J Mol Model (2002) 8:191–198
DOI 10.1007/s00894-002-0087-8

ORIGINAL PAPER

Guang Wu · Shao-Min Yan

Analysis of distributions of amino acids in the primary structure of tumor suppressor p53 family according to the random mechanism

Received: 22 March 2002 / Accepted: 19 April 2002 / Published online: 29 May 2002

Abstract It is well known that the evolutionary process leads to the majority of amino acids clustering in some regions rather than being homogenously distributed along a protein. Among numerous factors affecting the evolutionary process is chance, whose impact therefore should be present in a protein primary structure. The issue of how to measure the random distribution of amino acids in a primary structure is of importance for the understanding of protein structure and functions. In this study, we use the random principle as a tool to analyze and compare the distributions of amino acids in the primary structure of the p53 protein family. The results, for example, show that the amino acids are distributed more randomly in mouse p53 and less randomly in common tree shrew p53, the distribution ranks of amino acids are relatively lower in the functional regions (about 0.5 on average) than in the whole sequences (about 1.2 on average) except for mouse p53. From the probabilistic distribution view, the composition of human p53 is relatively stable in the functional regions rather than in the whole sequence, which may suggest one of the potential effects on the mutations inducing human cancers. In general, we can use the distribution probability to present quantitatively a type of distribution of amino acids in a protein, to compare quantitatively the magnitude of clusters between different proteins and to track the effect of chance on the evolutionary process.

G. Wu (✉)
Laboratoire de Toxicocinétique et Pharmacocinétique,
Faculté de Pharmacie,
Université de la Méditerranée Aix-Marseille II, Marseille, France
e-mail: guang.wu@pharma.novartis.com
Tel.: +41 061 696 7746, Fax: +41 061 696 6992

S.-M. Yan
Cattedra di Anatomia Patologica,
Dipartimento di Ricerche Mediche e Morfologiche,
Facoltà di Medicina e Chirurgia, Università degli Studi di Udine,
Udine, Italy

G. Wu
ICPP, WKL-135.1.16, Novartis Pharma AG, 4002 Basel,
Switzerland

Keywords Amino acids · Probability · p53 family

Introduction

An intriguing phenomenon is that the majority of amino acids cluster in some regions rather than being homogenously distributed along the primary structure of a protein, although this could simply be due to natural selection. Generally, the distribution of amino acids in a protein can be affected by many factors, such as adaptation, chance and history: (i) adaptation refers to the current selective benefit of having a particular amino acid configuration; (ii) chance affects amino acid distributions by mutation and the fixation of mutations, which is most easily seen to occur when selective differences between different amino acid distributions are non-existent (neutral evolution); and (iii) the effect of history can be seen as the starting point for subsequent adaptation.

With respect to the impact of chance on the distribution of amino acids in a protein, it is important to represent a type of distribution of amino acids in a protein quantitatively. If so, we can compare the magnitude of clusters in amino acids between different proteins quantitatively and track the impact of chance on the evolutionary process.

In the last few years, we have used two random approaches to analyze the primary structure of different proteins, [1, 2, 3, 4] which in fact provide us with ways to present and compare quantitatively the impact of chance on the random frequency of composition of amino acid pairs and triplets in the primary structure and the impact of chance on the distribution of amino acids and pairs in the primary structure of a protein. Obviously it is more important and meaningful to use our methods to analyze a protein family; in this manner we can get not only a general view of the impact of chance on the primary structure of proteins across different species, more important we can gain deep insight into protein structure and function.

In this study, we further develop our method to analyze the distributions of amino acids in all the proteins from the tumor suppressor p53 family in order to obtain

Table 1 All possible distributions of six 'F's in six parts

Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Distribution pattern
F	F	F	F	F	F	1, 1, 1, 1, 1, 1
	F	F	F	F	FF	0, 1, 1, 1, 1, 2
		F	F	FF	FF	0, 0, 1, 1, 2, 2
			FF	FF	FF	0, 0, 0, 2, 2, 2
		F	F	F	FFF	0, 0, 1, 1, 1, 3
			F	FF	FFF	0, 0, 0, 1, 2, 3
				FFF	FFF	0, 0, 0, 0, 3, 3
			F	F	FFFF	0, 0, 0, 1, 1, 4
				FF	FFFF	0, 0, 0, 0, 2, 4
				F	FFFFF	0, 0, 0, 0, 1, 5
					FFFFFF	0, 0, 0, 0, 0, 6

more understanding of p53 from the viewpoint of the molecular modeling, because p53 acts as a tumor suppressor in many tumor types by inducing growth arrest or apoptosis.

