Molecular Biology Today (2002) 3: 55-69.

Randomness in the Primary Structure of Protein: Methods and Implications

Guang Wu[1*], Shaomin Yan[2]

[1]Laboratoire de Toxicocinétique et Pharmacocinétique, Faculté de Pharmacie, Université de la Méditerranée Aix-Marseille II, Marseille, France
[2]Cattedra di Anatomia Patologica, Dipartimento di Ricerche Mediche e Morfologiche, Facoltà di Medicina e Chirurgia, Università degli Studi di Udine, Udine, Italy.

For correspondence: Guang Wu, ICPP, WKL-135.1.16, Novartis Pharma AG, CH-4002 Basel, Switzerland.

Abstract

It is no doubt that the evolutionary process is affected by chance, but the question is to what extent the chance plays its role. The random analysis can throw light on the underlying reasoning for the primary structure of proteins. With the use of random principles, we have explored three approaches to analyse protein primary structure, i.e. the randomness in the construction of amino-acid sequences, in the follow-up amino acid, and in the distribution of amino acid/amino acids. As the results, (i) we can evaluate the impact of chance on the composition of amino-acid sequences by comparing the measured probability/frequency with the predicted probability/frequency; (ii) we can evaluate the impact of chance on the follow-up amino acid by comparing the Markov transition probability with the predicted conditional probability; and (iii) we can evaluate the effect of chance on the distribution of amino acids by comparing the real distribution probability with the theoretical distribution probability. These approaches can be used to quantitatively analyse the primary structure of intra-protein as well as inter-proteins, thus we can get more insights into the mechanisms of protein construction, mutation, and evolutionary process. Also, these approaches may have some potential use for development of new drugs.

Introduction

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 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.

The patterns of amino-acid composition and mutations are archived via experimental methods and annotation (Barker *et al.*, 2001). 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 (Altschul *et al.*, 1990). There are several other approaches for the similarity analysis, such as fast Fourier transform (Benson, 1990), the statistical approach (Karlin *et al.*, 1991), linguistic approaches (Popov *et al.*, 1996; Searls, 1997), pattern graph (Jonassen, 1997) and so on.

With the advance in the experimental and theoretical studies, we always get new insights from the primary structure of proteins. For example, an intriguing phenomenon is that the majority of amino acids cluster in some regions rather than homogenously distribute along the primary structure of a protein, although this could simply be due to the nature selection. To our knowledge, there are three possible factors that can affect the distribution of amino acids in a protein, i.e. adaptation, chance and history. The adaptation refers to the current selective benefit of having a particular amino-acid configuration. The chance can affect amino acid distributions by mutation and the fixation of mutations, and is most easily seen to occur when selective differences between different amino-acid distributions are non-existent (neutral evolution). The effect of history can be seen as the starting point for subsequent adaptation.

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 (Everitt, 1999), but also because it can provide quantitatively measures to compare the composition and distribution of amino acids in a protein. Therefore, it is no doubt that the evolutionary process is affected by chance, but the question is to what extent the chance plays its role. During the last several years we applied three types of randomly probabilistic analyses on the protein primary structure, here we briefly review these methods, their meanings and potential uses.

Three types of randomness in protein primary structure

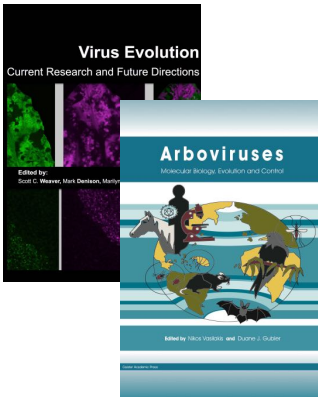
"Random" and "randomness" have been defined as follows (1) without a pattern: done, chosen, or occurring without a specific pattern, plan, or connection; (2) lacking regularity: with a pattern or in sizes that are not uniform or regular; (3) statistics equally likely: relating or belonging to a set in which all the members have the same probability of occurrence; (4) statistics have definite probability: relating to or involving variables that have undermined value but definite probability (Encarta World English Dictionary, http://dictionary.msn.com).

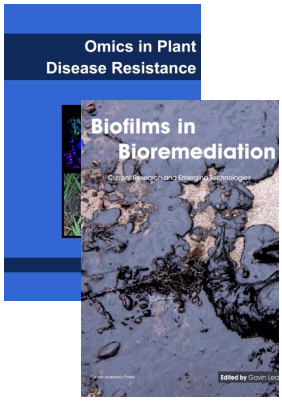
*For correspondence. Email guang.wu@pharma.novartis.com; Tel. +41 061 696 7746; Fax. +41 061 696 6992.

