

Random Analysis of Presence and Absence of Two- and Three-Amino-Acid Sequences and Distributions of Amino Acids, Two- and Three-Amino-Acid Sequences in Bovine p53 Protein

Guang Wu^{1,2,*} and Shaomin Yan³

¹Laboratoire de Toxicocinétique et Pharmacocinétique, Faculté de Pharmacie, Université de la Méditerranée Aix-Marseille II, Marseille, France

²Current Address: DMPK, WKL-135.1.16, Novartis Pharma AG, CH-4002 Basel, Switzerland

³Cattedra di Anatomia Patologica, Facoltà di Medicina e Chirurgia, Università degli Studi di Udine, Udine, Italy

Abstract

In this study we use five probabilistic procedures to analyse the bovine p53 protein. (1) We count each kind of two-, three- and multi-amino-acid sequences along bovine p53 protein from one terminal to the other and calculate their frequencies and probabilities. (2) We compare the amino-acid sequences in bovine p53 protein with the theoretical amino-acid sequences and determine which theoretical amino-acid sequences are present and absent. (3) We use the random principle to predict the frequencies of presence and absence of amino-acid sequences in bovine p53 protein from its amino acid composition and compare the predicted frequencies with the counted frequencies. (4) We use the random principle to predict the probability that an amino acid follows a preceding amino acid and compare the predicted probabilities with the probabilities occurred in bovine p53 protein. (5) We use the random principle to predict the distributions of amino acids, two- and three-amino-acid sequences in bovine p53 protein and compare the predicted distributions with the measured distributions.

Introduction

The relationship between various p53 proteins and their functions has been the objective of numerous experimental and theoretical studies. In this study, we use the following five probabilistic procedures to analyse the bovine p53 protein.

(1) We count each kind of two-, three- and multi-amino-acid sequences along bovine p53 protein from one terminal to the other and calculate their frequencies and

probabilities. The bovine p53 protein is composed of 386 amino acids (Dequiedt et al., 1995; Komori et al., 1996), we count the first and second amino acids as a two-amino-acid sequence, the second and third as another two-amino-acid sequence, the third and fourth, until the 385th and 386th, thus there are 385 two-amino-acid sequences. Furthermore, we count the first, second and third amino acids as a three-amino-acid sequence, the second, third and fourth as another three-amino-acid sequence, until the 384th, 385th and 386th, thus there are 384 three-amino-acid sequences. These considerations can be extended to other multi-amino-acid sequences. The rationale for adopting such a treatment of amino-acid sequences is because it is not yet known where a potential functional region begins and finishes, how non-functional regions mix with functional regions and how many amino acids a potential functional region has.

(2) We compare the amino-acid sequences in bovine p53 protein with the theoretical amino-acid sequences and determine which theoretical amino-acid sequences are present and absent. Ideally two amino acids in a two-amino-acid sequence can randomly be constructed from any one of 20 amino acids, thus there are 400 (20^2) kinds of theoretical two-amino-acid sequences (combinations). There are 385 two-amino-acid sequences in bovine p53 protein, if there was no repetition in them, there would be 385 kinds of two-amino-acid sequences, therefore one of the 400 kinds of theoretical two-amino-acid sequences would be expected to appear about 0.963 times ($385/400$). Clearly some of the 400 kinds of theoretical two-amino-acid sequences cannot appear in bovine p53 protein, a kind of two-amino-acid sequence being either present or absent in bovine p53 protein should be one of the 400 kinds of theoretical two-amino-acid sequences. Similarly, if three amino acids in a three-amino-acid sequence can randomly be constructed from any one of 20 amino acids, there are 8000 (20^3) kinds of theoretical three-amino-acid sequences. There are 384 three-amino-acid sequences in bovine p53 protein, if there was no repetition in them, there would be 384 kinds of three-amino-acid sequences, therefore one of the 8000 kinds of theoretical three-amino-acid sequences would be expected to appear about 0.048 times ($385/8000$). Naturally, many of the 8000 kinds of theoretical three-amino-acid sequences cannot appear in bovine p53 protein, a kind of three-amino-acid sequence being present or absent in bovine p53 protein should be one of the 8000 kinds of theoretical sequences. Similar reasoning can also be applied for other multi-amino-acid sequences.

