the predicted frequency of amino-acid pair VG is 2 ($18/146 \times 13/145 \times 145 = 1.603$), but its actual frequency is only one.

Actually, the permutation can apply to absent amino-acid pairs. For example, there are two cysteines (C) in human hemoglobin β-chain, and the amino-acid pair AC would not appear ($15/146 \times 2/145 \times 145 = 0.205$), which is true. Thus, this absence is randomly predictable. Also, there are 11 lysines (K) in human hemoglobin β-chain, and the amino-acid pair AK would appear once ($15/146 \times 11/145 \times 145 = 1.130$), however, AK does not exist in human hemoglobin β-chain, therefore its absence is randomly unpredictable.

1.3 Mutations in terms of predictable and unpredictable amino-acid pairs

Although a point mutation mutates a single amino acid, it changes the frequencies of two adjacent amino-acid pairs involved with the mutation, except for the mutation at the terminal of hemoglobin β-chain. For instance, a mutation at position 17 substitutes lycine (K) for glutamine (Q), which results in the amino-acid pairs GK and KV changed to GQ and QV, because the amino acid is G at position 16 and V at position 18. The actual frequency (AF) and predicted frequency (PF) are 3 and 1 for GK, 4 and 1 for KV, 1 and 2 for GQ, and 0 and 0 for QV, respectively. Thus, the sum of difference between actual frequency and predicted frequency $\Sigma(AF-PF)$ is 5, $(3-1)+(4-1)=5$, before the mutation, but -1, $(1-2)+(0-0)=-1$, after mutation.

1.4 Statistics

The Mann-Whitney U - test is used for comparison and $P < 0.05$ is considered statistically significant.

2 Results

The human hemoglobin β-chain consists of 146 amino acids, of which isoleucine (I) does not exist and there is only one methionine (M), therefore these 146 amino acids can theoretically construct 342 types of amino-acid pairs (19×18) and 145 adjacent amino-acid pairs. Of 342 theoretical types of amino-acid pairs, 229 types are absent including 193 predictable and 36 unpredictable. Thus, 145 adjacent pairs include only 113 theoretical types (342 − 229 = 113), so some pairs should appear more than once: 91 theoretical types once, 14 types twice, 6 types three times, and 2 types four times.

Consequently, these 113 present types include 46 predictable and 67 unpredictable, which correspond 51 predictable and 94 unpredictable among 145 adjacent pairs. Table 1 splits these data with respect to mutations. This is the first mutation pattern, that is, the mutation is far more likely to occur at unpredictable amino-acid pairs.

Table 1 Mutations and amino-acid pair predictability in human hemoglobin β-chain

Amino-acid pairs	Types No. (%)	Pairs No. (%)	Mutations No. (%)	Mutations /Types	Mutations /Pairs
Predictable	46 (40.71)	51 (35.17)	35 (14.34)	35/46 =0.76	35/51 =0.69
Unpredictable	67 (59.29)	94 (64.83)	209 (85.66)	209/67 =3.12	209/94 =2.22
Total	113 (100)	145 (100)	244 (100)	244/113 =2.16	244/145 =1.68

Table 2 shows the second mutation pattern with respect to the substituted amino-acid pairs, where the mutations occur. This table can be read as follows. The first column classifies the amino-acid pairs as predictable and unpredictable. The second and third columns show the amino-acid pairs in relation to their actual and predicted frequencies, for example, the first two cells in columns 2 and 3 indicate that the actual frequencies are equal to the predicated frequencies in both pairs Ⅰ and Ⅱ. The fourth and fifth columns indicate how many mutations occur in pairs Ⅰ and Ⅱ, for instance, 34 of 244 variants occur at the pairs whose actual frequencies are equal to their predicted ones (13.93%). The sixth column indicates the percent of 244 mutations occurring at predictable and unpredictable pairs.

Actually, Table 2 is the furthermore elaboration of Table 1, so both show that 85.66% of mutations occur at unpredictable pairs and 14.34% of mutations occur at predictable pairs. These results mean that 67 types of unpredictable pairs account for 85.66% mutations, whereas 46 types of predictable pairs account only for 14.34% mutations. Still we can see the ratio in Table 1 that the chance of occurring of mutations in unpredictable pairs is about 3.1-fold larger than in predictable pairs.

Moreover, the characteristic that one or both pairs whose actual frequency larger than predicted frequency governs the vast majority of the unpredictable pairs (the first four rows in unpredictable pairs in Table 2). The mutations narrow the difference between actual and predicted frequencies by means of reducing the actual frequency. This implies that the mutations lead to the construction of amino-acid pairs to be more randomly predictable. In other words, the mutations result in the construction of amino-acid pairs more easily naturally to occur. Yet, no mutations occur in the pairs whose actual frequency is smaller than their predicted one in both pairs, which suggests the difficulty for mutation to narrow the difference between actual and predicted frequencies by means of increasing the actual frequency.

