EXCLI Journal 2004;3:10-19 – ISSN 1611-2156
received: 22. April 2004, accepted: 25. April 2004, published: 26. April 2004

Original article:

Amino Acid Pairs Sensitive to Variants in Human Collagen α 1(I) Chain Precursor

Guang Wu* and Shaomin Yan

DreamSciTech Consulting Co. Ltd., 301, Building 12, Nanyou A-zone, Jiannan Road,
Shenzhen, Guangdong Province, CN-518054, China, Tel: +86 755 2202 9353; fax: +86 755 2520 8256. E-mail: hongguanglishibahao@yahoo.com (*corresponding author)

ABSTRACT

In this data-based theoretical analysis, we use our random approach to analyse amino acid pairs in human collagen α 1(I) chain precursor (CA11) in order to determine which amino acid pairs are more sensitive to 95 variants with missense mutant in human CA11 protein. The rationale of this study is based on our hypothesis and previous findings that harmful variance is more likely to occur at randomly unpredictable amino acid pair position rather than at randomly predictable positions. This is reasonable to argue that the randomly predictable amino acid pairs are less likely to be deliberately evolved, whereas the randomly unpredictable amino acid pairs are probably deliberately evolved in connection with protein function. The results show that all of 95 variants occurred at randomly unpredictable amino acid pairs and the chance of a variant occurring is markedly higher in randomly unpredictable amino acid pairs than in predictable ones. Thus, the randomly unpredictable amino acid pairs are more sensitive to variance in human CA11 protein. Also the results suggest that the human CA11 protein has a natural tendency towards variants.

Keywords: Collagen α1(I) chain; Ehlers-Danlos syndrome; Probability; Randomness; Variants.

INTRODUCTION

The frequency of amino acid pairs has been introduced to predict protein secondary structure content (Chou 1999; Liu and Chou 1999). In the past, we have used two probabilistic approaches to analyse the primary structure of proteins related to different diseases (for review see Wu and Yan 2002a). In general, our first approach can predict the presence and absence of amino acid sub-sequences in a protein primary structure. We argue that the randomly predictable present and absent sub-sequences were probably not deliberately evolved, whereas the randomly unpredictable present and absent sub-sequences were more likely to be deliberately evolved. Accordingly, our first approach can classify the present amino acid sub-sequences as randomly predictable and randomly unpredictable sub-sequences. We suggest that the randomly unpredictable amino acid sub-sequences are more related with protein function and harmful variants are more likely to occur at randomly unpredictable amino acid sub-sequences rather than at randomly predictable amino acid sub-sequences.

Recently we used our approach to analyse the amino acid pairs in human haemoglobin α chain (Wu and Yan 2003) and phenylalanine hydroxylase protein (Wu and Yan 2002b) to determine which amino acid sub-sequences more sensitive to variance. The results show that randomly unpredictable amino acid pairs are more sensitive to variance. An intriguing question is brought about whether these phenomena are occasional or they represent some general sense. Another interesting subject is weather or not the length of a protein affects the results of variance, as the composed number of amino acids is 141 in human haemoglobin α chain and 452 in phenylalanine hydroxylase protein, respectively. What about relatively big proteins? Thus, further studies are needed in order to obtain more information regarding this aspect.

Collagen is a major constituent of the extracellular matrix and synthesized as precursors in form of procollagen triple helices (Tang, 2001). The disorders of fibrillar collagen metabolism bring about Ehlers-Danlos syndrome (EDS), which is a clinically and genetically heterogeneous group of congenital connective tissue disorders affecting as many as 1 in 5000 individuals (Steinmann et al. 1993). EDS is characterized in its most common form by hyperextensibility of the skin, hypermobility of joints often resulting in dislocations, and tissue fragility exemplified by easy bruising, atrophic scars following superficial injury, and premature rupture of membranes during pregnancy (Byers 1994). There are three fundamental mechanisms causing EDS: deficiency of collagen-processing enzymes, dominant-negative effects of mutant collagen a-chain, and haploinsufficiency (Mao and Bristow 2001). Defects in COL1A1 gene result in arthrochalasia (EDS-VIIa) (Beighton et al. 1998).