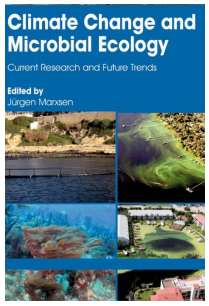
Further Reading

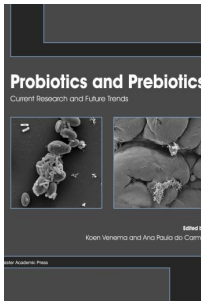
Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at **caister.com**

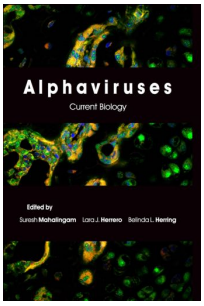
- **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof
- ***Aspergillus* and *Penicillium* in the Post-genomic Era**
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2
- **The Bacteriocins: Current Knowledge and Future Prospects**
  Edited by: RL Dorit, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins
- **Omics in Plant Disease Resistance**
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr
- **Acidophiles: Life in Extremely Acidic Environments**
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles
- **Climate Change and Microbial Ecology: Current Research and Future Trends**
  Edited by: J Marxsen (2016)
  www.caister.com/climate
- **Biofilms in Bioremediation: Current Research and Emerging Technologies**
  Edited by: G Lear (2016)
  www.caister.com/biorem
- **Microalgae: Current Research and Applications**
  Edited by: MN Tsaloglou (2016)
  www.caister.com/microalgae
- **Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives**
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma
- **Virus Evolution: Current Research and Future Directions**
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol
- **Arboviruses: Molecular Biology, Evolution and Control**
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo
- ***Shigella*: Molecular and Cellular Biology**
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella
- **Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment**
  Edited by: AM Romaní, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms
- **Alphaviruses: Current Biology**
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha
- **Thermophilic Microorganisms**
  Edited by: F Li (2015)
  www.caister.com/thermophile

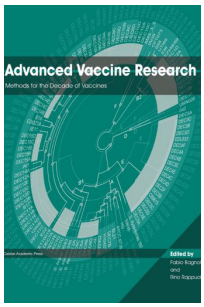


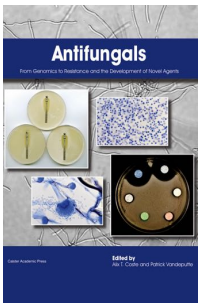












- **Flow Cytometry in Microbiology: Technology and Applications**
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow
- **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics
- **Epigenetics: Current Research and Emerging Trends**
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015
- ***Corynebacterium glutamicum*: From Systems Biology to Biotechnological Applications**
  Edited by: A Burkovski (2015)
  www.caister.com/cory2
- **Advanced Vaccine Research Methods for the Decade of Vaccines**
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines
- **Antifungals: From Genomics to Resistance and the Development of Novel Agents**
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals
- **Bacteria-Plant Interactions: Advanced Research and Future Trends**
  Edited by: J Murillo, BA Vinatzer, RW Jackson, et al. (2015)
  www.caister.com/bacteria-plant
- ***Aeromonas***
  Edited by: J Graf (2015)
  www.caister.com/aeromonas
- **Antibiotics: Current Innovations and Future Trends**
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics
- ***Leishmania*: Current Biology and Control**
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2
- ***Acanthamoeba*: Biology and Pathogenesis (2nd edition)**
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2
- **Microarrays: Current Technology, Innovations and Applications**
  Edited by: Z He (2014)
  www.caister.com/microarrays2
- **Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications**
  Edited by: D Marco (2014)
  www.caister.com/n2

Order from caister.com/order

Accordingly, it seems that the randomness has no relationship with the primary structure of protein as the protein has its aims and patterns which are the objectives of numerous studies. However if we consider the protein primary structure from statistical viewpoint, we have the possibility of studying the randomness in the primary structure of proteins, i.e. we can analyse the protein primary structure using the random principle. Table 1 lists the amino-acid composition in the proteins which we have previously studied using our approaches. After carefully looking at the protein sequences, it is natural to ask the following questions:

1 Why a type of amino acid is adjacent to a certain type of amino acid not to the others, for example, why alanine (A) and glutamic acid (E) appear together, but alanine (A) and arginine (R) do not appear together in human glutathione reductase (Tutic *et al*., 1990). Can the composition of amino-acid sequence be explained by the random principle? If so, how many amino-acid sequences can be explained by the random principle?

