J Mol Model (2003) 9:337–341
DOI 10.1007/s00894-003-0155-8

ORIGINAL PAPER

Guang Wu · Shaomin Yan

# Determination of amino acid pairs in human p53 protein sensitive to mutations/variants by means of a random approach

Received: 13 March 2003 / Accepted: 30 June 2003 / Published online: 30 August 2003


**Abstract** This is the continuation of our studies using random approaches to analyse the p53 protein family. In this data-based theoretical analysis, we use the random approach to analyse the amino acid pairs in human p53 protein in order to determine which amino acid pairs are more sensitive to 190 human p53 mutations/variants. The rationale of this study is based on our hypothesis and findings that a harmful mutation is more likely to occur at randomly unpredictable amino acid pairs, and a harmless mutation is more likely to occur at randomly predictable amino acid pairs. This is because we argue that the randomly predictable amino acid pairs should not be deliberately evolved, whereas the randomly unpredictable amino acid pairs should be deliberately evolved with a connection to protein function. The results show, for example, that 93.16% of 190 mutations/variants occur at randomly unpredictable amino acid pairs. Thus, the randomly unpredictable amino acid pairs are more sensitive to mutations/variants in human p53 protein. The results also suggest that the human p53 protein has a tendency for the occurrence of mutation/variants.

**Keywords** Human p53 · Mutations · Probability randomness · Randomness

## Introduction

The human p53 protein plays an extremely important role in tumour suppression and apoptosis. Mutations/variants in the p53 protein are the most frequently observed genetic alternations in human cancer. [1] Various factors including toxic and environmental ones can lead to mutations/variants in human p53 protein. However, it is still difficult to draw a general rule as to which amino acid sub-sequence is more sensitive to mutations/variants and which amino acid sub-sequence is less sensitive to mutations/variants. If such a general rule can be drawn, then we can not only gain more insight into the relationship between the p53 protein and its related tumours, but more importantly, we can pay more attention to these sensitive sub-sequences in order to protect them from mutations/variants. Moreover, we can in principle even predict the possible sub-sequences sensitive to currently unknown mutations/variants.

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. Currently two explanations are commonly proposed to explain why some amino acids are mutated more frequently than others. The first is targeted mutagenesis, which defines "hotspot" sites sensitive to endogenous and exogenous mutagens. [2, 3, 4] The second is function selection, which indicates the disruption of p53 functions may depend upon the position of the mutation/variant in the protein. [5, 6, 7] However, these explanations still did not answer the question as to why some amino acid sequences are sensitive to mutations/variants. Probably the probabilistic approach can contribute its understanding to this problem, because in the past we have used two probabilistic approaches to analyse the primary structure of p53 proteins with the hope that these approaches can throw light into p53 and its related tumours. [8, 9, 10, 11] In general, our first approach can predict the amino acid sub-sequences that are present and absent in a protein primary structure. We argue that the randomly predictable present and absent sub-sequences should not be deliberately evolved, whereas the randomly unpredictable present and absent sub-sequences should 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-

G. Wu (✉) · S. Yan
Dream Science and Technology Consulting Co. Ltd.,
Shenzhen City, Guangdong Province, China
e-mail: hongguanglishibahao@yahoo.com

*Present address:*
G. Wu, c/o Mr. Yongqing Yan,
Shenzhen-Dic. Ltd., 1035 Nanshan Road,
Nanshan District, 518052 Shenzhen City, China

sequences are more related to protein function, and the mutations/variants in these sub-sequences may lead to dysfunction of protein. More recently we have found that a mutation that leads to the dysfunction of rat monoamine oxidase B is located in a randomly unpredictable amino-acid pair and another mutation, which does not affect rat monoamine oxidase B function, is located in a randomly predictable amino acid pair. [12]

In this study, we attempt to use our first random approach to analyse amino acid pairs in human p53 protein with its 190 mutations/variants in order to determine which amino acid pairs are more sensitive to mutations/variants.