So far, 97 variants are documented in the Swiss-Protein data bank (Bairoch and Apweiler 2000). Of all variants, 95 belong to missense point mutations and the rest 2 are small deletions of one amino acid. In this study, we attempt to use our random approach to analyse amino acid pairs in human CA11 protein with its 95 variants in order to determine which amino acid pairs are more sensitive to the variants.

MATERIALS AND METHODS

The amino acid sequence of the human CA11 protein and its 95 variants with missense point mutants was obtained from the Swiss-Protein data bank (access number P02452, due to the limitation of space, we will not cite the numerous references related to human CA11 protein) (Bairoch and Apweiler 2000). The detailed calculations and rationales have already been published in a number of our previous studies (for a review, see Wu and Yan 2002a). Briefly, the calculation procedure with its examples is as follows.

Amino acid pairs in human CA11 protein

The human CA11 protein consists of 1464 amino acids. The first and second amino acids are counted as an amino acid pair, the second and third as another amino acid pair, the third and fourth, until the 1463rd and 1464th, thus there is a total of 1463 amino acid pairs. In general, there are 20 types of amino acids, any amino acid pair can be composed from any of 20 types of amino acids so, theoretically, there are 400 (20^2) possible amino acid pairs. Again there are 1463 amino acid pairs in human CA11 protein, which are more than 400 types of theoretical amino acid pairs, clearly some of 400 types of theoretical amino acid pairs should appear more than once. Meanwhile we may expect that some of 400 kinds of theoretical amino acid pairs are absent from human CA11 protein.

Actual frequency and randomly predicted frequency in human CA11 protein

The randomly predicted frequency is calculated according to a simple permutation principle (Feller 1968). For example, there are 141 alanines (A) and 28 asparagines (N) in human CA11, the predicted frequency of amino acid pair "AN" would be 3 (141/1464×28/1463×1463=2.697). Actually we can find three "AN"s in human CA11, so the actual frequency of "AR" is 3. Hence we have three relationships between the actual and predicted frequencies, i.e. the actual frequency is smaller, equal to and larger than the predicted frequency, respectively.

Randomly predictable present amino acid pairs

As described in the last section, the predicted frequency of randomly presence of amino acid pair “AN” would be 3 and “AN” does appear three times in human CA11, so the presence of “AN” is randomly predictable.

Randomly unpredictable present amino acid pairs

There are 71 arginines (R) in human CA11, the frequency of random presence of amino acid pair “AR” would be 7 (141/1464×71/1463×1463=6.838), i.e. there would be seven “AR”s in human CA11. But actually the “AR” appears ten times, so the presence of “AR” is randomly unpredictable. In this case the actual frequency of “AR” is larger than the predicted frequency of “AR”. In other case the actual frequency is smaller than the predicted frequency. For example, the predicted frequency of “AA” is 13 (141/1464×140/1463×1463=13.484), whereas the actual frequency of “AA” is 6.

Randomly predictable absent amino acid pairs

There are 6 tryptophans (W) in human CA11, the frequency of random presence of “RW” would be 0 (71/1464×6/1463×1463=0.291), i.e. the amino acid pair “RW” would not appear in this protein, which is true in the real situation. Thus the absence of “RW” is randomly predictable.

Randomly unpredictable absent amino acid pairs

The frequency of random presence of “RR” would be 3 (71/1464×70/1463×1463=3.395), i.e. there would be three “RR”s in human CA11. However no “RR” appears in this protein, therefore the absence of “RR” from human CA11 is randomly unpredictable.

Variants in randomly predictable and unpredictable amino acid pairs

A variant with point mutation results in two amino acid pairs being replaced by another two pairs. After calculating the predicted frequency and comparing with the actual frequency, it can be determined that the original amino acid pairs belong to predictable/unpredictable amino acid pairs.