2 Why a type of amino acid is more likely to follow a preceding amino acid, for example, why methionine (M) follows threonine (T), but does not follow leucine (L) in human monoamine oxidase type B (Grimsby *et al*., 1991).

3 Why a type of amino acids clusters in some regions rather than homogenously distributes along the protein sequence. For example, there are seven phenylalanines ("F"s) in *Citrobacter Freundii* β-lactamase (Lindberg and Normark, 1986), why these seven "F"s do not homogenously distribute along *Citrobacter Freundii* β-lactamase.

Simple explanation of three types of randomness in protein primary structure

- The randomness in the adjacency of amino acid (the construction of amino-acid sequences) can be simplified as such a random experiment. For example, we know that human tumour necrosis factor precursor is composed of 233 amino acids (Pennica *et al*., 1984), each type of amino acid contributes differently to these 233 amino acids (Table 1). If these 233 amino acids were mixed together, we randomly take two amino acids from these 233 amino acids, then what a chance we can obtain "AN". After taking the first pair of "AN" out, what a chance we can obtain the second pair of "AN" from 231 remaining amino acids. These random processes can proceed until no amino acid remains. Similarly what a chance we cannot obtain "AE" from 233 amino acids, then from 231 amino acids and so on. These random processes can also apply to three-amino-acid, four-amino-acid and multi-amino-acid sequences.
- The randomness in amino-acid following a preceding amino acid/amino acids (follow-up amino acid) is similar to the random experiment, however the difference is that we take two amino acids each time in the above experiment for two-amino-acid sequence, while in this experiment we fix the first amino acid "A", for example, in human tumour necrosis factor precursor, then randomly take another amino acid to see what a chance the second amino acid is "E". As two-letter words in English language, how large is

Table 1 Amino-acid composition from different proteins

Protein	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	Total
Human AML-1[a]	43	32	16	18	3	15	19	31	12	12	34	9	13	15	53	55	33	3	15	22	453
Human ADH1[b]	31	9	11	17	16	20	7	36	8	26	29	32	10	15	20	23	22	2	4	36	374
Human CTGF[c]	22	22	9	20	39	17	7	28	3	10	21	24	13	12	28	20	22	2	8	22	349
Human GR[d]	53	24	17	21	10	30	11	46	16	29	43	34	17	16	28	34	32	4	13	44	522
Human MAO-B[e]	34	31	17	20	9	42	13	42	13	28	52	31	15	14	29	20	35	11	23	40	519
Human PAF-AH-α[f]	21	22	11	26	10	31	10	25	16	16	27	34	11	14	13	37	27	12	10	36	409
Human TNF-α[g]	19	14	7	7	4	16	13	17	4	12	30	8	2	10	15	20	10	2	7	16	233
Human tyrosinase[h]	30	27	27	30	17	27	23	34	17	22	55	17	15	31	33	51	19	14	23	17	529
Human TAT[i]	30	21	19	22	17	30	13	30	9	31	45	25	15	15	32	33	16	5	15	31	454
Bovine p53[j]	21	27	17	19	13	31	11	18	9	8	37	20	9	11	44	38	22	4	11	16	386
Mouse p53[k]	24	24	8	17	12	32	14	24	11	9	35	24	11	13	38	35	24	3	12	20	390
Sheep p53[l]	20	27	14	16	11	34	13	21	10	8	35	19	10	13	39	42	19	4	10	17	382
AMPC-CITFR[m]	41	13	14	12	2	19	22	31	6	21	35	24	11	9	22	25	21	13	15	25	381

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.

AML-1 – acute myeloid leukaemia 1 protein; ADH1 – alcohol dehydrogenase α-chain; CTGF – connective tissue growth factor; GR – glutathione reductase; MAO-B – monoamine oxidase type B; PAF-AH- α – platelet-activating factor acetylhydrolase α -subunit; TNF- α – tumour necrosis factor precursor; TAT – tyrosine aminotransferase; AMPC-CITFR – *Citrobacter Freundii* β-lactamase.

[a] Levanon *et al*., 1994; [b] Matsuo and Yokoyama, 1989; [c] Bradham *et al*., 1991; [d] Tutic *et al*., 1990; [e] Grimsby *et al*., 1991; [f] Lo Nigro *et al*., 1997; [g] Pennica *et al*., 1984, [h] Takeda *et al*., 1989; [i] Rettenmeier *et al*., 1990; [j] Dequiedt *et al*., 1995a; [k] Jenkins *et al*., 1984; [l] Dequiedt *et al*., 1995b; [m] Lindberg and Normark, 1986.

the probability that the letter "e" appears given the first letter "w".