(3) We use the random principle to predict the frequencies of presence and absence of amino-acid

*For correspondence. Email guang.wu@pharma.novartis.com; Tel. +41 61 696 7746; Fax. +41 61 696 6992.

Table 1. Distributions of 4 'W's in 4 parts of bovine p53 protein and their distribution probabilities

Part 1	Part 2	Part 3	Part 4	Probability
W	W	W	W	0.09375
	W	W	WW	0.56250
		WW	WW	0.14063
		W	WWW	0.18750
			WWWW	0.01563

sequences in bovine p53 protein from its amino acid composition and compare the predicted frequencies with the counted frequencies. For example, (i) There are 21 alanines (A) in bovine p53 protein, the frequency of a random construction of 'AA' would be expected to be 1.088 ($21/386 \times 20/385 \times 385$), i.e. the 'AA' would appear once, which is true, so its presence is predictable. (ii) There are 37 leucines (L) and 4 tryptophans (W) in bovine p53 protein, the frequency of a random construction of 'LW' would be expected to be 0.383 ($37/386 \times 4/385 \times 385$), i.e. the 'LW' would not appear, but it appears twice, so its presence is unpredictable. (iii) There are 17 asparagines (N) in bovine p53 protein, the frequency of a random construction of 'NW' would be expected to be 0.176 ($17/386 \times 4/385 \times 385$), i.e. the 'NW' would not appear, which is true, so its absence is predictable. (iv) The frequency of a random construction of 'AN' would be expected to be 1.469 ($21/386 \times 27/385 \times 385$), i.e. the 'AN' would appear once, but it is absent, so its absence is unpredictable.

(4) We use the random principle to predict the probability that an amino acid follows a preceding amino acid and compare the predicted probabilities with the probabilities occurred in bovine p53 protein (Markov transition probability). For example, (i) There are 31 glutamic acids (E) in bovine p53 protein, thus an 'E' would have a probability of 0.081 ($31/385$) in following a preceding amino acid, which is true in following an 'L', so an 'E' in following an 'L' is predictable. (ii) An 'A' would have a probability of 0.052 ($20/385$) in following a preceding amino acid, which is not true in following an 'A', so an 'A' in following an 'A' is unpredictable.

(5) We use the random principle to predict the

distributions of amino acids, two- and three-amino-acid sequences in bovine p53 protein and compare the predicted distributions with the measured distributions. For example, (i) There are 2 two-amino-acid sequence of 'AE' among 385 two-amino acid sequences in bovine p53 protein. Our intuition with regard to their randomness may suggest that there would be one 'AE' in the first half of bovine p53 protein and another 'AE' in the second half of bovine p53 protein, which is true, i.e. one 'AE' is at positions 6 and 7, and another 'AE' is at positions 196 and 197, so the distribution of 'AE' is predictable although we are still not able to predict their exact position. (ii) There are 4 'W's in bovine p53 protein and we can group the bovine p53 protein into four parts, each contains about 97 amino acids ($386/4 = 96.5$) and our intuition with regard to randomness may once again suggest that each part would contain a 'W'. If we do not distinguish four parts and four 'W's because we are only interested in how many 'W's occur in any part, for this purpose there is no difference between four 'W's and between four parts, we have 5 distributions of 'W's' (Table 1). The distribution of 'W's in bovine p53 protein is in fact that two parts contain 0 'W', one part contains 1 'W' and one part contains 3 'W's. So the distribution of 'W's in bovine p53 protein is neither our prediction nor the highest probabilistic distribution, and this distribution is unpredictable.