Difference between actual and predicted frequencies

For the numerical analysis, we calculate the difference between actual frequency (AF) and predicted frequency (PF) of affected amino acid pairs, i.e. $\sum$(AF−PF). For instance, a variant at position 350 substitutes “G” for “R” which results in two amino acid pairs “VG” and “GA” changing to “VR” and “RA”, because the amino acid is “V” at position 349 and “A” at position 351. The actual frequency and predicted frequency are 12 and 12 for “VG”, 63 and 37 for “GA”, 3 and 2 for “VR”, and 2 and 7 for “RA”, respectively. Thus the difference between actual and predicted frequencies is 26 with regard to the original amino acid pairs, i.e. (12−12)+(63−37), and –4 for the mutant amino acid pairs, i.e. (3−2)+(2−7). In this way, we can compare the frequency difference in the amino acid pairs affected by variants.

RESULTS

General information on amino acid pairs and variants in human CA11 protein

Of 400 types of theoretical amino acid pairs, 135 are absent from human CA11 protein including 74 randomly predictable and 61 randomly unpredictable. Consequently 1463 amino acid pairs in human CA11 include only 265 types of theoretical amino acid pairs (400-35=265), i.e. some amino acid pairs should appear more than once (Table 1).

Of 265 types of theoretical amino acid pairs in human CA11 protein, 70 types are randomly predictable and 195 are randomly unpredictable. As mentioned above, some types of amino acid pairs appear more than once, thus, of 1463 amino acid pairs in human

CA11, 124 pairs are randomly predictable and 1339 pairs are randomly unpredictable. Therefore the number of variants occurring with respect to these amino acid pairs in human CA11 protein can be detected by probability. (Table 2).

Table 1: Appearance of theoretical kinds of amino acid pairs in human CA11 protein

Times of appearance	Number of theoretical kinds of amino acid pairs
0	135
1	96
2	63
3	36
4	14
5	7
6	5
7	7
8	2
9	2
10	5
11	4
12	3
13	1
14	1
15	2
17	1
18	1
20	2
21	2
22	1
23	1
26	1
34	1
44	1
46	1
49	1
61	1
63	1
127	1
130	1

Table 2: Occurrence of variants with respect to randomly predictable and unpredictable amino acid pairs in human CA11 protein

CA11	Types		Pairs		Variants		Ratio	
	number	%	number	%	number	%	Variants/Types	Variants/Pairs
Predictable	70	26.42	124	8.48	0	0	0/70=0	0/124=0
Unpredictable	195	73.58	1339	91.52	95	100	95/195=0.49	95/1339=0.07
Total	265	100	1463	100	95	100	95/265=0.36	95/1463=0.06

Variants of human CA11 protein in randomly predictable and unpredictable present amino acid pairs

As mentioned in materials and methods section, in general, a point mutant protein leads to two amino acid pairs being original by another two and their actual frequency can be smaller, equal to or larger than the predicable frequency. Tables 3 and 4 detail the situations related to original and mutant amino acid pairs, respectively and the relationship between their actual and predicted frequencies.

Table 3 can be read as follows. The first column classifies the original amino acid pairs into randomly predictable and unpredictable. The second and third columns show in which type of amino acid pairs the variant occurs, for example, the first two cells in columns 2 and 3 indicate that the actual frequencies are equal to the predicated frequencies in both amino acid pairs I and II. The fourth and fifth columns indicate how many variants occur in amino acid pairs I and II. No variant occurs at both amino acid pairs whose actual frequencies are equal to predicted frequencies. The sixth column indicates the percentage of 95 variants occurring at predictable and unpredictable amino acids.

Tables 2 and 3 show that all variants occur at randomly unpredictable amino acid pairs and no variant occurs in randomly predictable amino acid pairs. These results mean that 195 types of randomly unpredictable present amino acid pairs account for all of 95 variants in human CA11 protein, whereas 70 types of randomly predictable present amino acid pairs do not account for any variants. Still we can see the ratio in Table 2 that the chance of occurring of variants in unpredictable amino acid pairs is much larger than in predictable amino acid pairs. These phenomena strongly support our rationale that harmful variants are more likely to occur at randomly unpredictable amino acid pair positions rather than at randomly predictable. Thus the randomly unpredictable amino acid pair positions are more sensitive to the variants.