With the distribution probability as a tool to measure the impact of chance on protein primary structure, we can compare the magnitude of clusters in amino acids quantitatively between proteins across the p53 family. Moreover, we can compare the distributions of amino acids in different functional regions to gain a deeper insight into the p53 family, as there are five functional regions, i.e. (i) transcription activation, (ii) DNA-binding, (iii) nuclear localization signal, (iv) oligomerization and (v) repression of DNA-binding.

Materials and methods

Tumor suppressor p53 protein family

Currently the p53 family includes 28 proteins with complete sequences in the Swiss-Protein data bank (http://srs.ebi.ac.uk). [5] These proteins are from (i) bovine (access number Q29628), [6, 7] (ii) cat (access number P41685), [8, 9] (iii) chicken (access number P10360), [10] (iv) dog (access number Q29537), [11, 12] (v) African clawed frog (access number P07193), [13, 14] (vi) Chinese hamster (access number O09185), [15] (vii) golden hamster (access number Q00366), [16] (viii) human (access number P04637), [17, 18] (ix) cynomolgus monkey (access number P56423), [19] (x) green monkey (access number P13481), [20] (xi) rhesus monkey (access number P56424), [21] (xii) mouse (access number P02340), [22, 23] (xiii) guinea pig (access number Q9WUR6), [24] (xiv) pig (access number Q9TUB2), [25] (xv) rabbit (access number Q95330), [26] (xvi) rat (access number P10361), [27, 28] (xvii) sheep (access number P51664), [29] (xviii) common tree shrew (access number Q9TTA1), [30] (xix) woodchuck (access number O36006), [31] (xx) barbel (access number Q9W678), [32] (xxi) channel catfish (access number O93379), [33] (xxii) Medaka fish (access number P79820), [34] (xxiii) European flounder (access number O12946), [35] (xxiv) xiphophorus helleri (access number O57538), [36] (xxv) southern platyfish (access number Q92143), [36] (xxvi) Congo puffer (access number Q9W679), [32] (xxvii) rainbow trout (access number P25035) [37] and (xxviii) zebra fish (access number P79734). [38]

Calculation of distribution probability

The calculation method has previously been published in this journal; [4] we therefore briefly describe it through an example. There are six phenylalanines (F) among 363 amino acids from the African clawed frog p53 protein. We can group this protein into six parts, each contains about 61 amino acids (363/6=60.5), and then calculate the distribution probability according to the calculation of occupancy problems of subpopulations and partitions. [39] The distribution probability is $r!/(q_0!\times q_1!\times...\times q_n!)\times r!/(r_1!\times r_2!\times...\times r_n!)\times n^{-r}$ for any type of distribution.

In the equation,! is the factorial function. r is the number of a type of amino acid, we have r=6 for 'F's because there are six 'F's in African clawed frog. n is the number of grouped parts in which amino acids can distribute, thus n=6 for 'F's. q is the number of parts with the same number of amino acids, when six 'F's appear in each of six parts in our example, we have q_0=0, q_1=6, q_2=0, q_3=0, q_4=0, q_5=0 and q_6=0, i.e. 0 part has 0 'F', six parts have 1 'F', 0 part has two 'F's, 0 part has three 'F's, 0 part has four 'F's, 0 part has five 'F's and 0 part has six 'F's. r_1, r_2,..., r_n are the number of amino acids in parts 1, 2,..., n. When six 'F's appear in each of six parts, we have r_1=1, r_2=1, r_3=1, r_4=1, r_5=1 and r_6=1.

Tables 1 and 2 detail the calculation using this equation with respect to the distributions of six amino acids in six parts. The first six columns in Table 1 show that the protein is divided into six parts, and the first six cells in each row show a possible configuration of amino acids; the seventh column is the numeric presentation of each configuration; in Table 2 the first column shows the details of the calculation of the distribution probability (the first and second parentheses correspond to q and r in the equation); the second column is the distribution probability and the third column is the distribution rank.

Ranking distribution probability

In order to avoid too many decimal values, we rank the distribution probabilities according to a descending

Table 2 Calculation of distribution probability and rank with respect to distribution patterns in Table 1

Calculation of probability ($6!\times6!\times6^{-6}$ divided by)	Probability	Rank
(0!×6!×0!×0!×0!×0!×0!)×(1!×1!×1!×1!×1!×1!)	0.015432	5
(1!×4!×1!×0!×0!×0!×0!)×(0!×1!×1!×1!×1!×2!)	0.231481	2
(2!×2!×2!×0!×0!×0!×0!)×(0!×0!×1!×1!×2!×2!)	0.347222	1
(3!×0!×3!×0!×0!×0!×0!)×(0!×0!×0!×2!×2!×2!)	0.038580	4
(2!×3!×0!×1!×0!×0!×0!)×(0!×0!×1!×1!×1!×3!)	0.154321	3
(3!×1!×1!×1!×0!×0!×0!)×(0!×0!×0!×1!×2!×3!)	0.154321	3
(4!×0!×0!×2!×0!×0!×0!)×(0!×0!×0!×0!×3!×3!)	0.006430	7
(3!×2!×0!×0!×1!×0!×0!)×(0!×0!×0!×1!×1!×4!)	0.038580	4
(4!×0!×1!×0!×1!×0!×0!)×(0!×0!×0!×0!×2!×4!)	0.009645	6
(4!×1!×0!×0!×0!×1!×0!)×(0!×0!×0!×0!×1!×5!)	0.003858	8
(5!×0!×0!×0!×0!×0!×1!)×(0!×0!×0!×0!×0!×6!)	0.000129	9