## Materials and methods

The amino acid sequence of the human p53 protein and its 190 mutations/variants with point mutations were obtained from the Swiss-Protein data bank (access number P04637, due to the limitation of space, we will not cite the numerous references related to the human p53 protein). [13] The detailed calculations and their rationales with examples are described in the following.

### Amino acid pairs in human p53 protein

The human p53 protein is composed of 393 amino acids; we count the first and second amino acids as an amino acid pair, the second and third as another amino acid pair, the third and fourth, until the 392nd and 393rd, thus there is a total of 392 amino acid pairs. As there are 20 types of amino acids, any amino acid pair can be composed of any of 20 types of amino acids, so theoretically there are 400 ($20^2$) kinds of amino acid pairs. Again there are 392 amino acid pairs in human p53 protein, less than the 400 kinds of theoretical amino acid pairs. Clearly some of the 400 kinds of theoretical amino acid pairs are absent from the human p53 protein.

### Randomly predicted frequency and actual frequency in human p53 protein

The randomly predicted frequency is governed by the simple permutation principle. [14] For example, there are 26 arginines (R) and 30 glutamic acids (E) in human p53 protein. The random frequency of amino acid pair "RE" would be 2 (26/393×30/392×392=1.985). Actually we can find two "RE" in human p53 protein, so the actual frequency of "RE" is 2. Hence we have three relationships between actual and randomly predicted frequencies, i.e. the actual frequency is smaller than, equal to or larger than the randomly predicted frequency.

### Randomly predictable present amino acid pairs

As described in the last section, the frequency of random presence of the amino acid pair "RE" would be 2 and "RE" really appears twice in human p53 protein, so the presence of "RE" is randomly predictable.

### Randomly unpredictable present amino acid pairs

There are 24 alanines (A) in human p53 protein, the frequency of random presence of amino acid pair "AA" would be 1 (24/393×23/392×392=1.405), i.e. there would be one "AA" in human p53 protein. But actually the "AA" appears three times in human p53 protein, so the presence of "AA" is randomly unpredictable. This illustrates the case that the actual frequency of "AA" is larger than the randomly predicted frequency of "AA". Another case is that the actual frequency is smaller than the randomly predicted frequency, for example, the randomly predicted frequency of "RP" is 3 (26/393×45/392×392=2.9771), while the actual frequency is 1.

### Randomly predictable absent amino acid pairs

There are four tryptophans (W) in human p53 protein, the frequency of random presence of "AW" would be 0 (24/393×4/392×392=0.2443), i.e. the amino acid pair "AW" would not appear in human p53 protein, which is true in the real situation. Thus the absence of "AW" is randomly predictable.

### Randomly unpredictable absent amino acid pairs

The frequency of random presence of "AR" would be 2 (24/393×26/392×392=1.5878), i.e. there would be two "AR" in human p53 protein. However there are no "AR" in human p53 protein, therefore the absence of "AR" from human p53 protein is randomly unpredictable.

### Mutations/variants in randomly predictable and unpredictable amino acid pairs

Our rationale for determination of mutations/variants in randomly predictable and unpredictable present amino acid pairs in human p53 protein is based on the finding of our previous study, [12] which is described as follows. There are two mutations in rat monoamine oxidase B. The first mutation occurs at position 139 changing leucine (L) to histidine (H), and the amino acids at positions 138 and 140 are proline (P) and "A", and thus this mutation leads to four amino acid pairs changed, i.e. "PL" → "PH" and "LA" → "HA". As "PL" and "LA" are randomly predictable amino acid pairs according to our random analysis, consequently we would not expect the first mutation to lead to a substantial change in enzymatic activity, which is true in the real situation. The second mutation occurs at position 199 changing "I" to "F" leading the changes in amino acid pairs as "II" → "IF" and "IS" → "FS". As the "IS" is randomly unpredictable amino acid pairs according to our random analysis, we would expect the second mutation to lead to a substantial change in enzymatic activity, and this expectation also is true in the real situation. In this manner we hope to determine whether a mutation/variant occurs at randomly predictable or unpredictable amino acid pairs in human p53 protein in order to gain more insight into the relationship between mutations/variants and sensitivity of amino acid pairs.