When looking at the unpredictable amino acid pairs in Table 3, 98.95% of these pairs are characterised by one or both original pairs whose actual frequency is larger than their predicted frequency (the first three rows in unpredictable pairs). Comparing with the normal human CA11 protein, the impact of variants is to narrow the difference between actual and predicted frequencies by means of reducing the actual frequency which implies that the variants associated with the construction of amino acid pairs is randomly predictable. In other words, the variants result in the construction of amino acid pairs which are more likely to be naturally evolved. Additionally, only one variant occurs in the amino acid pairs whose actual frequency is smaller than predicted frequency in both pairs. This interesting phenomenon suggests that it is difficult for variants to narrow the difference between actual and predicted frequencies by means of increasing the actual frequency. Commonly, reduction of actual frequency would lead to the construction of amino acid pairs against natural direction.

Table 4 can be read as follows. The first and second columns indicate the actual and predicted situations in amino acid pairs I and II, the third and fourth columns indicate the number of variants occurring at amino acid pairs I and II and their percents, the fifth column shows total classifications.

Table 4 shows that 38.95% of variants result in one or both mutant amino acid pairs are absent in normal human CA11 protein (AF=0). Furthermore 82.11% of variants target one or both mutant amino acid pairs with their actual frequency smaller than predicted frequency (†). These phenomena indicate that the amino acid pairs in mutant CA11 protein are more randomly constructed.

Frequency difference of amino acid pairs affected by variants

The difference between actual and predicted frequencies represents a measure of randomness of construction of amino acid pairs, i.e. the smaller the difference, the more random the construction of amino acid pairs. In particular, (i) the larger the positive difference, the more randomly unpredictable amino acid pairs are present; and (ii) the larger the negative difference, the more randomly unpredictable amino acid pairs are absent.

Considering all 95 variants, the difference between actual and predicted frequencies is 50.84±3.69 (mean±SE, ranging from –41 to 109) in original amino acid pairs. This means that the variants occur in the amino acid pairs which appear more than their predicted frequency. Meanwhile, the difference between actual and predicted frequencies is –1.53±0.59 (mean±SE, ranging from –23 to 31) in mutant amino acid pairs. This implies that the mutant amino acid pairs are more randomly constructed in the variants of CA11 protein, as their actual and predicted frequencies are about the same. Striking statistical difference is found between the original and mutant amino acid pairs (P<0.0001). Figure 1 shows the distribution of difference between actual and predicted frequencies.

Table 3: Classification of original amino acid pairs induced by variants in human CA11 protein

CA11	Amino acid pairs		Variants		Total
	I	II	number	%	%
Predictable	AF=PF	AF=PF	0	0	0
Unpredictable	AF>PF	AF>PF	58	61.05	100.00
	AF>PF	AF=PF	7	7.37	
	AF>PF	AF<PF	28	29.47	
	AF>PF	– –	1	1.05	
	AF<PF	AF=PF	0	0	
	AF<PF	AF<PF	1	1.05	

AF: actual frequency; PF: predicted frequency.

Table 4: Classification of mutant amino acid pairs induced by variants in human CA11 protein