The last three procedures are important from the views of evolution and probability. With respect to the amino-acid sequences in bovine p53 protein, the functional regions in bovine p53 protein should deliberately be evolved and conserved, thus the presence of amino-acid sequences in these regions is unlikely to be predictable by a random mechanism; but the non-functional regions in bovine p53 protein are unlikely to be deliberately evolved and conserved, thus the presence of amino-acid sequences in these regions is possibly predictable by a random mechanism. With respect to the amino-acid sequence being absent from bovine p53 protein, some amino-acid sequences may deliberately be eliminated from bovine p53 protein during its evolutionary process, thus their absence should not be predictable by a random mechanism; whereas some amino-acid sequences may not deliberately be eliminated, thus their absence could be predictable by a random mechanism. Similarly, the predictable distributions should not be deliberately evolved and conserved, otherwise they should be deliberately evolved and conserved.

Table 2. Calculation of distribution probabilities of 4 'W's in 4 parts referring to Table 1.

Distribution	Calculation of probability ($4! \times 4! \times 4^{-4}$ divided by)	Probability
1, 1, 1, 1	$(0! \times 4! \times 0! \times 0! \times 0!) \times (1! \times 1! \times 1! \times 1!)$	0.09375
0, 1, 1, 2	$(1! \times 2! \times 1! \times 0! \times 0!) \times (0! \times 1! \times 1! \times 2!)$	0.56250
0, 0, 2, 2	$(2! \times 0! \times 2! \times 0! \times 0!) \times (0! \times 0! \times 2! \times 2!)$	0.14063
0, 0, 1, 3	$(2! \times 1! \times 0! \times 1! \times 0!) \times (0! \times 0! \times 1! \times 3!)$	0.18750
0, 0, 0, 4	$(3! \times 0! \times 0! \times 0! \times 1!) \times (0! \times 0! \times 0! \times 4!)$	0.01563

Table 3. Counted frequency (CF) and predicted frequency (PF) and the first order Markov chain transition probabilities (MP) of two-amino-acid sequences, which have a difference of ≥ 2 between counted and predicted frequencies in bovine p53 protein.

Sequence	CF	PF	MP	Sequence	CF	PF	MP	Sequence	CF	PF	MP
AP	6	2	0.286	AS	0	2	0.000	AT	3	1	0.143
RA	4	1	0.148	RL	1	3	0.037	RV	3	1	0.111
NL	5	2	0.294	NP	0	2	0.000	DG	3	1	0.167
DP	0	2	0.000	DS	4	2	0.222	CP	5	1	0.385
EK	0	2	0.000	GR	3	1	0.167	GN	3	1	0.167
GF	3	1	0.167	GS	0	2	0.000	LD	4	2	0.108
LW	2	0	0.054	LV	0	2	0.000	KR	4	1	0.200
KE	0	2	0.000	KK	6	1	0.300	FR	3	1	0.273
PA	4	2	0.091	PR	1	3	0.023	PQ	3	1	0.068
PG	4	2	0.091	PP	9	5	0.205	PT	1	3	0.023
SR	1	3	0.026	SN	0	2	0.000	SD	4	2	0.105
SC	4	1	0.105	SQ	3	1	0.079	SL	1	4	0.026
SS	7	4	0.184	TI	2	0	0.091	TP	1	3	0.045
VR	3	1	0.187	VL	0	2	0.000	VS	0	2	0.000
VV	3	1	0.000								

A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; E, glutamic acid; Q, glutamine; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine.

Results

Two-amino-acids sequences and their first order Markov chain transition probabilities

When counting 385 two-amino-acid sequences in bovine p53 protein, we found that 120 kinds of two-amino-acid sequences appear once, 51 kinds twice, 21 kinds three times, 13 kinds four times, 4 kinds five times, 2 kinds six times, 1 kind seven times and 1 kind nine times. When comparing the two-amino-acid sequences in bovine p53 protein with the 400 kinds of theoretical two-amino-acid sequences, we found 187 kinds of theoretical two-amino-acid sequences absent.