### Difference between actual and randomly predicted frequencies

For the numerical analysis, we calculate the difference between the actual frequency (AF) and the predicted frequency (PF) of the affected amino acid pairs, i.e. $\sum$(AF–PF). For instance, a variant at position 240 substitutes serine (S) for isoleucine (I), which results in two amino acid pairs "NS" and "SS" changing to "NI" and "IS", because the amino acid is "N" at position 239 and "S" at position 241. The actual frequency and randomly predicted frequency are 2 and 1 for "NS", 7 and 4 for "SS", 0 and 0 for "NI", and 0 and 1 for "IS", respectively. Thus the difference between the actual and predicted frequencies is 4 with regard to the mutated amino acid pairs, i.e. (2−1)+(7−4), and −1 with regard to the original amino acid pairs, i.e. (0–0)+(0−1). In this way we can compare the frequency difference in the amino acid pairs affected by mutations/variants.

Statistical inference

The actual and predicted frequencies can be compared as follows. Generally each of the 20 types of amino acid has a chance of 1/20 ($p$=0.05) to repeat once, and an amino acid pair has a chance of 1/400 ($p$=0.0025) to repeat once in the protein primary structure. In the case of human p53 protein, there are 45 prolines ("P"s), the most abundant amino acid, and 4 tryptophans ("W"s), the least amino acid. If the first amino acid is "P", then the chance of the second amino acid being "P" is 44/392 ($p$=0.1122>0.05). If the first amino acid is "W", then the chance of the second amino acid being "W" is 3/392 ($p$=0.0077<0.01). Thus, the chance of the first amino acid pair of "PP" is 45/393×44/392 ($p$=0.0129<0.05), the chance of second amino acid pair of "PP" is 43/391×42/390 ($p$=0.0118<0.05), and the chance of the third amino acid pair of "PP" is 41/389×40/388 ($p$=0.0109<0.05). If we consider the least frequent amino acid "W", the chance of first amino acid pair of "WW" is 4/393×3/392 ($p$=0.0001<0.001), and the chance of second amino acid pair of "WW" is 2/391×1/390 ($p$=0.0001<0.001). Therefore the probability would be less than 0.05 if difference between any amino acid pairs is equal to or larger than one.

## Results

### General information on amino acid pairs in human p53 protein

Of 400 kinds of theoretical amino acid pairs, 190 kinds are absent from human p53 protein including 88 randomly predictable and 102 randomly unpredictable. Consequently 392 amino acid pairs in human p53 protein include only 210 kinds of theoretical amino acid pairs (400−190=210), i.e. some amino acid pairs should appear more than once. Actually, of 392 amino acid pairs in human p53 protein, 108 kinds of theoretical amino acid pairs appear once, 59 kinds twice, 24 kinds three times, 11 kinds four times, 3 kinds five times, 2 kinds six times, 2 kinds seven times and 1 kind nine times. Of 210 kinds of theoretical amino acid pairs in 392 amino acid pairs in human p53 protein, 90 kinds are randomly predictable and 120 kinds are randomly unpredictable. Therefore, we would like to find how many human p53 mutations/variants occur in the 90 kinds of randomly predictable amino acid pairs and in the 120 kinds of randomly unpredictable amino acid pairs.

### Mutations/variants in human p53 protein randomly predictable and unpredictable present amino acid pairs

As mentioned in the Materials and methods section, a point mutant protein leads to two amino acid pairs being replaced by another two, and their actual frequencies can be smaller than, equal to or larger than the randomly predicable frequencies. Tables 1 and 2 detail the situations related to original and mutated amino acid pairs, respectively, and the relationship between their actual and randomly predicted frequencies.