• The randomness in the distribution of amino acids along a protein is similar to that if we divide the *Citrobacter Freundii* β-lactamase into seven parts, each part would contain how many "F"s, which is similar to that how many letters we would expect to receive per day if we would randomly receive seven letters per week, or we have seven balls to throw randomly into seven holes.

In this review, an attempt has been made to give a general view on the methods, which are far more related to the probabilistic calculations and are not familiar with the experimental scientists, used in our previous studies and their implications.

Methods for the calculations

The calculations with respect to the first two examples are relatively simple, while the calculation with respect to the third example is somewhat difficult. First of all, it is necessary to count the numbers of amino acid, two-, three- and multi-amino-acid sequences along a protein sequence. Then the possible and predicted probabilities and frequencies are calculated for analysing of the construction of amino-acid sequences; the predicted conditional probability and Markov chain transition probability are calculated for analysing of follow-up amino acid; and the distribution probability and rank are calculated for analysing of distributions of amino acids along a protein sequence.

Counting amino-acid sequences

The amino-acid sequences of proteins can be obtained from public data bank, for example, the Swiss-Protein (Bairoch and Apweiler, 1999).

Along a protein sequence, any two amino acids in order can construct a two-amino-acid sequence, i.e. the first and second; the second and third; etc. Furthermore, any three amino acids in order can construct a three-amino-acid sequence, i.e. the first, second and third; the second, third and fourth; etc. The similar consideration can be deduced for multi-amino-acid sequences. Thus the total numbers of constructed amino-acid sequences can be counted. For example, there are 207 amino acids in the human membrane-bound form of DNA methyltransferase (Hayakawa *et al.*, 1990), totally there are 206 two-amino-acid sequences, 205 three-amino-acid sequences, 204 four-amino-acid sequences and so on. Consequently, these amino-acid sequences are grouped in order to know how many times the same sequences repeat in the protein.

The reason for grouping amino acid and amino-acid sequences 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 finding properties or active sites (PROSITE: a dictionary of protein sites and patterns user manual, http://www.expasy.ch/prosite/). At present, we know that a single amino-acid "word" is not constructed by three amino acids as DNA by three codons, but we do not know how many amino acids construct a single amino-acid "word". Also we do not know whether there are "punctuation" and "space" in a protein sequence, so we do not know where an amino-acid "word" begins and finishes and at this stage an amino-acid "word" can begin and finish anywhere.

Calculating possible amino-acid sequences

In an ideally random situation, two amino acids in a two-amino-acid sequence could be constructed from any one of 20 amino acids, there would be 400 (20^2) possible sequences (combinations). Naturally, any two-amino-acid sequence in a protein should be one of 400 possible sequences, and any two-amino-acid sequence which does not appear in a protein also should be one of 400 possible sequences. Similarly, if three amino acids in a three-amino-acid sequence could be randomly constructed from any one of 20 amino acids, there would be 8000 (20^3) possible sequences. Naturally, any three-amino-acid sequence in a protein should be one of 8000 possible sequences, and any three -amino-acid sequence which does not appear in a protein also should be one of 8000 possible sequences. The similar deduction can be applied to more-than-three-amino-acid sequences, thus there are 160 000 (20^4) possible four-amino-acid sequences, 3 200 000 (20^5) possible five-amino-acid sequences and so on.

A caution should be given to the least number of amino acids. For example, there are 7600 (20^2 x 19) and 144400 (20^2 x 19^2) possible three- and four-amino-acid sequences, respectively, in human alcohol dehydrogenase α-chain, which contains only two tryptophans ("W"s). In human tumour necrosis factor, there are 7200 (20^2 x 18) and 129600 (20^2 x 18^2) possible three- and four-amino-acid sequences, respectively, because there are only 2 methionines ("M"s) and 2 tryptophans ("W"s) in this protein.

Calculating predicted probability and frequency

There are 25 alanines ("A"s) and 7 arginines ("R"s) in the human membrane-bound form of DNA methyltransferase. For two-amino-acid sequences, the predicted probability for "AA", "AR", "RR" and "RA" are 25/207 x 24/206, 25/207 x 7/206, 7/207 x 6/206 and 7/207 x 25/206. For three-amino-acid sequences, the predicted probability for "AAA" is 25/207 x 24/206 x 23/205.