Table 3 shows the third mutation pattern with respect to the substituting amino-acid pairs after mutation. This table can be read as follows. The first and second columns indicate the actual and predicted situations in pairs Ⅰ and Ⅱ, the third and fourth columns indicate the number of mutations occurs at pairs Ⅰ and Ⅱ and their percents, the fifth column show the total classifications. 79.51% of mutations lead one or both pairs that are absent in the original human β-hemoglobin (AF = 0 in Table 3). Furthermore, 31.97% of mutations generate one or both substituting pairs whose actual frequency is smaller than their predicted one (*).

Table 2 Substituted amino-acid pairs during mutation in human hemoglobin β-chain

Predictability	Amino-Acid Pair		Mutations		Total
	Ⅰ	Ⅱ	number	(%)	(%)
Predictable	AF=PF	AF=PF	34	13.93	14.34
	AF=PF	—	1	0.41	
Unpredictable	AF>PF	—	4	1.64	85.66
	AF>PF	AF>PF	102	41.80	
	AF>PF	AF=PF	92	37.71	
	AF>PF	AF<PF	2	0.82	
	AF<PF	AF=PF	9	3.69	
	AF<PF	AF<PF	0	0	

AF: actual frequency, PF: predicted frequency.

Table 3 Substituting amino-acid pairs after mutation in human hemoglobin β-chain

Amino-Acid Pair		Mutations		Total
I	II	Number	(%)	(%)
AF=0,PF>0	AF=0,PF>0	13*	5.33	79.51
AF=0,PF>0	AF=PF=0	25*	10.25	
AF=0,PF>0	AF=PF>0	21*	8.61	
AF=0,PF>0	AF<PF,AF≠0	1*	0.41	
AF=0,PF>0	AF>PF	12*	4.92	
AF=PF=0	AF=PF=0	70	28.69	
AF=PF=0	—	2	0.82	
AF=PF=0	AF=PF>0	22	9.02	
AF=PF=0	AF<PF,AF≠0	1*	0.41	
AF=PF=0	AF>PF	27	11.07	
AF<PF,AF≠0	AF<PF,AF≠0	0*	0	20.49
AF<PF,AF≠ 0	AF=PF>0	5*	2.05	
AF<PF,AF≠0	AF>PF	0*	0	
AF=PF>0	AF=PF>0	15	6.15	
AF>PF	AF>PF	7	2.87	
AF>PF	—	3	1.23	
AF=PF>0	AF>PF	20	8.20	

AF: actual frequency, PF: predicted frequency; * : indicates the amino acid pairs with their actual frequency smaller than predicted frequency and the total of these amino acid pairs are 78 (31.79%).

Figure 1 demonstrates the fourth mutation pattern related to the difference between actual and predicted frequencies, which actually represents as a measure of randomness of construction of pairs, i. e. the smaller the difference, the more random the construction of pairs. In particular, (i) the larger the positive difference, the more unpredictable pairs present; and (ii) the larger the negative difference, the more unpredictable pairs absent.

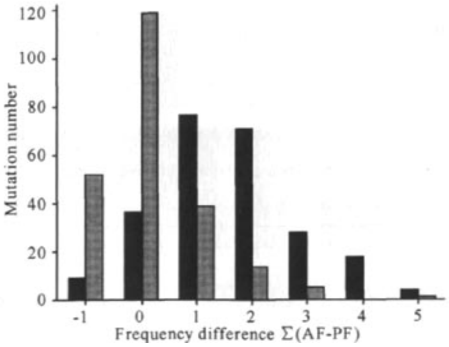

Fig. 1 Frequency difference between substituted (black bars) and substituting (gray bars) amino-acid pairs induced by mutations in human hemoglobin β-chain.

AF, actual frequency; PF, predicted frequency; significantly statistical difference is found between the substituted and substituting pairs ($P < 0.0001$).

Considering all 244 mutations, the mean ± SE is 1.58 ± 0.08 (ranging from −1 to 5) for the difference between actual and predicted frequencies in the substituted amino-acid pairs. This means that the mutations target the pairs, which appear more than their predicted frequency. Meanwhile, the mean ± SE is 0.03 ± 0.07 (ranging from −2 to 5) for the difference between actual and predicted frequencies in amino-acid pairs generated by mutations. This implies that the substituting amino-acid pairs are more randomly constructed in the variants of human hemoglobin β-chain, as their actual and predicted frequencies are about the same. Striking statistical difference is found between the substituted and substituting pairs ($P < 0.0001$).

3 Discussion

Numerous studies have been done in the transcription and post-transcription regulation of human hemoglobin β-chain gene, using it as a basis to explain the mechanisms by which human hemoglobin β-chain mutations downregulate expression causing a quantitative deficiency of human hemoglobin β-chain[14~16]. Also some human hemoglobin β-chain variant hemoglobinopathies are associated with human hemoglobin α-chain variants[17,18]. There is increasing evidence that the most common monogenetic conditions in humans have evolved under pressure from malaria[19]. That every population has a different set of thalassaemia mutations suggests that this selective force is fairly recent, otherwise the same mutations would appear throughout the tropical world[20].