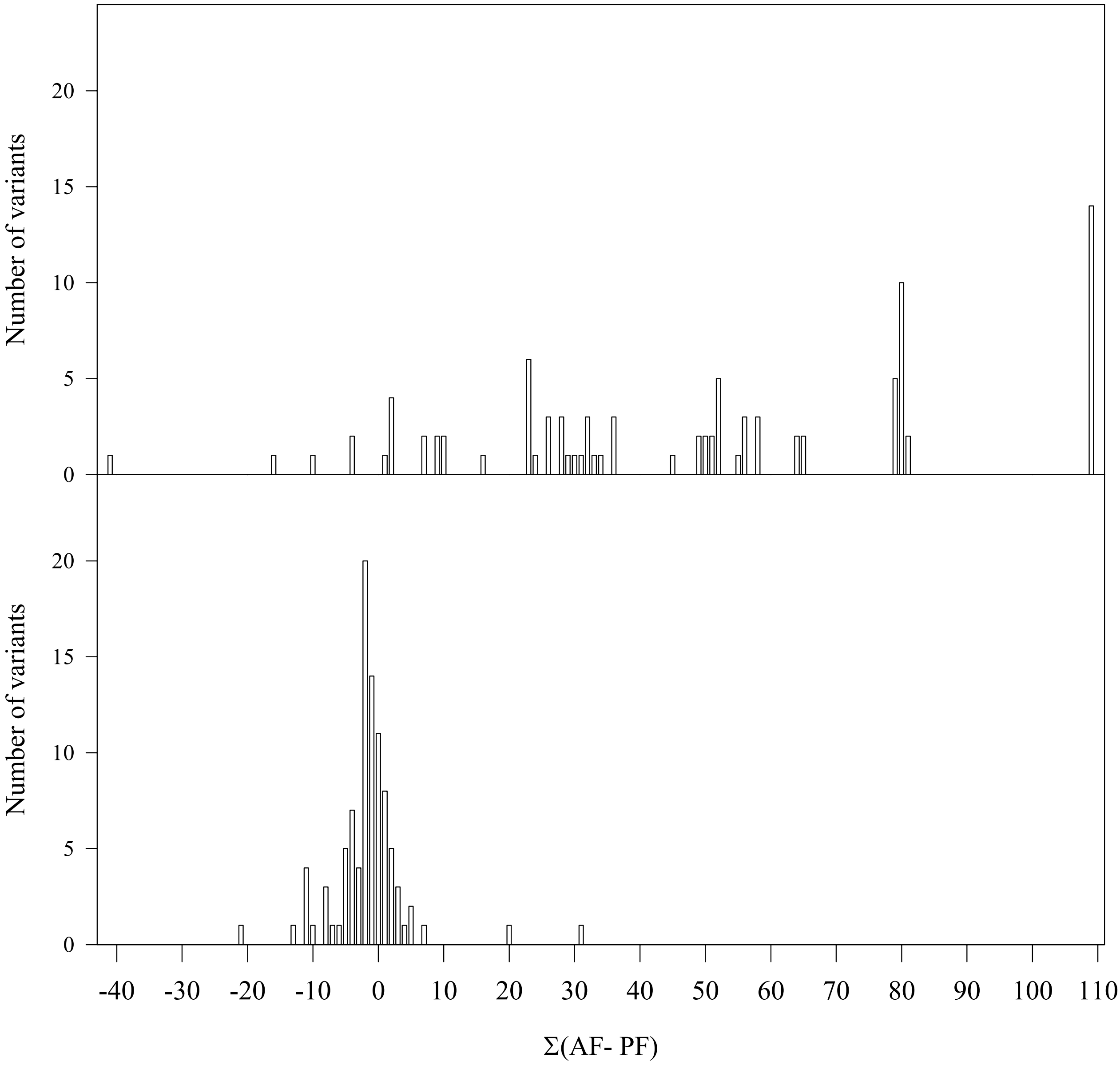
Amino acid pairs		Variants		Total
I	II	Number	%	%
AF=0, PF>0	AF=0, PF>0	5†	5.26	38.95
AF=0, PF>0	AF=PF=0	1†	1.05	
AF=0, PF>0	AF=PF>0	7†	7.37	
AF=0, PF>0	AF<PF, AF≠0	9†	9.47	
AF=0, PF>0	AF>PF	13†	13.68	
AF=PF=0	AF=PF=0	0	0	
AF=PF=0	AF=PF>0	0	0	
AF=PF=0	AF<PF, AF≠0	1†	1.05	
AF=PF=0	AF>PF	1	1.05	
AF<PF, AF≠0	AF<PF, AF≠0	14†	14.74	61.05
AF<PF, AF≠0	AF=PF>0	13†	13.68	
AF<PF, AF≠0	AF>PF	16†	16.84	
AF=PF>0	AF=PF>0	1	1.05	
AF>PF	AF>PF	7	7.37	
AF>PF	– –	1	1.05	
AF=PF>0	AF>PF	6	6.32	

† indicates the variants which target one or both mutant amino acid pairs with their actual frequency smaller than predicted one (totally 82.11%).

