EXCLI Journal 2004;3:20-28 – ISSN 1611-2156
received: 2. May 2004, accepted: 12. May 2004, published: 28. May 2004

Original article:

Susceptible Amino Acid Pairs in Variants in Human Collagen α 1(III) 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 a previous study we presented a technique to differentiate between randomly predictable and randomly unpredictable amino acid pairs. We have shown that all of 95 variants in human collagen α 1(I) chain precursor (CA11) occurred at randomly unpredictable amino acid pairs. To study whether this is unique for the human collagen α 1(I) chain precursor (CA11) or whether this is a more general principle we examined another human collagen precursor, the human collagen α 1(III) chain precursor (CA13), in which 106 variants have been documented. Interestingly, all of 106 variants occurred at randomly unpredictable amino acid pairs in human CA13. In conclusion, the observation that randomly unpredictable amino acid pairs are more susceptible to variation seems to be a more general principle for human collagen chain precursor proteins.

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

INTRODUCTION

This is a further study on the human collagen chain precursor using our probabilistic approach. In our previous study published in this journal (Wu and Yan 2004a), we have shown that all of 95 variants in human collagen α 1(I) chain precursor (CA11) occurred at randomly unpredicted amino acid pairs as our approach classifies the amino acid pairs in a protein into the randomly predictable and the unpredictable.

However, the effort needs to extend to other human collagen precursors in order to determine whether or not similar results can be obtained. This effort is important because 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 result in the Ehlers-Danlos syndrome, which is a clinically and genetically heterogeneous group of congenital connective tissue disorders affecting as many as 1 in 5000 individuals (Steinmann et al. 2001). The Ehlers-Danlos syndrome 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 Ehlers-Danlos syndrome: deficiency of collagen-processing enzymes, dominant-negative effects of mutant collagen α-chain, and haploinsufficiency (Mao and Bristow 2001).

Among various human collagen precursors, we are interested in the human collagen α 1(III) chain precursor (CA13), in which so far 106 variants have been documented. Therefore, the human CA13 provides us with the opportunity of seeing the patterns of variants in terms of randomly predictable and unpredictable amino acid pairs. If the human CA13 can follow a similar pattern, we would expect that the pattern plays an important role in the mutation process, and we could predict the possible future variants according to this pattern. Thus the aim of this study is to analyse amino acid pairs in human CA13 with its 106 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 CA13 and its 106 variants with missense point mutants were obtained from the Swiss-Prot data bank (accession no. P02461) (Bairoch and Apweiler 2000). The calculations have already been published in this journal (Wu and Yan 2004a). Briefly, the calculation procedure with its examples is as follows.

Amino acid pairs in human CA13

The human CA13 consists of 1466 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 1465th and 1466th, thus there are 1465 amino acid pairs. In general, there are 20 kinds of amino acids, any amino acid pair can be composed from any of 20 kinds of amino acids, so there are 400 types of theoretically possible amino acid pairs. Again there are 1465 amino acid pairs in human CA13, which are more than 400 types of possible amino acid pairs, clearly some of 400 types of possible amino acid pairs should appear more than once. Meanwhile we may expect that some of 400 types of possible amino acid pairs are absent from human CA13.

Actual frequency and randomly predicted frequency in human CA13

The randomly predicted frequency is calculated according to a simple permutation principle (Feller 1968). For example, there are 115 alanines (A) in human CA13, the predicted frequency of amino acid pair "AA" would be 9 (115/1466×114/1465×1465=8.943). Actually we can find nine "AA"s in human CA13, so the actual frequency of "AA" is 9. Hence we have three relationships between the actual and predicted frequencies, i.e. the actual frequency is smaller than, 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 "AA" would be 9 and "AA" does appear nine times in human CA13, so the presence of "AA" is randomly predictable.

Randomly unpredictable present amino acid pairs

There are 413 glycines (G) in human CA13, the frequency of random presence of amino acid pair "AG" would be 32 (115/1466×413/1465×1465=32.398), i.e. there would be 32 "AG"s in human CA13. But actually the "AG" appears 45 times, so the presence of "AG" is randomly unpredictable. In this case the actual frequency of "AG" is larger than the predicted frequency of "AG". In other case the actual frequency is smaller than the predicted frequency. For example, there are 55 aspartic acids (D) in human CA13, the predicted frequency of "AD" is 4 (115/1466×55/1465×1465=4.314), whereas the actual frequency of "AD" is only 2.