Table 1 can be read as follows. The first column classifies the amino acid pairs into randomly predictable and unpredictable. The second and third columns show in which type of amino acid pairs the mutation/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 amino acid pairs I and II. The fourth and fifth columns indicate how many mutations/variants occur in amino acid pairs I and II, for example, 13 of 190 (6.84%) mutations/variants occur at amino acid pairs whose actual frequencies are equal to their predicted frequencies. The sixth column indicates the percentage of the 190 mutations/variants occurring at predictable and unpredictable amino acids.

**Table 1** Classification of original amino acid pairs induced by mutations/variants in human p53 protein

| | Pair I[a,b] | Pair II[a] | Mutations/variants | | Total (%) |
|---|---|---|---|---|---|
| | | | Number | % | |
| Predictable | AF=PF | AF=PF | 13 | 6.84 | 6.84 |
| Unpredictable | AF>PF | AF>PF | 89 | 46.84 | 93.16 |
| | AF>PF | AF=PF | 77 | 40.53 | |
| | AF>PF | AF<PF | 9 | 4.74 | |
| | AF<PF | AF=PF | 2 | 1.05 | |
| | AF<PF | AF<PF | 0 | 0 | |

[a] AF: actual frequency
[b] PF: predicted frequency

**Table 2** Classification of mutated amino acid pairs induced by mutations/variants in human p53 protein

| Pair I | Pair II | Mutations/variants | | Total (%) |
|---|---|---|---|---|
| | | Number | % | |
| AF=0, PF>0 | AF=0, PF>0 | 18† | 9.47 | 63.15 |
| AF=0, PF>0 | AF=PF=0 | 9† | 4.74 | |
| AF=0, PF>0 | AF=PF>0 | 19† | 10.00 | |
| AF=0, PF>0 | AF<PF, AF≠0 | 11† | 5.79 | |
| AF=0, PF>0 | AF>PF | 34† | 17.89 | |
| AF=PF=0 | AF=PF=0 | 10 | 5.26 | |
| AF=PF=0 | AF=PF>0 | 3 | 1.58 | |
| AF=PF=0 | AF<PF, AF≠0 | 1† | 0.53 | |
| AF=PF=0 | AF>PF | 15 | 7.89 | |
| AF<PF, AF≠0 | AF<PF, AF≠0 | 2† | 1.05 | 36.85 |
| AF<PF, AF≠0 | AF=PF>0 | 14† | 7.37 | |
| AF<PF, AF≠0 | AF>PF | 6† | 3.16 | |
| AF=PF>0 | AF=PF>0 | 19 | 10.00 | |
| AF>PF | AF>PF | 8 | 4.21 | |
| AF=PF>0 | AF>PF | 21 | 11.05 | |

Table 1 tells us that 93.16% of mutations/variants occur at randomly unpredictable amino acid pairs and 6.84% of mutations/variants occur in randomly predictable amino acid pairs. These results mean that 120 kinds of randomly unpredictable present amino acid pairs account for 93.16% of the mutations/variants in human p53 protein, whereas the 90 kinds of randomly predictable present amino acid pairs account for only 6.84%. These results strongly support our rationale that harmful mutations/variants are more likely to occur at randomly unpredictable present amino acid pairs, which therefore are more sensitive to mutations/variants.

When looking at the unpredictable pairs in Table 1, we find that the vast majority of these pairs are characterized