The predicted frequency is the rounded integral value of the production of predicted probability and total number of amino-acid sequences, so the predicted frequency for "AA" is 3 (25/207 x 24/206 x 206). Although the predicted frequency is less accurate than the predicted probability, however, the predicted frequency is easier to use for the more-than-two-amino-acid sequences, in which the predicted probability is extremely low.

C*alculating presence and absence of amino-acid sequences by random frequencies*.

Again an example is the human membrane-bound form of DNA methyltransferase. (i) There are 25 alanines ("A"s) and 7 arginines ("R"s) in this enzyme, the frequency of a random construction of "AR" would be expected to be 0.845 (25/207 x 7/206 x 206), i.e. the "AR" would be expected to appear once, which is true in the real situation, so the presence of "AR" is randomly predictable. (ii) There are 4 tryptophans ("W"s) in this enzyme, the frequency of a random construction of "AW" would be expected to be 0.483

(25/207 x 4/206 x 206), i.e. the "AW" would be expected to be absent, but the "AW" appears twice in the real situation, so the presence of "AW" is randomly unpredictable. (iii) As there are 7 "R"s in this enzyme, the frequency of a random construction of "RR" would be expected to be 0.203 (7/207 x 6/206 x 206), i.e. the "RR" would be expected to be absent, which is true, so the absence of "RR" is randomly predictable. (iv) As there are 13 glutamic acids ("E"s) in this enzyme, the frequency of a random construction of "AE" would be expected to be 1.570 (25/207 x 13/206 x 206), i.e. the "AE" would be expected to appear twice, but the "AE" is absent, so the absence of "AE" is randomly unpredictable. In this way, we can compare the impact of chance on the composition of amino-acid sequences in a protein.

Calculating predicted conditional probability

In an ideally random situation, each amino acid could be possible to follow a preceding amino acid, thus the probability of follow-up amino acid is 1/20. In human tumour necrosis factor, there are 19 alanines ("A"s) and 14 arginines ("R"s), the predicted conditional probabilities for "AA" and "RA" are 18/232 and 19/232 for the second amino acid of "A" in two-amino-acid sequences to follow an "A" and a "R"; the predicted conditional probabilities for "AR" and "RR" are 14/232 and 13/232 for the second amino acid of "R" to follow an "A" and a "R". The predicted conditional probability of the third amino acid of "A" in a three-amino-acid sequence to follow "AA" is 17/231.

Calculating Markov chain transition probability

The Markov chain is to calculate the transition probability from one state to another state (Ash, 1965; Feller, 1968; Csiszar, 1981; van der Lubbe, 1997). For a two-amino-acid sequence, an amino acid has how large probability to follow a certain preceding amino acid, which constructs a conditional probability (the first order Markov chain), i.e. the probability of an amino acid occurs in a two-amino-acid sequence given a certain kind first amino acid [*P*(second amino acid|first amino acid)]. 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 kinds first two amino acids [*P*(third amino acid|first and second amino acids)].

Calculating follow-up amino acid by random principle

By comparing predicted conditional probability with Markov transition probability, it can be obtained that how many percentages of the Markov transition probability can be explained by a purely random mechanism. For example, there are 58 asparagines ("N"s) among 1434 amino acids in human nitric-oxide synthase type I form (Fujisawa *et al*., 1994; Hall *et al*., 1994). A "N" would have the probability of 0.040 (58/1433) in following a preceding "E" according to a purely random mechanism, which is true in the real situation, so the fact that a "N" follows a preceding "E" can be predicted by a purely random mechanism. By contrast, a "N" would have the probability of 0.040 (58/1433) in following a proceeding "A" according to a purely random mechanism. But the "N" has the probability of 0.057 in following a proceeding "A" in the real situation, thus the fact that a "N" follows a proceeding "A" cannot be predicted by a purely random mechanism.