Randomly predictable absent amino acid pairs

There are 60 arginines (R) and 7 tryptophans (W) in human CA13, the frequency of random presence of "RW" would be 0 (60/1466×7/1465×1465=0.286), i.e. the amino acid pair "RW" would not appear in

human CA13, which is true in the real situation. Thus the absence of "RW" is randomly predictable.

Randomly unpredictable absent amino acid pairs

There are 31 threonines (T) in human CA13, the frequency of random presence of "AT" would be 2 (115/1466×31/1465×1465=2.432), i.e. there would be two "AT"s in human CA13. However no "AT" appears in human CA13, therefore the absence of "AT" from human CA13 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. ∑(AF−PF). For instance, a variant at position 726 substitutes "G" for "R" which results in two amino acid pairs "QG" and "GM" changing to "QR" and "RM", because the amino acid is "Q" at position 725 and "M" at position 727. The actual frequency and predicted frequency are 18 and 12 for "QG", 5 and 5 for "GM", 1 and 2 for "QR", and 0 and 1 for "RM", respectively. Thus the difference between actual and predicted frequencies is 6 with regard to the original amino acid pairs, i.e. (18−12)+(5−5), and −2 for the mutant amino acid pairs, i.e. (1−2)+(0−1). 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 CA13

Of 400 types of possible amino acid pairs, 131 are absent from human CA13 including 57 randomly predictable and 74 randomly unpredictable. Consequently 1465 amino acid pairs in human CA13 include only 269 types of possible amino acid pairs (400−131=269), i.e. some amino acid pairs should appear more than once (Table 1).

Of 269 types of possible amino acid pairs in human CA13, 71 types are randomly predictable and 198 are randomly unpredictable. As mentioned above, some types of amino acid pairs appear more than once, thus, of 1465 amino acid pairs in human CA13, 142 pairs are randomly predictable and 1323 pairs are randomly unpredictable. Therefore the number of variants occurring with respect to these amino acid pairs in human CA13 can be detected by probability (Table 2).

Variants of human CA13 in randomly predictable and unpredictable present amino acid pairs

As mentioned in materials and methods section, in general, a point mutation leads to two amino acid pairs being replaced by another two and their actual frequency can be smaller than, 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 1: Appearance of possible types of amino acid pairs in human CA13

Appearance	Possible types of amino acid pairs
0	131
1	103
2	55
3	37
4	13
5	11
6	10
7	7
9	2
10	2
11	3
12	2
13	3
14	2
17	1
18	1
19	3
20	1
21	1
23	1
26	1
27	1
28	1
34	1
40	1
42	1
45	1
50	1
53	1
109	1
153	1

Table 2: Occurrence of variants with respect to randomly predictable and unpredictable amino acid pairs in human CA13

Human CA13	Types		Pairs		Variants		Ratio	
	number	%	number	%	number	%	variants/types	variants/pairs
Predictable	71	26.39	142	9.69	0	0	0/71=0	0/142=0
Unpredictable	198	73.61	1323	90.31	106	0	106/198=0.54	106/1323=0.08
Total	269	100	1465	100	106	100	106/269=0.39	106/1465=0.07

Table 3: Classification of original amino acid pairs induced by variants in human CA13

CA13	Amino acid pairs		Variants		Total
	I	II	number	%	%
Predictable	AF=PF	AF=PF	0	0	0
Unpredictable	AF>PF	AF>PF	88	83.02	100
	AF>PF	AF=PF	5	4.72	
	AF>PF	AF<PF	13	12.26	
	AF<PF	AF=PF	0	0	
	AF<PF	AF<PF	0	0	

AF: actual frequency; PF: predicted frequency.