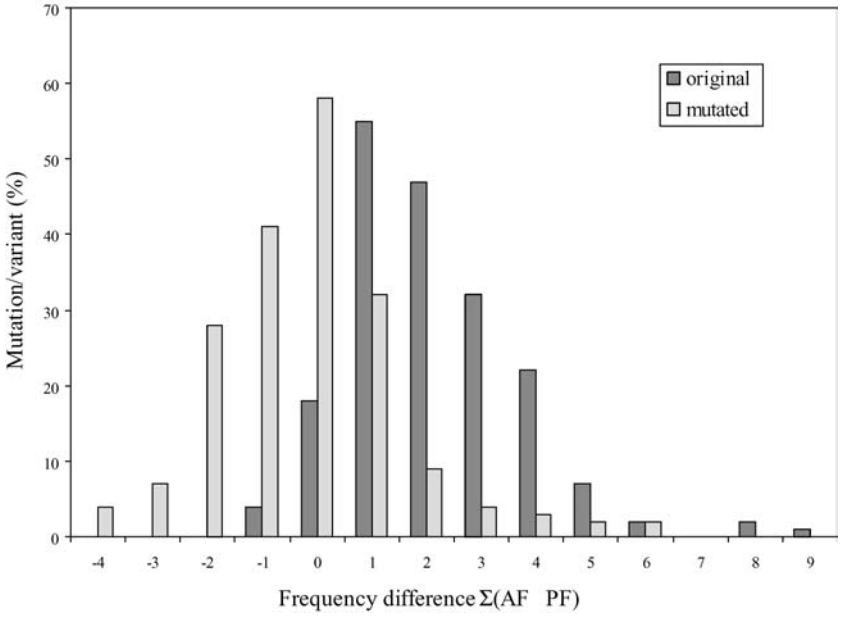


**Fig. 1** Frequency difference between mutated and original amino acid pairs induced by mutations/variants

by one or both original pairs whose actual frequencies are larger than their predicted frequency (the first three rows in unpredictable pairs). Comparing each mutation/variant, we find that the impact of mutations/variants is to narrow the difference between actual and predicted frequencies by reducing the actual frequency. This means that the mutations/variants lead to the construction of amino acid pairs to be randomly predictable. In other words, the mutations/variants lead to the construction of amino acid pairs that occur more easily naturally. Interestingly, no mutations/variants occur in the amino acid pairs, whose actual frequency is smaller than predicted frequency in both pairs. This suggests that it is difficult for mutations/variants to narrow the difference between actual and predicted frequencies by means of increasing the actual frequency. However, the way of reducing actual frequency would lead to the construction of amino acid pairs against the natural direction.

Table 2 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 mutations/variants that occur at amino acid pairs I and II and their percentages, the fifth column is the total of our classifications.

Table 2 shows that 63.15% of mutations/variants result in one or both mutated amino acid pairs which are absent in normal human p53 protein (AF=0). Table 2 also tells us that 59.47% of mutations/variants target one or both mutated amino acid pairs with their actual frequency smaller than their predicted frequency (indicated by † in Table 2). These phenomena indicate that the amino acid pairs in mutant p53 proteins are more randomly constructed.

## Frequency difference of amino acid pairs affected by mutations/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 190 mutations/variants, the mean±SD (ranging from −1 to 9) is 2.1±1.4 for the difference between actual and predicted frequencies in original amino acid pairs. This means that the mutations/variants occur in the amino acid pairs that appear more often than their predicted frequency. Meanwhile, the mean±SD (ranging from −4 to 6) is −0.2±0.4 for the difference between actual and predicted frequencies in the mutated amino acid pairs. This means that the mutated amino acid pairs are randomly constructed in the mutant p53 proteins, as their actual and predicted frequencies are about the same. A striking statistical difference is found between the mutated and original amino acid pairs ($P<0.0001$). Figure 1 shows the distribution of differences between actual and predicted frequencies.

Figure 2 displays the mean difference between actual and predicted frequencies with respect to different tumours. Generally the differences are remarkable large between mutated and original amino acid pairs in each tumour. However, no statistical significance is found due to the few cases documented in the databank.