When using the random principle to predict two-amino-acid sequences on the base of amino acid composition of bovine p53 protein, we found that the random principle can predict the presence of 80 two-amino-acid sequences in bovine p53 protein and the absence of 98 kinds of theoretical two-amino-acid sequences from bovine p53 protein.

The unpredictable two-amino-acid sequences are important, because they should deliberately be developed. Statistically, the sequences having a difference of ≥ 2 between counted and predicted frequencies are particularly interesting, because the predicted frequency is the rounded value of predicted probability and the difference being equal to 1 is probably attributed to the rounding error. Table 3 shows these two-amino-acid sequences in bovine p53 protein.

When using the random principle to predict the probability that an amino acid follows a preceding amino acid in two-amino-acid sequences on the base of amino acid composition of bovine p53 protein and comparing with

the first order Markov transition probability, we found 3 probabilities predicable. Table 3 shows the first order Markov transition probabilities of two-amino-acid sequences which have a difference of ≥ 2 between counted and predicted frequencies.

Three-amino-acids sequences and their second order Markov chain transition probabilities

When counting 384 three-amino-acid sequences in bovine p53 protein, we found that 345 kinds of three-amino-acid sequences appear once, 16 kinds twice, 1 kind three times and 1 kind four times. When comparing the three-amino-acid sequences in bovine p53 protein with the 8000 kinds of theoretical three-amino-acids sequences, we found 7637 kinds of theoretical three-amino-acids sequences absent.

When using the random principle to predict three-amino-acid sequences on the base of amino acid composition of bovine p53 protein, we found that the random principle can predict the absence of 7637 kinds of theoretical three-amino-acid sequences from bovine p53 protein. Table 4 shows the three-amino-acid sequences which have a difference of ≥ 2 between counted and predicted frequencies.

With regard to the prediction of probability that an amino acid follows a preceding amino acid in three-amino-acid sequences, we did not find any predicted probability identical with the second order Markov transition probability. Table 4 shows the second order Markov transition probabilities of three-amino-acid sequences which have a difference of ≥ 2 between counted and predicted frequencies.

Table 4. Counted frequency (CF) and predicted frequency (PF) and the second order Markov chain transition probability (MP) of three-amino-acid sequences, which have a difference of ≥ 2 between counted and predicted frequencies in bovine p53 protein.

Sequence	CF	PF	MP	Sequence	CF	PF	MP	Sequence	CF	PF	MP
APA	2	0	0.333	APP	2	0	0.333	NLR	2	0	0.400
NLL	3	0	0.600	DAL	2	0	1.000	ELN	2	0	0.500
EPP	2	0	0.500	GNL	2	0	0.667	LDG	2	0	0.500
LLP	2	0	0.500	LSS	2	0	0.500	KKP	2	0	0.333
PAT	2	0	0.500	PEP	2	0	0.500	PLS	2	0	0.667
PPP	4	1	0.444	SAP	2	0	1.000	SPS	2	0	0.500

Distributions of amino acids

Table 5 shows amino acids, their numbers, their measured distribution probabilities, their maximum theoretical distribution probabilities, their random ranks against theoretical distribution probabilities and their 'even' distribution probabilities. We arrange all the probabilities in an descending order, we rank the highest probability as one. Table 5 shows that 'M' has the highest random rank while 'P' has the lowest random rank, and any measured distribution probability is far away from the 'even' distribution probability, i.e. none of amino acids distributes homogeneously along the bovine p53 protein.

Distributions of two- and three-amino-acid sequences

Table 6 shows the appeared-more-than-twice two- and three-amino acid sequences, their numbers, their

measured distribution probabilities, their maximum theoretical distribution probabilities and their random ranks against theoretical distribution probabilities in bovine p53 protein. It can be seen that 43% (19/44) of two- and three-amino acid sequences follow a probabilistic simplest pathway to distribute along the bovine p53 protein.