order; thus the highest distribution probability is ranked as one. Again in our examples, there are 11 possible distributions for six amino acids in six parts in Table 1, but we rank them only to nine because the same probability can be calculated from different distributions (ranks 3 and 4 in Table 2). In general, the smaller the rank, the larger the distribution probability. Although there are many possible distributions for a type of amino acid in a protein (such as 11 possible distributions for six amino acids in Table 1), a protein adopts only one possible distribution during its evolutionary process. Therefore there is only one distribution probability/rank for each type of amino acid and a maximum of 20 types of distribution probability/rank in a protein.

Distribution rank in primary structure

As the composed numbers of amino acids are different from one another even in the same protein family, we standardize the distribution rank in a protein using the distribution ranks per amino acid and for each type of amino acid. These can be calculated by dividing the sum of distribution ranks by the number of amino acids and the sum of distribution rank of each type of amino acid by the number of the corresponding type of amino acids. Using human p53 as an example, this protein contains 393 amino acids and the sum of distribution ranks for all 20 types of amino acids is 685; thus the distribution rank per amino acid is 685/393=1.734. Similarly, there are 45 prolines ('P's) in human p53 and the distribution rank for 'P's is 121; thus the distribution rank per each 'P' is 121/45=2.69. In this manner, we can compare the distribution ranks on primary structure across a protein family.

Comparison between proteins and between different functional regions

As each type of amino acid has a specific distribution probability/rank in a protein, this probability is not similar to the probability obtained from statistical inference, we therefore cannot treat the distribution probability in the sense of statistical significance. However, we can understand the likelihood of occurrence of a type of distribution, for example, the distribution of 'F', 'F', 'FF' and 'FF' is 22.5 times (0.347222/0.015432) more likely to occur than the distribution of 'F', 'F', 'F', 'F', 'F' and 'F' (Tables 1 and 2).

Comparison of distribution probability with homogenous distribution probability

As each type of amino acid has a certain distribution probability in a protein, we can compare the distribution probability with the homogenous distribution probability in order to determine whether a distribution of amino acids is different from a homogenous distribution. For example, the probability of homogenous distribution of 22 amino acids in a protein is 3.2921×10^{-9}, while the real distribution probability of 22 threonines ('T's) is 0.058542 in pig p53, so the chance of homogenous distribution of 'T's is quite small.

Results

Figure 1 shows the distribution rank per amino acid in proteins across the p53 family. It can be seen that mouse p53 has the lowest distribution rank, whereas the p53 from common tree shrew has the highest distribution rank, whose value increases 3.8-fold compared to that of mouse p53. This difference implies that chance has more impact on the amino acid distribution in mouse p53 and less impact in common tree shrew p53. The results also suggest that the distribution of amino acids is relatively unpredictable in common tree shrew p53.

Figure 2 shows the distribution rank per amino acid in different functional regions across the p53 family. In the transcription activation region, the lowest distribution rank is calculated from rabbit p53, indicating that chance has more impact on its amino acid distribution in this region, while the highest distribution rank is calculated from channel catfish and zebra fish p53 (about 75% higher than that of rabbit p53), showing that chance has less impact on the amino acid distribution in this region.

In the DNA-binding region, human and common tree shrew have the lowest value of distribution rank, which