DISCUSSION

Currently two explanations are commonly proposed to explain why some amino acids are mutated more frequently than the others. The first is targeted mutagenesis, which defined the “hotspot” sites sensitive to endogenous and exogenous mutagens (Rideout et al. 1990; Montesano et al. 1997; Hainaut and Pfeifer 2001). The second is the function selection, which indicates the disruption of protein functions may depend upon the position of the mutation/variant in the protein (Ory et al. 1994; Forrester et al. 1995; Aas et al. 1996). However, these explanations still do not answer why some amino acid sub-sequences are sensitive to variants.

Figure 1: Frequency difference between original (upper panel) and mutant (lower panel) amino acid pairs induced by variants from human CA11 protein.



This problem can be assessed from different approaches such as empirical (regression analysis), experimental (artificial and natural mutations), and computation (multiple sequence comparisons and alignments), etc. The probabilistic approach can contribute considerable understanding to this problem. By means of a random approach to estimate the variants from human haemoglobin α chain (Wu and Yan 2003) and phenylalanine hydroxylase protein (Wu and Yan 2002b), we found that 94% of variants occur in randomly unpredictable amino acid pairs. In this study we, correspondingly, analyse the amino acid pairs in human CA11 protein to determine which amino acid pairs are more sensitive to variants. The results show that all of 95 variants occur in randomly unpredictable amino acid pairs, which further confirm our hypothesis that the randomly unpredictable amino acid pairs are more sensitive to variants.

Based on our previous studies, our argument is that the functional amino acid pairs are more likely to be deliberately evolved and thus the actual frequency should be different from the randomly predicted frequency. As the randomly predicted frequency is the highest potential for construction of amino acid pairs, it is important to find whether or not a variant leads to the actual frequency to approach the randomly predicted frequency. If so, the protein has a natural trend to mutate; if not, the protein does not have a natural trend to mutate. The present study demonstrates that 98.95% of variants bring about one or both original amino acid pairs whose actual frequency is larger than predicted frequency, that 38.95% of variants result in one or both mutant amino acid pairs which are absent in normal human CA11 protein (AF = 0) and that 82.11% of variants lead to one or both mutant amino acid pairs with their actual frequency smaller than predicted frequency. All of these results reveal that the human CA11 protein has a natural trend towards variance.

With respect to randomly unpredictable absent and present amino acid pairs, the difference between actual and predicted frequencies is interesting, because the randomly predicable absent and present frequency represents the more likely naturally occurring event, i.e. the construction of amino acid pairs should be the least energy- and time-consuming. Thus the difference between actual and predicted frequencies should be engineered by the evolutionary process, i.e. the larger the difference, the larger the impact by the evolutionary process. Diminishing of difference between actual and predicted frequencies has been shown in this study (Figure 1), thus the variants of human CA11 protein in fact represent a degeneration process inducing Ehlers-Danlos syndromes.

The most common type of CA11 variants is the substitution of a glycine residue by another amino acid, which phenomenon could be explained by our analyses, because many amino acid pairs with glycine have very big difference between their actual and predicted frequencies, such as 24 in “AG”, 25 in “RG”, 53 in “PG”, 56 in “GP”, –97 in “GG”, and so on. The current study highlights the changes in the frequencies of amino acid pairs in mutant human CA11 protein. In general, point mutations modify the configuration of amino acid pairs in a protein, which could target the changes in the secondary structure contents and consequently affect biologic functions of the protein. Variants with such a glycine being original are thought to interfere with the normal folding of the mutant α chain into triple helices with other collagen α chain (Kuvaniemi et al. 1997). Thus, our approach may provide useful insight into the molecular mechanisms of arthrochalasia.