Discussion

The primary structure of proteins also provides the basis for studies and modelling of (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 base for higher level structures, etc.

Table 5. Amino acids, their numbers, measured distribution probability (MDP), maximum theoretical distribution probability (MTDP), random ranks against theoretical distribution probability (RRTDP) and 'even' distribution probability (EDP) in bovine p53 protein.

Amino acid	Number	MDP	MTDP	RRTDP	EDP
A	21	0.0013	0.0954	55	8.7446×10^{-9}
R	27	0.0012	0.0668	84	2.4556×10^{-11}
N	17	0.1098	0.1280	2	4.2997×10^{-7}
D	19	0.0895	0.1118	2	6.1486×10^{-8}
C	13	0.0617	0.1544	5	2.0560×10^{-5}
E	31	0.0147	0.0535	20	4.8174×10^{-13}
Q	11	0.1010	0.2020	4	1.3991×10^{-4}
G	18	0.0117	0.1246	22	1.6272×10^{-7}
H	9	0.0197	0.1967	10	9.3666×10^{-4}
I	8	0.0841	0.2523	4	2.4033×10^{-3}
L	37	0.0297	0.0395	6	1.3040×10^{-15}
K	20	0.0001	0.0965	100	2.3202×10^{-8}
M	9	0.1967	0.1967	1	9.3666×10^{-4}
F	11	0.1077	0.2020	3	1.3991×10^{-4}
P	44	0.0008	0.0308	167	1.2962×10^{-18}
S	38	0.0071	0.0373	37	4.8612×10^{-16}
T	22	0.0512	0.0878	6	3.2921×10^{-9}
W	4	0.1875	0.5625	2	9.3750×10^{-2}
Y	11	0.1616	0.2020	2	1.3991×10^{-4}
V	16	0.0568	0.1362	6	1.1342×10^{-6}

Table 6. Appeared-more-than-twice two- and three-amino acid sequences, measured distribution probability (MDP), maximum theoretical distribution probability (MTDP) and random ranks against theoretical distribution probability (RRTDP) in bovine p53 protein.

Amino acid pair/triplet	Number	MDP	MTDP	RRTDP
AL	3	0.1111	0.6667	3
AP	6	0.1543	0.3472	3
AT	3	0.1111	0.6667	3
RA	4	0.5625	0.5625	1
RR	3	0.6667	0.6667	1
RE	3	0.1111	0.6667	3
RV	3	0.6667	0.6667	1
NL	5	0.2880	0.3840	2
DG	3	0.6667	0.6667	1
DS	4	0.5625	0.5625	1
CP	5	0.3840	0.3840	1
EE	3	0.6667	0.6667	1
EL	4	0.1406	0.5625	3
EP	4	0.5625	0.5625	1
ES	3	0.2222	0.2222	2
GR	3	0.1111	0.6667	3
GN	3	0.6667	0.6667	1
GF	3	0.6667	0.6667	1
LN	3	0.6667	0.6667	1
LD	4	0.5625	0.5625	1
LE	3	0.6667	0.6667	1
LL	4	0.1875	0.5625	2
LK	3	0.1111	0.6667	3
LP	3	0.6667	0.6667	1
LS	4	0.0156	0.5625	5
KR	4	0.0156	0.5625	5
KK	6	0.1543	0.3472	3
FR	3	0.2222	0.6667	2
PA	4	0.0156	0.5625	5
PE	4	0.1406	0.5625	3
PQ	3	0.2222	0.6667	2
PG	4	0.0938	0.5625	4
PL	3	0.6667	0.6667	1
PP	9	0.1475	0.1967	4
PS	5	0.2880	0.3840	2
SD	4	0.5625	0.5625	1
SC	4	0.1875	0.5625	2
SQ	3	0.1111	0.6667	3
SP	5	0.3840	0.3840	1
SS	7	0.1285	0.3213	3
VR	3	0.6667	0.6667	1
VV	3	0.1111	0.6667	3
NLL	3	0.6667	0.6667	3
PPP	4	0.5625	0.5625	1