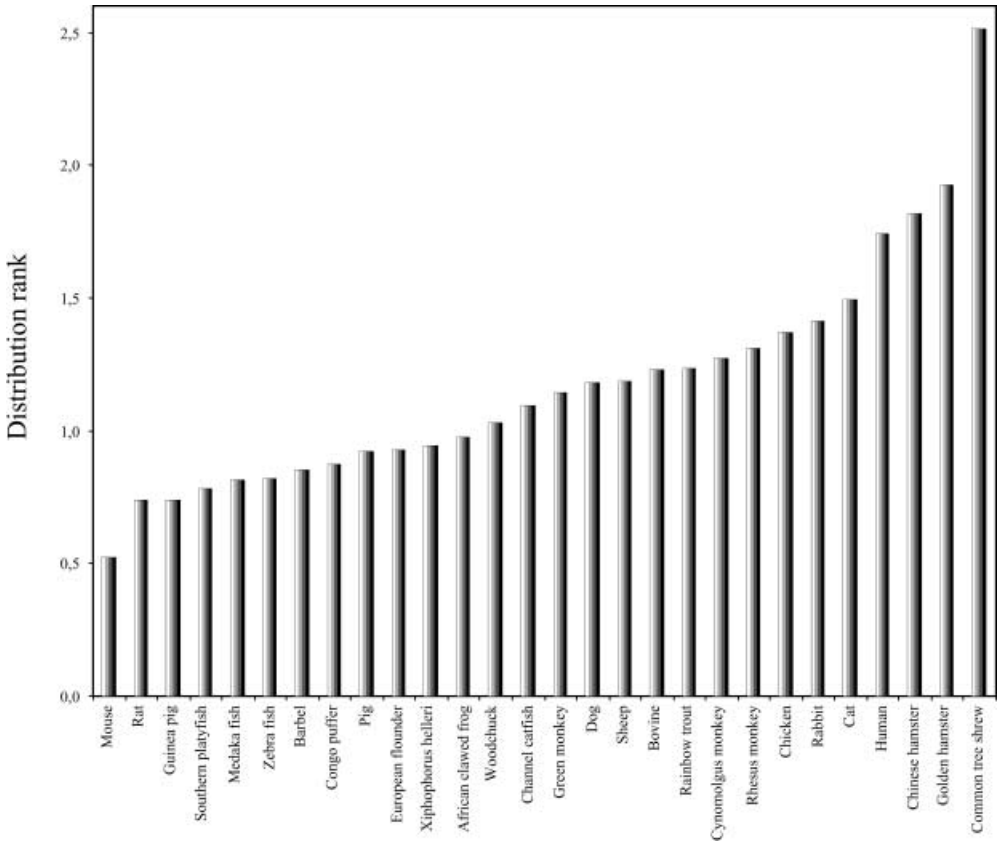

Fig. 1 Distribution rank per amino acid in the whole sequence across the p53 family

means that their distribution is probabilistically simple. By contrast, the highest distribution rank is from channel catfish p53 (about 1.3-fold higher), excluding random distributions of amino acids in this region.

In the nuclear localization signal region, the lowest distribution rank comes from cat p53, while the highest value is calculated from the p53 of barbel and European flounder, which is increased by 77% over that of cat p53.

In the oligomerization region the distribution rank is similar throughout the whole family and the biggest difference is only 40%. This value is higher in European flounder, chicken and common tree shrew and lower in barbel, Congo puffer, Medaka fish and woodchuck.

In the repression of the DNA-binding region, the bovine p53 shows the lowest value of distribution rank, whereas the mouse and rat p53s have the highest value (75% higher). These results indicate that the distribution of amino acids is more random in the former but not in the latter.

Figure 3 shows the distribution rank for each type of amino acid in the proteins across the p53 family. Although this measure is different from species to species and different from one type of amino acid to another, we can see that the distribution probability shifts from homogeneity to heterogeneity. The change in this value is highly consistent with the distribution rank per amino acid in the whole sequence. Among the p53 family, about half of the types of amino acid give relatively low values of distribution rank, indicating that

Table 3 Homogenous distribution probability with respect to the different numbers of amino acids

Number of amino acids	Probability	Number of amino acids	Probability
2	0.5000	3	0.2222
4	0.0938	5	0.0384
6	0.0154	7	0.0061
8	0.0024	9	0.0009
10	0.0004	11	0.0001
12	5.3723×10^{-5}	13	2.0560×10^{-5}
14	7.8454×10^{-6}	15	2.9863×10^{-6}
16	1.1342×10^{-6}	17	4.2997×10^{-7}
18	1.6272×10^{-7}	19	6.1486×10^{-8}
20	2.3202×10^{-8}	21	8.7446×10^{-9}
22	3.2921×10^{-9}	23	1.2381×10^{-9}
24	4.6520×10^{-10}	25	1.7464×10^{-10}
26	6.5511×10^{-11}	27	2.4556×10^{-11}
28	9.1985×10^{-12}	29	3.4435×10^{-12}
30	1.2883×10^{-12}	31	4.8174×10^{-13}
32	1.8004×10^{-13}	33	6.7255×10^{-14}
34	2.5112×10^{-14}	35	9.3724×10^{-15}
36	3.4966×10^{-15}	37	1.3040×10^{-15}
38	4.8612×10^{-16}	39	1.8116×10^{-16}
40	6.7491×10^{-17}	41	2.5136×10^{-17}
42	9.3585×10^{-18}	43	3.4834×10^{-18}
44	1.2962×10^{-18}	45	4.8222×10^{-19}
46	1.7935×10^{-19}	47	6.6691×10^{-20}
48	2.4793×10^{-20}	49	9.2150×10^{-21}
50	3.4243×10^{-21}		