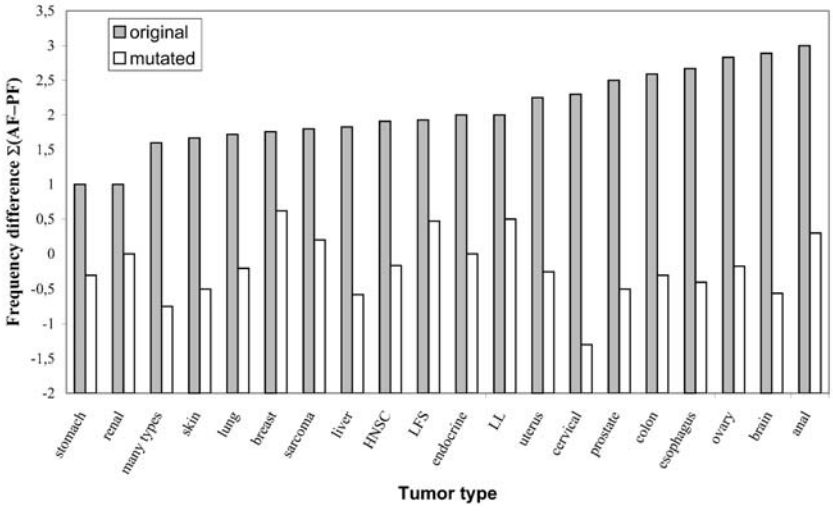


**Fig. 2** Mean difference between actual and predicted frequencies of mutated and original amino acid pairs induced by mutations/variants in different tumour types. HNSC: head and neck squamous cancer; L.L.: leukaemia and lymphoma; LFS: Li-Fraumeni syndrome

## Discussion

In this study we use the random approach to analyse the amino acid pairs in human p53 protein to determine which amino acid pairs are more sensitive to mutations/variants. The results confirm our hypothesis that the randomly unpredictable amino acid pairs are more sensitive to mutations/variants. This data-based theoretical analysis may provide a clue for preventing human p53 from mutations/variants, and throw light on the nature of p53 mutations/variants.

Based on our previous studies, our argument is that the functional amino acid pairs should be deliberately evolved, and thus the actual frequency should be different from the randomly predicted frequency. As the randomly predicted frequency is the highest chance for construction of amino acid pairs, it is important to find whether the mutation/variant leads to the actual frequency approaching the randomly predicted frequency. If so, we can understand that the human p53 protein has a natural trend to mutations/variants; if not, we can understand the human p53 protein does not have a natural trend to the mutation/variant.

In our another study, [11] we found that the human p53 protein is very unstable among p53 protein family across different species using a random distribution approach. In fact only human p53 can accommodate so many mutations/variants. This may be due to the unstable nature of human p53 protein.

With respect to randomly unpredictable absent and present amino acid pairs, we are interested in the difference between actual and predicted frequencies, because the randomly predictable absent and present frequency represents the naturally easiest 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.

## Supplementary material

Data for Figs. 1 and 2 are available as supplementary material.

## References

1. Guimaraes DP, Hainaut P (2002) Biochimie 84:83–93
2. Rideout WM, Coetzee GA, Olumi AF, Jones PA (1990) Science 249:1288–1290
3. Hainaut P, Pfeifer GP (2001) Carcinogenesis 22:367–374
4. Montesano R, Hainaut P, Wild CP (1997) J Natl Cancer Inst 89:1844–1851
5. Forrester K, Lupold SE, Ott VL, Chay CH, Band V, Wang XW, Harris CC (1995) Oncogene 10:2103–2111
6. Ory K, Legros Y, Auguin C, Soussi T (1994) EMBO J 13:3496–3504
7. Aas T, Borresen AL, Geisler S, Smith-Sorensen B, Johnsen H, Varhaug JE, Akslen LA, Lonning PE (1996) Nat Med 2:811–814
8. Wu G (2000) Human Exp Toxicol 19:535–539
9. Wu G (2000) J Biochem Mol Biol Biophys 4:179–85
10. Wu G, Yan SM (2002) Mol Biol Today 3:31–37
11. Wu G, Yan SM (2002) J Mol Model 8:191–198
12. Wu G, Yan SM (2001) Biomol Engi 18:23–27
13. Bairoch A, Apweiler R (2000) Nucleic Acids Res 28:45–48
14. Feller W (1968) An introduction to probability theory and its applications, 3rd edn, vol I. Wiley, New York