One of the rationales to conduct our study is that a good signature pattern of a protein must be as short as possible and many short sequences (not more than four or five residues long conserved sequence) are often diagnostics of certain binding properties or active sites (PROSITE 2000). While the multiple sequence comparisons and alignments can assign a function to a

protein based on its sequence similarity to other sequences within a family of proteins, our methods explore these short sequences further from the randomly probabilistic viewpoint and connect them with a possibly evolutionary process.

For the prediction of presence and absence of two- and three-amino-acid sequences, we may have several implications and applications. For example, the random

analysis divide the bovine p53 protein into predictable and unpredictable regions, i.e. 'random' and 'non-random' regions. We therefore can determine whether a mutation occurs in predictable or unpredictable regions, and deduce that a mutation in predictable region is unlikely to lead to a substantial change in protein function but a mutation in unpredictable region is likely to lead to a substantial change in protein function. Unfortunately we still do not have any available data to prove this in bovine p53 protein, however this has been proven in our study on the rat monoamine oxidase (Wu and Yan, 2001).

For the prediction of distributions of amino acids, two- and three-amino-acid sequences, we notice that almost all of amino acids, two- and three-amino-acid sequences do not homogeneously distribute along the bovine p53 protein but heterogeneously distribute along the bovine p53 protein. In fact the homogenous distribution has an impossible probability to occur in Table 5, thus a heterogeneous distribution with a higher probability of occurrence may construct several clusters which are the base for the protein three-dimensional structure and function. Another implication is that it is likely that a variant develops along the probabilistic simple pathway and thus is more easily to occur spontaneously if a variant in bovine p53 protein leads to an increased distribution probability, otherwise along the probabilistic difficult pathway and is more difficult to occur spontaneously.

Three functional regions have been documented in bovine p53 protein, i.e. (i) the aspartic acid/glutamic acid domain from position 1 to position 59, (ii) the nuclear localisation signal domain from position 304 to position 316 and (iii) the post-translational modification of a residue (phosphorylation) at position 380. Although the aspartic acid/glutamic acid and nuclear localisation signal domains are longer than the definition given in (PROSITE 2000) and the post-translational modification of a residue is shorter than the definition given in (PROSITE 2000), the aspartic acid/glutamic acid domain contains 'NLL' and 'LLP' with a difference of ≥ 2 between counted and predicted frequencies and the nuclear localisation signal domain contains 'KK' and 'SS' with a difference of ≥ 2 between counted and predicted frequencies. Our methods suggest that these four amino-acid sequences would have a particular importance. If a more exactly functional property of bovine p53 protein would have been documented, our methods may definitely determine them.

In the past we have used the first four probabilistic procedures in several occasions (Wu, 2000a,b; Wu and Yan, 2000), however we did not reason our analysis to this level. Our probabilistic approaches may provide the insight into the underlying mechanism for the protein configuration and evaluation.

Experimental Procedures

The amino acid sequence of the bovine p53 protein was obtained from the SWISS-PROT Protein Sequence Database, access number Q29628 (Bairoch and Apweiler, 1999).

Counting two- and three-amino-acid sequences

The two- and three-amino-acid sequences in bovine p53 protein were counted as described in Introduction, however no more-than-three-amino-acid sequences was counted, because no repetition regarding the more-than-three-amino-acid sequences was found, thus each more-than-three-amino-acid sequence is unique.

Numbers of theoretical two- and three-amino-acid sequences

As all 20 kinds of amino acids are present in bovine p53 protein and the number of each kind of amino acid is at least larger than 3, thus there are 400 (20^2) and 8000 (20^3) kinds of theoretical sequences for two- and three-amino-acid sequences.