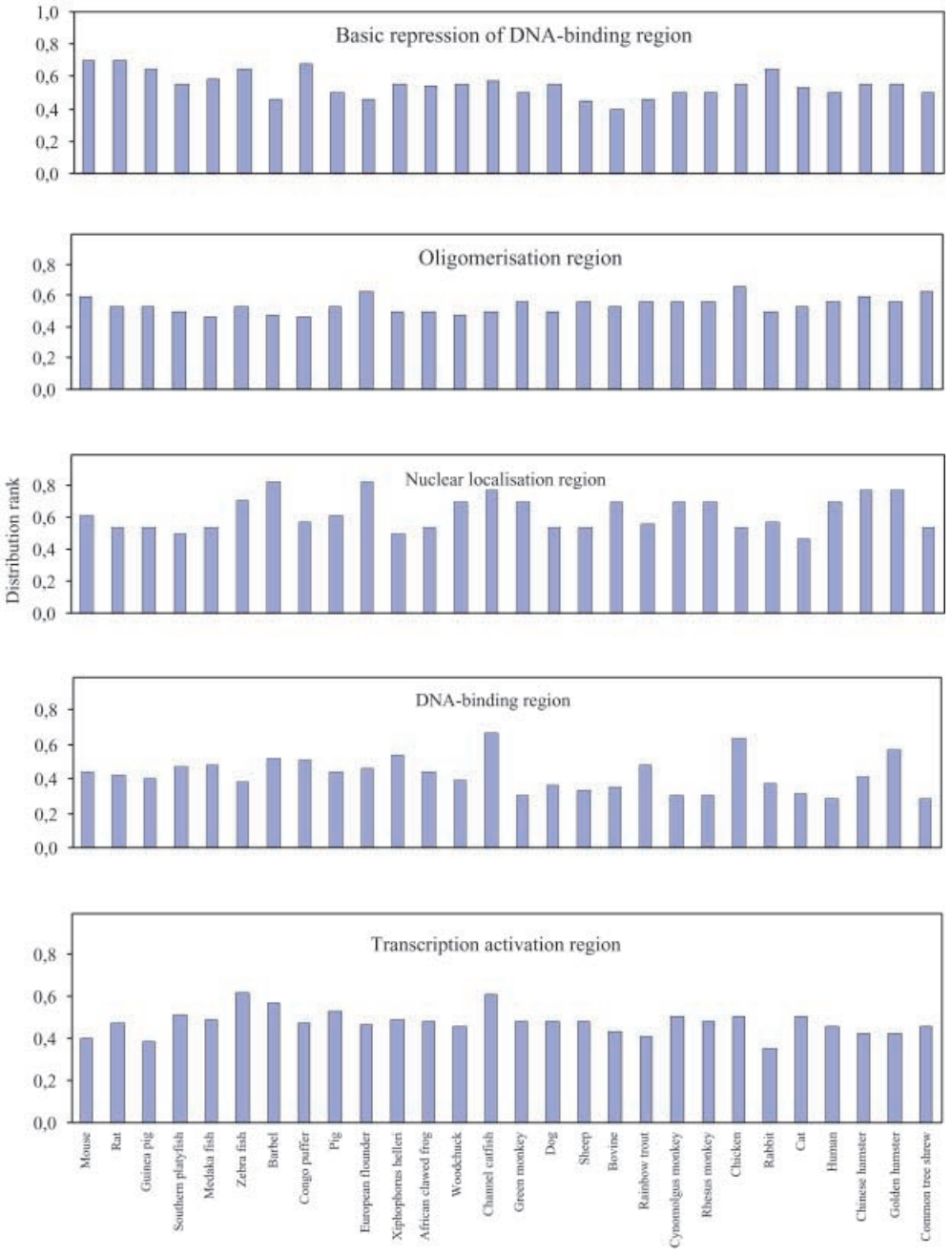

Fig. 2 Distribution rank per amino acid in different functional regions across the p53 family

these amino acids distribute more randomly in the sequences, whereas the amino acids with high values of distribution rank imply that their distribution has been less affected by chance.

Table 3 shows the homogenous distribution probability with respect to the different numbers of amino acids. It can be seen that the homogenous distribution probability is very low; thus it is not easy for a protein to adopt the homogenous distribution for a certain type of amino acid.

We can compare the distribution probability among different types of amino acids with the use of the current approach (Fig. 3). Some types of amino acid, such as asparagines (N), have a very low rank value, which indicates that the distribution of these amino acids is highly random in the p53 family. The amino acids with a high rank value are not randomly distributed and result more likely from a deliberate evolutionary process, for example, alanines (A) in common tree shrew may link to some special functional units. Also, the homogenous distribution of a type of amino acid appears with a very low probability (Table 3). Thus, the tendency for amino acids to cluster in the protein primary structure results from the necessity of both function and randomness.