The current study is in good agreement with our previous studies in human hemoglobins[21~24], which suggest that the mutation pattern can be found using the numeric property of amino acids in a protein, because this property reflects another aspect that the physicochemical property cannot replace. This would possibly be one of the reasons of why most hematological diseases are irreversible, because their hemoglobin structure is more probabilistically stable, and the occurrence of many hemoglobinopathies is a natural process, because the hemoglobin needs to be more probabilistically stable.

Methodologically the amino-acid pairs sensitive to mutations in a protein can be predicted using the random approach, thus it would be possible to firstly detect the unpredictable amino-acid pairs in human hemoglobin in order to screen the diseases efficiently. Meanwhile it would be possible to modify the unpredictable amino-acid pairs to prevent human hemoglobin from mutating.

References:

[1] Agarwal N, Gordeuk R V, Prchal J T. Genetic mechanisms underlying regulation of hemoglobin mass[J]. Adv Exp Med Biol, 2007, 618: 195-210.

[2] Vichinsky E. Hemoglobin e syndromes[J]. Hematology Am Soc Hematol Educ Program, 2007(1): 79-83.

[3] Panigrahi I, Marwaha R K, Kulkarni K. The expanding spectrum of thalassemia intermedia[J]. Hematology, 2009, 14: 311-314.

[4] Vichinsky E P. Alpha thalassemia major-new mutations, intrauterine management, and outcomes[J]. Hematology Am Soc Hematol Educ Program, 2009: 35-41.

[5] Steinberg M H. Sickle cell anemia, the first molecular disease: overview of molecular etiology, pathophysiology, and therapeutic approaches[J]. Scientific World J, 2008, 8: 1295-1324.

[6] Rideout W M, Coetzee G A, Olumi A F, et al. 5-Methylcytosine as an endogenous mutagen in human LL receptor and p53 genes[J]. Science, 1990, 249: 1288-1290.

[7] Montesano R, Hainaut P, Wild C P. Hepatocellular carcinoma: from gene to public health[J]. J Natl Cancer Inst, 1997, 89: 1844-1851.

[8] Hainaut P, Pfeifer G P. Patterns of p53 G T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke[J]. Carcinogenesis, 2001, 22: 367-374.

[9] The UniProt Consortium. The universal protein resource (UniProt) in 2010[J]. Nucleic Acids Res, 2010, 38: D142-D148.

[10] Wu G, Yan S. Randomness in the primary structure of protein: methods and implications[J]. Mol Biol Today, 2002, 3: 55-69.

[11] Wu G, Yan S. Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint[J]. Acta Pharmacol Sin, 2006, 27: 513-526.

[12] Wu G, Yan S. Lecture notes on computational mutation[M]. New York: Nova Science Publishers, 2008.

[13] Yan S, Wu G. Creation and application of computational mutation[J]. J Guangxi Acad Sci, 2010, 17(2): 145-150.

[14] Kutlar A. Sickle cell disease: a multigenic perspective of a single gene disorder[J]. Hemoglobin, 2007, 31: 209-24.

[15] Ho P J, Thein S L. Gene regulation and deregulation: a β globin perspective[J]. Blood Rev, 2000, 14: 78-93.

[16] Cao A, Galanello R. Beta-thalassemia[J]. Genet, Med, 2010, 12: 61-76.

[17] Krauss J S. The proportion of hybrid heterodimers in homozygous or doubly heterozygous beta chain variant haemoglobinopathies associated with alpha chain haemoglobin variants[J]. Ann Clin Lad Sci, 2000, 30: 391-394.

[18] Ataga K I, Cappellini M D, Rachmilewitz EA. Beta-thalassaemia and sickle cell anaemia as paradigms of hypercoagulability[J]. Br J Haematol, 2007, 139: 3-13.

[19] Denic S, Nicholls M G. Genetic benefits of consanguinity through selection of genotypes protective against malaria[J]. Hum Biol, 2007, 79: 145-58.

[20] Flint J, Harding R T, Boyce A J, et al. The population genetics of the haemoglobinipathies[J]. Baillieres Clin Haematol, 1998, 11: 1-51.

[21] Wu G. The first and second order Markov chain analysis on amino acids sequence of human haemoglobin β-chain and its three variants with low O_2 affinity[J]. Comp Haematol Int, 1999, 9: 148-151.

[22] Wu G, Yan S M. Prediction of two- and three-amino acid sequence of human acute myeloid leukemia 1 protein from its amino acid composition[J]. Comp Haematol Int, 2000, 10: 85-89.

[23] Wu G, Yan S M. Determination of amino acid pairs in human haemoglobulin β-chain sensitive to variants by means of a random approach[J]. Comp Clin Pathol, 2003, 12: 21-25.

[24] Gao N, Yan S, Wu G. Pattern of positions sensitive to mutations in human haemoglobin β-chain[J]. Protein Pept Lett, 2006, 13: 101-107.

（责任编辑：尹　闯）