Calculating predicted probability and frequency

The predicted probability was calculated according to the random mechanism as described in the Introduction. For example, there are 21 'A's and 27 arginines (R) in bovine p53 protein. For two-amino-acid sequences at any position, the predicted probabilities for 'AA', 'AR', 'RR' and 'RA' are $21/386 \times 20/385$, $31/386 \times 27/385$, $27/386 \times 26/385$ and $27/386 \times 21/385$. For three-amino-acid sequences, the predicted probability for 'AAA' is $21/386 \times 20/385 \times 19/384$. The numbers of predicted probabilities are identical to the numbers of theoretical kinds of two- and three-amino-acid sequences. The predicted frequency is the rounded integral value of the product of predicted probability and total number of amino-acid sequences, thus the predicted frequency for 'AA' is 1 ($21/386 \times 20/385 \times 385$).

Calculating predicted probability that an amino acid follows a preceding amino acid

The predicted probability for an amino acid in following a preceding amino acid is calculated according to the random mechanism as described in the Introduction. For example, there are 21 'A's and 27 'R's in bovine p53 protein. The predicted probabilities for the second amino acid of 'A' in two-amino-acid sequences of 'AA' and 'RA' are $20/385$ and $21/385$, the predicted probabilities for the second amino acid of 'R' in two-amino-acid sequences of 'AR' and 'RR' are $27/385$ and $26/385$, and the predicted probability for the third amino acid of 'A' in a three-amino-acid sequence of 'AAA' is $19/384$. The numbers of predicted probabilities are identical to the numbers of theoretical kinds of two- and three-amino-acid sequences.

Calculating Markov transition probability

The Markov chain is to calculate the transition probability from one state to another state (Feller, 1968). For a two-amino-acid sequence, an amino acid has a certain probability in following a certain preceding amino acid, which constructs a probability (the first order Markov chain), i.e. the probability of an amino acid occurs in a two-amino-acid sequence given a certain first amino acid [$P(\text{second amino acid}|\text{first amino acid})$]. The calculation of this probability is the transition from the state of one-amino-acid sequence (if it can be called as a sequence) to the state of two-amino-acid sequence, and the state of two-amino-acid sequence is only dependent on the state of

one-amino-acid sequence. For a three-amino-acid sequence, the second order Markov chain can be defined, i.e. the probability of an amino acid occurs in a three-amino-acid sequence given a certain first two amino acid [$P(\text{third amino acid} | \text{first and second amino acids})$]. The calculation of this probability is the transition from the state of two-amino-acid sequence to the state of three-amino-acid sequence, and the state of three-amino-acid sequence is only dependent on the state of two-amino-acid sequence.

Calculation of distribution probabilities of amino acids, two- and three-amino-acid sequence

The calculation of distributions of amino acids and amino acid pairs is according to the calculation of occupancy problems of subpopulations and partitions (Feller, 1968). For each of distributions of amino acids and amino acid pairs, the probability is $n! / (q_0! \times q_1! \times \dots \times q_n!) \times r! / (r_1! \times r_2! \times \dots \times r_n!) \times n^{-r}$.

In the equation, ! is the factorial function, i.e. $n! = n \times (n-1) \times (n-2) \times \dots \times 1$, for example, $4! = 4 \times 3 \times 2 \times 1 = 24$ and $0! = 1$ by definition. r is the number of a given kind of amino acid or amino acid pair, for example, we have $r = 4$ for 'W' and $r = 21$ for 'A' because there are 4 'W's and 21 'A's in bovine p53 protein. n is the number of grouped parts in bovine p53 protein for a given kind of amino acid or amino acid pair, for example, $n = 4$ for 'W', in fact we actually have $r = n$ in this study. r_1, r_2, \dots, r_n are the number of a given kind of amino acid or amino acid pair in part 1, 2, ... n , for example, when 4 'W's appear in each of 4 parts, we have $r_1 = 1, r_2 = 1, r_3 = 1$ and $r_4 = 1$. q is the number of parts with the same number of amino acid or amino acid pair, for example, when 4 'W's appear in each of 4 parts, we have $q_0 = 0, q_1 = 4, q_2 = 0, q_3 = 0$ and $q_4 = 0$, i.e. 0 part has 0 'W', 4 parts have 1 'W', 0 part has 2 'W's, and 0 part has 3 'W's, 0 part has 4 'W's.