Table 4: Classification of mutant amino acid pairs induced by variants in human CA13

Amino acid pairs		Variants		Total
I	II	Number	%	%
AF=0, PF>0	AF=0, PF>0	2†	1.89	32.08
AF=0, PF>0	AF=PF=0	1†	0.94	
AF=0, PF>0	AF=PF>0	4†	3.77	
AF=0, PF>0	AF<PF, AF≠0	20†	18.87	
AF=0, PF>0	AF>PF	7†	6.60	
AF=PF=0	AF=PF=0	0	0	
AF=PF=0	AF=PF>0	0	0	
AF=PF=0	AF<PF, AF≠0	0†	0	
AF=PF=0	AF>PF	0	0	
AF<PF, AF≠0	AF<PF, AF≠0	36†	33.96	67.92
AF<PF, AF≠0	AF=PF>0	6†	5.66	
AF<PF, AF≠0	AF>PF	24†	22.64	
AF=PF>0	AF=PF>0	1	0.94	
AF>PF	AF>PF	1	0.94	
AF=PF>0	AF>PF	4	3.77	

† indicates the variants which target one or both mutant amino acid pairs with their actual frequency smaller than predicted one (totally 94.34%).

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 106 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 198 types of randomly unpredictable present amino acid pairs account for all of 106 variants in human CA13, whereas 71 types of randomly predictable present amino acid pairs do not account for any variant. Still we can see the ratio in Table 2 that the chance of occurring of variants in unpredictable amino acid pairs is 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, all 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 CA13, 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 are randomly predictable. In other words, the variants result in the construction of amino acid pairs, which are more likely to be naturally evolved. No 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 32.08% of variants result in one or both mutant amino acid pairs are absent in normal human CA13 (AF=0). Furthermore 94.34% of variants form one or both mutant amino acid pairs with their actual frequency smaller than predicted frequency (†).

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 106 variants, the difference between actual and predicted frequencies is 62.75±3.45 (mean±SE, ranging from –1 to 106) in original amino acid pairs. This means that the variants occur in the amino acid pairs that appear more than their predicted frequency. Meanwhile, the difference between actual and predicted frequencies is –4.50±0.54 (mean±SE, ranging from –18 to 19) in mutant amino acid pairs. This implies that the mutant amino acid pairs are more randomly constructed in the variants of CA13, 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.

DISCUSSION

Using a probabilistic model we recently have shown that all of the 95 protein variants of human CA11 occurred at randomly unpredicted amino acid pairs. The present study was performed to analyze, whether a similar pattern of variants can also be observed for another collagen precursor, the human CA13. The human CA13 occurs in most soft connective tissues along with the human CA11, this might suggests that the human CA13 would follow a similar pattern of variants. However the human CA13 and human CA11 are functionally different. Defects in the human CA13 are the cause of type IV Ehlers-Danlos syndrome that is the most severe form of the disease and often produces life-threatening consequences, such as rupture of the arteries, bowel, or uterus. Still, defects in the human CA13 may be a cause of Gottron type acrogeria whose characteristics are atrophy and mottled-type hyperpigmentation of the acral skin, resulting in an aged appearance. While defects in the human CA11 are a cause of osteogenesis imperfecta type IV, which is charaterized by normal sclerae, moderate to mild deformity and variable short stature. Dentinogenesis imperfecta is common and hearing loss occurs in some patients. The human CA11 is also involves in pathogenesis of dermatofibrosarcoma protuberans through a chromosomal translocation t(17;22)(q22;q13), and dermatofibrosarcoma protuberans is an uncommon, locally aggressive, but rarely metastasizing tumour of the deep dermis and subcutaneous tissue, and occurs during early or middle adult life and is most frequently located on the trunk and proximal extremities (Beighton et al. 1998).