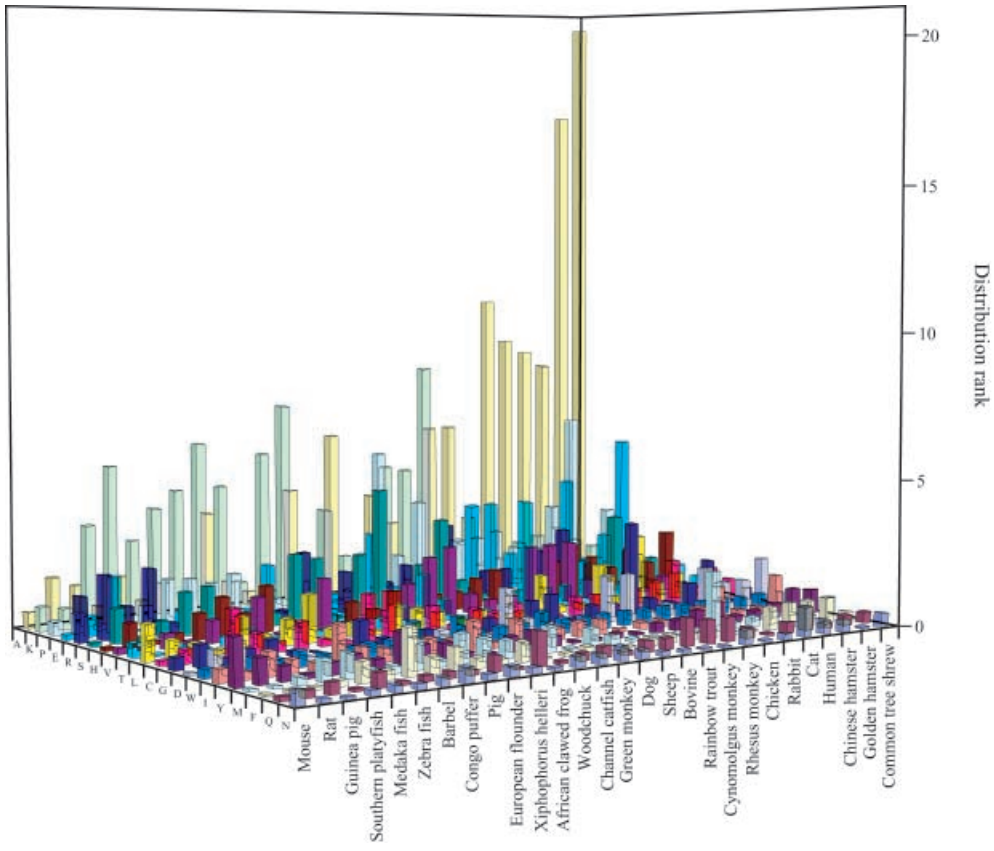

Fig. 3 Distribution rank per each type of amino acid in the proteins across the p53 family

Discussion

With the advance in experimental and theoretical studies, we always can gain new insights from the primary structure of proteins. The random analysis can throw light on the underlying reasoning for the primary structure of proteins, not only because pure chance is now considered to lie at the very heart of nature, [40] but also because it can provide a quantitative measure to compare the magnitude of clusters of amino acids in a protein.

The primary structure of proteins is the basis for higher level structures and protein functions; thus the primary structure always provides the basis for studying and modeling (i) the patterns of amino acid composition, (ii) the patterns of natural and artificial mutations, (iii) the similarity within a protein family, (iv) the similarity between protein families, (v) the mechanism for construction of higher level structures, (vi) the topological basis for higher level structures, etc. The patterns of amino acid composition and mutations are archived via experimental methods and annotation. [41] The most popular analysis of the similarity within a protein family and between protein families is archived via multiple sequence comparisons and alignments using standard software, for example, BlastP. [42] There are several other approaches for similarity analysis, such as fast Fourier transform, [43] the statistical approach, [44] linguistic approaches [45, 46] and pattern graphs. [47] Particularly worth mentioning are two series of studies using statistical methods to analyze and predict locations of amino acids, distribution, structure, etc. [48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58]

However, the general difference between previous approaches and ours is that the previous studies concentrated on the description of primary structures while we are trying to explain the reason for the construction of primary structure. For example, the speed of evolution in a species must catch up with the environmental changes, otherwise the species would die. This requires the natural selection to be efficient. For example, the synthesis of a protein should be less energy- and time-consuming. The choice of the amino acid sequences with a high probability of occurrence is certainly a way to be efficient, which can be viewed as the effect of chance on the evolutionary process. However, at this stage we are still not aware whether or not nature does so, although we observed that the distribution probabilities for several types of amino acids are at or very near to the probabilistically simplest distribution in this study. For example, five types of amino acids occur with the highest probability in mouse and rat p53. There is the possibility that the distributions with a high distribution probability may not be deliberately evolved and conserved, whereas the distributions with a low distribution probability should be deliberately evolved and conserved, because nature is clever enough to spend only the necessary energy and time during evolution.