Taking both our previous and present studies into account, the changes in amino acid frequencies induced by variants reveal similar ways in proteins with different length, which encourages us to explore an interesting aspect regarding which amino acid sub-sequences are more or less sensitive to variants. If such a general rule could be drawn, then we could gain not only more insight into the relationship between protein variants and its related disorders but, more importantly we could offer more attention to these sensitive sub-sequences in order to prevent them from variants. Moreover the possible sub-sequences sensitive to the, currently, unknown variants could be predicted.

REFERENCES

Aas T, Borresen AL, Geisler S, Smith-Sorensen B, Johnsen H, Varhaug JE, Akslen LA and Lonning PE. Specific p53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. *Nat Med* 1996;2:811–4

Bairoch A and Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;28:45–8

Beighton P, De Paepe A, Steinmann B, Tsipouras P and Wenstrup RJ. Ehlers-Danlos syndromes: revised nosology, Villefranche, 1997. Ehlers-Danlos National Foundation (USA) and Ehlers-Danlos Support Group (UK). *Am J Med Genet* 1998;77:31–7

Byers PH. Recent advances and current understanding of the clinical and genetic heterogeneity. *J Invest Dermatol* 1994;103:47S–52S

Chou KC. Using pair-coupled amino acid composition to predict protein secondary structure content. *Protein Engin* 1999;12:1041–50

Feller W. An introduction to probability theory and its applications, 1968, 3rd ed., Vol, I. John Wiley and Sons, New York

Forrester K, Lupold SE, Ott VL, Chay CH, Band V, Wang XW and Harris CC. Effects of p53 mutants on wild-type p53-mediated transactivation are cell type dependent. *Oncogene* 1995;10:2103–11

Hainaut P, Pfeifer GP. Patterns of p53 G→T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. *Carcinogenesis* 2001;22:367–74

Kuvaniemi H, Tromp G and Prockop DJ. Mutations in fibrillar collagens (types I, II, III, and IV), fibril-associated collagen (type IV), and network-forming collagen (type X) cause a spectrum of diseases of bone, cartilage, and blood vessels. *Hum Mutat* 1997;9:300–15

Liu W and Chou KC. Prediction of protein secondary structure content. *J Protein Chemi* 1999;18:473–80

Mao JR and Bristow J. The Ehlers-Danlos syndrome: on beyond collagens. *J Clin Invest* 2001;107:1063–9

Montesano R, Hainaut P and Wild CP. Hepatocellular carcinoma: from gene to public health. *J Natl Cancer Inst* 1997;89:1844–51

Ory K, Legros Y, Auguin C and Soussi T. Analysis of the most representative tumour-derived p53 mutants reveals that changes in protein conformation are not correlated with loss of transactivation or inhibition of cell proliferation. *EMBO J* 1994;13:3496–504

Rideout WM, Coetzee GA, Olumi AF and Jones PA. 5-Methylcytosine as an endogenous mutagen in human LL receptor and p53 genes. *Science* 1990;249:1288–90

Steinmann B, Royce P and Superti-Furga A. The Ehlers-Danlos Syndrome. In: Connective Tissue and Its Heritable disorders. 1993, Royce P and Steinmann B, editors. Wiley-Liss. New York, USA. 351–407

Tang BL. ADAMTS: a novel family of extracellular matrix proteases. *Int J Biochem Cell Biol* 2001;33:33–44

Wu G, Yan S Randomness in the primary structure of protein: methods and implications. *Mol Biol Today* 2002a;3:55-69

Wu G and Yan S Estimation of amino acid pairs in human phenylalanine hydroxylase protein sensitive to variants by means of a random approach. *Peptides* 2002b;23:2085–90

Wu G and Yan S Determination of amino acid pairs in human haemoglobin α chain sensitive to variants by means of a random approach. *Comp Clin Path* 2003;12:21–5