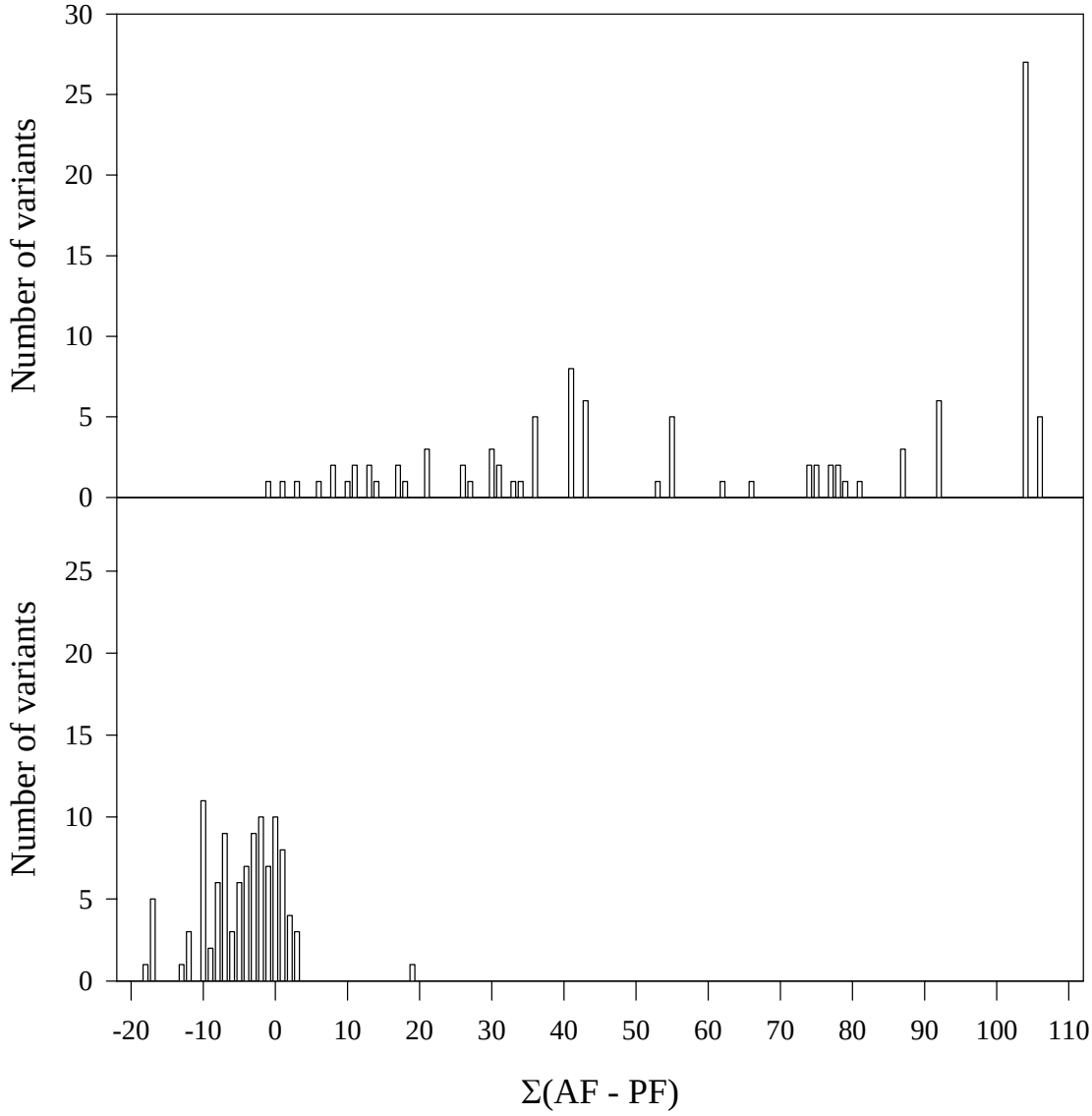

Figure 1: Frequency difference between original (upper panel) and mutant (lower panel) amino acid pairs induced by variants from human CA13.

In addition to the functional difference between human CA13 and CA11, the percentage identity of human CA13 with respect to human CA11 is 61%, which is less than human CA12 (human collagen α 1(II) chain precursor) with 70% identity when we perform the multiple sequence comparisons and alignments using BlastP program. However, the human CA12 has only about 45 variants despite of its similarity with human CA11.

Our results show that the human CA13 does follow a similar pattern of variants as human CA11 does, i.e. (i) all the variants occur at randomly unpredictable amino acid pairs (Tables 2 and 3), (ii) most of variants form one or both mutant amino acid pairs with their actual frequency smaller than predicted frequency (Table 4), and (iii) the variants reduce the difference between the actual and predicted frequencies (Figure 1).