Although the distribution rank of amino acids varies from species to species, our results demonstrate that this value is relatively lower in the functional regions (about 0.5 on average) than in the whole sequences (about 1.2 on average), except for mouse p53. Also, this measure shows less difference between different species in all of the functional regions as opposed to the whole sequence. Therefore, the functional regions (especially the DNA-binding region) are not only conserved through the evolutionary process of the species, but also arranged in a probabilistically simple way, which guarantees the functional base of the p53 protein. Taking five functional regions into account, the DNA-binding region is larger (about 190 amino acids) than the others. Mutations in this region may lead to p53 dysfunction and induce cancers, which is true in humans as the majority of such mutations are found in the DNA-binding region. [59, 60]

Subsequently, it is natural to ask the question why there are so many mutations in human p53 that lead to various cancers in humans. The current results may give information on this issue. In mouse p53, the distribution of amino acids is probabilistically simple, as the distribution rank is about the same in the whole sequence and in different functional regions as well, which may suggest that the primary structure of mouse p53 is one of the more stable among the p53 family. By contrast, the distribution probability of amino acids in human p53 shows dual characteristics: the distribution rank is lower in the functional regions but higher in the whole sequence (more than two times higher than that in mouse p53). From the distribution probabilistic viewpoint, the composition of human p53 is relatively more stable in the functional regions than in the whole sequence. This contradictory feature may contribute to one of the potential effects on mutation.

We did not compare our results with those obtained from multiple sequence comparisons and alignments, not only because the multiple sequence comparisons and alignments are a routine approach easily used for all the proteins in a databank, but more important is that each approach has its own advantage and addresses specific issues. In general, multiple sequence comparisons and alignments are focused on finding and ranking the similarity among all the proteins in a databank, while our approach tries to use the random mechanism as the cause to explain the distribution of amino acids.

We also did not use the information entropy to express our results, because the information entropy does not deal with a particular distribution of amino acids along a protein, although it deals with the numbers of amino acids.

With our first random approach to analyzing the primary structure of mouse and sheep p53, [1, 2] we attempted to understand why a certain type of amino acid rather than the other types is associated with a type of amino acid in amino acid pairs and triplets. Our first approach [3] and the approach in this study are attempts to explore the quantitative tools to measure, compare and explain the primary structures of proteins.