Table 2 shows the calculation using this equation with respect to the distributions of 4 'W's in bovine p53 protein. Compared with Table 1, we can understand not only which distribution is more likely to occur, but also the comparison between them, for example, the distribution of 'W', 'W', and 'WW' is 36 times ($0.56250/0.01563$) more likely to occur than the distribution of 'WWWW'.

Statistical comparison

The counted and predicted frequencies, and the Markov transition and predicted probabilities are compared based on the following statistical consideration. In the general case of this type of analysis, an amino acid, a two-amino-acid sequence and a three-amino-acid sequence have the chances of $1/20$ ($P = 0.05$), $1/400$ ($P = 0.0025$) and $1/8000$ ($P = 0.000125$) to repeat once, respectively. In case of bovine p53 protein, 44 'P's are the most abundant amino acid and 4 'W's are the least abundant amino acid. If the first amino acid is 'P', the chance for the second 'P' is $43/385$ ($P = 0.1117 > 0.05$). If the first amino acid is 'W', the chance for the second 'W' is $3/385$ ($P = 0.0078 < 0.01$). With respect to two- and three-amino-acid sequences, the chance for the first 'PP' is $44/386 \times 43/385$ ($P = 0.0127 < 0.05$) and the chance for the second 'PP' is $42/384 \times 41/383$ ($P = 0.0117 < 0.05$), the chance for the first 'PPP' is $44/386 \times 43/385 \times 42/384$ ($P = 0.0014 < 0.01$) and the

chance for the second 'PPP' is $42/383 \times 41/382 \times 40/381$ ($P = 0.0012 < 0.01$). When considering the two- and three-amino-acid sequences constructed by other less abundant amino acids in bovine p53 protein, the chance of appearance and repetition would be further smaller, therefore all the probabilities of two- and three-amino-acid sequences are less than 0.05. The statistical difference is much less than 0.05 when two repetitions are used in Tables.

References

- Bairoch, A., and Apweiler, R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* 27: 49–54.
- Dequiedt, F., Kettmann, R., Burny, A., and Willems, L. 1995. Nucleotide sequence of the bovine P53 tumor-suppressor cDNA. *DNA Seq.* 5: 261–264.
- Feller, W. 1968. *An Introduction to Probability Theory and Its Applications.* 3rd ed, New York: John Wiley and Sons, Vol I.
- Komori, H., Ishiguro, N., Horiuchi, M., Shinagawa, M., and Aida, Y. 1996. Predominant p53 mutations in enzootic bovine leukemic cell lines. *Vet. Immunol. Immunopathol.* 52: 53–63.
- PROSITE: A dictionary of protein sites and patterns user manual, <http://www.expasy.ch/prosite/>.
- Wu G. 2000a. Frequency and Markov chain analysis of amino-acid sequence of human glutathione reductase. *Biochem. Biophys. Res. Commun.* 268: 823–826.
- Wu G. 2000b. Frequency and Markov chain analysis of amino-acid sequence of human tumor necrosis factor. *Cancer Lett.* 153: 145–150.
- Wu G., and Yan, S.M. 2000. Prediction of two- and three-amino-acid sequences of *Citrobacter Freundii* β -lactamase from its amino acid composition. *J. Mol. Microbiol. Biotechnol.* 2: 277–281.
- Wu G, Yan SM. 2001. Prediction of presence and absence of two- and three-amino-acid sequence of human monoamine oxidase B from its amino acid composition according to the random mechanism. *Biomolecular Engineering (former Genetic Analysis: Biomolecular Engineering)* 18: 23–27.