If we compare the human CA13 with the proteins that have more variants than other proteins in Swiss-Prot data bank (Table 5), we find that the CA11, CA13 and CA54 (human collagen α5(IV) chain precursor) are likely to follow the similar pattern of variants, because no variants are found in their randomly predictable amino acid pairs. Thus we would predict that the new variants would

occur at the randomly unpredictable amino acid pairs.

Table 5: Comparison with proteins that have more variants.

Protein	Number of variants	Ratio				References
		Predictable		Unpredictable		
		Variant/type	Variant/Pair	Variant/type	Variant/Pair	
ATP7	125	0.03	0.01	0.48	0.10	Wu and Yan 2004b
BTK	112	0.08	0.05	0.55	0.22	Wu and Yan 2003a
CA11	95	0	0	0.46	0.07	Wu and Yan 2004a
CA13	106	0	0	0.54	0.08	This study
CA54	151	0	0	0.78	0.10	Wu and Yan 2003b
F9	99	0.07	0.06	0.60	0.28	Wu and Yan 2003c
GLCM	109	0.07	0.05	0.64	0.27	Wu and Yan 2003d
HBA	133	0.21	0.18	2.05	1.30	Wu and Yan 2003e
LDLR	127	0.05	0.03	0.57	0.18	Wu and Yan 2002a
Human p53	190	0.14	0.10	1.48	0.66	Wu and Yan 2003f
PAH	187	0.12	0.09	1.26	0.54	Wu and Yan 2002b
VHL	109	0.16	0.13	1.03	0.62	Wu and Yan 2003g

ATP7: human copper-transporting ATPase 2; BTK: human Bruton's tyrosine kinase; CA54: human collagen α5(IV) chain precursor; F9: human coagulation factor IX precursor; GLCM: human β-glucocerebrosidase; HBA: human haemoglobin chain; LDLR: human low-density lipoprotein receptor; PAH: human phenylalanine 4-hydroxylase; VHL: human Von Hippel-Lindau disease tumor suppressor.

REFERENCES

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

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

Mao JR, and Bristow J. The Ehlers-Danlos syndrome: on beyond collagens. *J Clin Invest* 2001;107:1063–9
Steinmann B, Royce P, Superti-Furga A. The Ehlers-Danlos Syndrome. In: Connective Tissue and Its Heritable disorders. Royce P, Steinmann B, editors. Wiley-Liss. New York, USA. 1993; p 351–407

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

Wu G, Yan S. Determination of amino acid pairs sensitive to variants in human low-density lipoprotein receptor precursor by means of a random approach. *J Biochem Mol Biol Biophys* 2002a;6:401–6

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, Yan S. Determination of amino acid pairs sensitive to variants in human Bruton's tyrosine kinase by means of a random approach. *Mol Simul* 2003a;29:249–54

Wu G, Yan S. Analysis of amino acid pairs sensitive to variants in human collagen α5(IV) chain precursor by means of a random approach. *Peptides* 2003b;24:347–52

Wu G, Yan S. Determination of amino acid pairs sensitive to variants in human coagulation factor IX precursor by means of a random approach. *J Biomed Sci* 2003c;10:451–4

Wu G, Yan S. Determination of amino acid pairs sensitive to variants in human β-glucocerebrosidase by means of a random approach. *Protein Eng* 2003d;16:195–9

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* 2003e;12:21–5

Wu G, Yan S. Determination of amino acid pairs in human p53 protein sensitive to mutations/variants by means of a random approach. *J Mol Model* 2003f;9:337–41

Wu G, Yan S. Determination of amino acid pairs in Von Hippel-Lindau disease tumour suppressor (G7 protein) sensitive to variants by means of a random approach. *J Appl Res* 2003g (in press)

Wu G, Yan S. Amino acid pairs sensitive to variants in human collagen α 1(I) chain precursor. *EXCLI J* 2004a;3:10–9

Wu G, Yan S. Determination of amino acid pairs sensitive to variants in human copper-transporting ATPase 2. *Biochem Biophys Res Commun* 2004b; 319:27–31