References

1. Wu G (2000) Human Exp Toxicol 19:535–539
2. Wu G (2000) J Biochem Mol Biol Biophys 4:179–185
3. Wu G, Yan SM (2001) J Mol Model 5:120–124
4. Wu G, Yan SM (2001) J Mol Model 7:318–323
5. Bairoch A, Apweiler R (1999) Nucleic Acids Res 27:49–54
6. Dequiedt F, Kettmann R, Burny A, Willems L (1995) DNA Seq 5:261–264
7. Komori H, Ishiguro N, Horiuchi M, Shinagawa M, Aida Y (1996) Vet Immunol Immunopathol 52:53–63
8. Okuda M, Umeda A, Matsumoto Y, Momoi Y, Watari T, Goitsuka R, O'Brien SJ, Tsujimoto H, Hasegawa A (1993) J Vet Med Sci 55:801–805
9. Okuda M, Umeda A, Sakai T, Ohashi T, Momoi Y, Youn HY, Watari T, Goitsuka R, Tsujimoto H, Hasegawa A (1994) Int J Cancer 58:602–607
10. Soussi T (1988) Nucleic Acids Res 16:11383
11. Kraegel SA, Pazzi KA, Madewell BR (1995) Cancer Lett 92:181–186
12. Veldhoen N, Milner J (1998) Oncogene 16:1077–1084
13. Soussi T, de Fromentel CC, Mechali M, May P Kress M (1987) Oncogene 1:71–78
14. Hoever M, Clement JH, Wedlich D, Montenarh M, Knoechel W (1994) Oncogene 9:109
15. Lee H, Larner JM, Hamlin JL (1997) Gene 184:177–183
16. Legros Y, McIntyre P, Soussi T (1992) Gene 112:247–250
17. Zakut-Houri R, Bienz-Tadmor B, Givol D, Oren M (1985) EMBO J 4:1251–1255
18. Lamb P, Crawford L (1986) Mol Cell Biol 6:1379–1385
19. Khan MA, Hansen C, Welsh JA, Bennett WP (1996) Submitted to the EMBL/GenBank/DDBJ databases
20. Rigaudy P, Eckhardt W (1989) Nucleic Acids Res 17:8375–8375
21. Kay HD, Mountjoy CP, Wu G, Cornish KG, Smith LJ (1994) Gene 138:223–226
22. Zakut-Houri R, Oren M, Bienz B, Lavie V, Hazum S, Givol D (1983) Nature 306:594–597
23. Bienz B, Zakut-Houri R, Givol D, Oren M (1984) EMBO J 3:2179–2183
24. D'erchia AM, Pesole G, Tullo A, Saccone C, Sbisa E (1999) Genomics 58:50–64
25. Burr PD, Argyle DJ, Reid SWJ, Nasir L (1998) Submitted to the EMBL/GenBank/DDBJ databases
26. le Goas F, May P, Ronco P, Caron de Fromentel C (1997) Gene 185:169–173
27. Soussi T (1988) Nucleic Acids Res 16:11384–11384
28. Hulla JE, Schneider RP (1993) Nucleic Acids Res 21:713–717
29. Dequiedt F, Kettmann R, Burny A, Willems L (1995) DNA Seq 5:255–259
30. Park U, Lee Y (1999) Submitted to the EMBL/GenBank/DDBJ databases
31. Feitelson MA, Ranganathan PN, Clayton MM, Zhang SM (1997) Oncogene 15:327–336
32. Bhaskaran A, May D, Rand-Weaver M, Tyler CR (1998) Submitted to the EMBL/GenBank/DDBJ databases
33. Luft JC, Bengten E, Clem LW, Miller NW, Wilson MR (1998) Comput Biochem Physiol, B 120:675–682
34. Krause MK, Rhodes LD, van Beneden RJ (1997) Gene 189:101–106
35. Cachot J, Galgani F, Vincent F (1998) Comput Biochem Physiol, B 121:235–242
36. Kazianis S, Gan L, Della Coletta L, Santi B, Morizot DC, Nairn RS (1998) Gene 212:31–38
37. de Fromentel CC, Padkel F, Chapus A, Baney C, May P, Soussi T (1992) Gene 112:241–245
38. Cheng R, Ford BL, O'Neal PE, Mathews CZ, Bradford CS, Thongtan T, Barnes DW, Hendricks JD, Bailey GS (1997) Mol Mar Biol Biotechnol 6:88–97
39. Feller W (1968) An introduction to probability theory and its applications, 3rd edn, vol I. Wiley, New York, pp 38–40

40. Everitt BS (1999) Chance rules: an informal guide to probability, risk, and statistics. Springer, Berlin Heidelberg New York
41. Barker WC, Garavelli JS, Hou Z, Huang H, Ledley RS, McGarvey PB, Mewes HW, Orcutt BC, Pfeiffer F, Tsugita A, Vinayaka CR, Xiao C, Yeh LS, Wu C (2001) Nucleic Acids Res 29:29–32
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) J Mol Biol 215:403–410
43. Benson DC (1990) Nucleic Acids Res 18:6305–6310
44. Karlin S, Bucher P, Brendel V, Altschul SF (1991) Annu Rev Biophys Biophys Chem 20:175–203
45. Popov O, Segal DM, Trifonov EN (1996) Biosystems 38:65–74
46. Searls DB (1997) Comput Appl Biosci 13:333-344
47. Jonassen I (1997) Comput Appl Biosci 13:509–522
48. Tanaka S, Scheraga HA (1975) Macromolecules 8:623–631
49. Tanaka S, Scheraga HA (1976) Macromolecules 9:142–159
50. Tanaka S, Scheraga HA (1976) Macromolecules 9:159–167
51. Tanaka S, Scheraga HA (1976) Macromolecules 9:168–182
52. Tanaka S, Scheraga HA (1976) Macromolecules 9:812–833
53. Tanaka S, Scheraga HA (1977) Macromolecules 10:9–20
54. Tanaka S, Scheraga HA (1977) Macromolecules 10:305–316
55. Tanaka S, Scheraga HA (1977) Proc Natl Acad Sci USA 72:1320–1323
56. Nishikawa K, Kubota Y, Ooi T (1983) J Biochem 94:997–1007
57. Nishikawa K, Kubota Y, Ooi T (1983) J Biochem 94:991–995
58. Nishikawa K, Ooi T (1986) J Biochem 100:1043–1047
59. Hoolstein M, Sidransky D, Vogelstein B, Harris CC (1991) Science 253:49–53
60. de Vries EMG, Ricke DO, de Vries TN, Hartmann A, Blaszyk H, Liao D, Soussi T, Kovach JS, Sommer SS (1996) Hum Mutat 7:202–213