Calculating amino-acid distributions along a protein primary structure

The calculation of amino-acid distributions is similar to the calculation of occupancy problems of subpopulations and partitions (Feller, 1968). For each of distributions of amino acids, the probability is $r!/(q_0! \times q_1! \times \ ... \times q_n!) \times r!/(r_1! \times r_2! \times ... \times r_n!) \times n^{-r}$. In the equation, ! is the factorial function, i.e. $n! = n \times (n-1) \times (n-2) \times ... \times 1$, for example, $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5040$ and $0! = 1$ by definition. r is the number of a given kind amino acid or amino acid pair, for example, we have $r = 7$ for "F"s because there are 7 "F"s in *Citrobacter Freundii* β-lactamase. n is the number of grouped parts in *Citrobacter Freundii* β-lactamase sequence for a given amino acid or amino acid pair, for example, $n = 7$ for "F", in fact we actually have $r = n$ in this type of calculation. r_1, r_2 ... r_n are the number of a given kind of amino acid in part 1, 2, ... n, for example, when 7 "F"s appear in each of 7 parts, we have $r_1 = 1$, $r_2 = 1$, $r_3 = 1$, $r_4 = 1$, $r_5 = 1$, $r_6 = 1$ and $r_7 = 1$. q is the number of parts with the same number of amino acid, for example, when 7 "F"s appear in each of 7 parts, we have $q_0 = 0$, $q_1 = 7$, $q_2 = 0$, $q_3 = 0$, $q_4 = 0$, $q_5 = 0$, $q_6 = 0$ and $q_7 = 0$, i.e. 0 part contains 0 "F", 7 parts contain 1 "F", 0 part contains 2 "F"s, 0 part contains 3 "F"s, 0 part contains 4 "F"s, 0 part contains 5 "F"s, 0 part contains 6 "F"s and 0 part contains 7 "F"s.

Table 2 details the calculation using this equation with respect to the distributions of seven amino acids in seven parts. The first seven columns show that the protein is divided into seven parts, and the first seven cells in each row show a possible configuration of amino acids; the eighth column is the numeric presentation of each configuration; the ninth column shows the details of calculation of distribution probability (the first and second parenthesis correspond to q and r in the equation); the tenth column is the distribution probability and the eleventh column is the distribution rank.

Ranking possible distribution probabilities

In order to avoid too many decimal values, we rank the possible distribution probabilities according to a descending order, thus the highest distribution probability is ranked as one. For example, there are 15 types of possibly theoretical distributions for seven amino acids in seven parts, but we rank them only to 13 because the same probability can be calculated from different distributions (rank 4 and 6 in Table 2).

Calculating distribution ranks

As the composed number of amino acids is different one another even in the same protein family, we standardise the impact of chance on a protein using the distribution ranks per amino acid and per each type of amino acid. These can be calculated by dividing the sum of ranks by the number of amino acids and the sum of ranks of each type of amino acids by the number of corresponding type of amino acids. Taking the mouse bcl-2 as an example, this protein contains 236 amino acids and the sum of ranks

Table 2. All possible distributions of 7 'F's in 7 parts with their distribution probability and rank.

Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Distribution	Calculation of probability (7!x 7!x 7^{-7} divided by)	Probability	Rank
F	F	F	F	F	F	F	1, 1, 1, 1, 1, 1 ,1	(0!x 7!x 0!x 0!x 0!x 0!x 0!x 0!x)x(1!x 1!x 1!x 1!x 1!x 1!x 1!x)	0.006120	8
	F	F	F	F	F	FF	0, 1, 1, 1, 1, 1, 2	(0!x 5!x 1!x 0!x 0!x 0!x 0!x 0!x)x(0!x 1!x 1!x 1!x 1!x 1!x 2!x)	0.128518	3
		F	F	F	FF	FF	0, 0, 1, 1, 1, 2, 2	(2!x 3!x 2!x 0!x 0!x 0!x 0!x 0!x)x(0!x 0!x 1!x 1!x 1!x 2!x 2!x)	0.321295	1
			F	FF	FF	FF	0, 0, 0, 1, 2, 2, 2	(3!x 1!x 3!x 0!x 0!x 0!x 0!x 0!x)x(0!x 0!x 0!x 1!x 2!x 2!x 2!x)	0.107098	4
		F	F	F	F	FFF	0, 0, 1, 1, 1, 1, 3	(2!x 4!x 0!x 1!x 0!x 0!x 0!x 0!x)x(0!x 0!x 1!x 1!x 1!x 1!x 3!x)	0.107098	4
			F	F	FF	FFF	0, 0, 0, 1, 1, 2, 3	(3!x 2!x 1!x 1!x 0!x 0!x 0!x 0!x)x(0!x 0!x 0!x 1!x 1!x 2!x 3!x)	0.214197	2
				FF	FF	FFF	0, 0, 0, 0, 2, 2, 3	(4!x 0!x 2!x 1!x 0!x 0!x 0!x 0!x)x(0!x 0!x 0!x 0!x 2!x 2!x 3!x)	0.026775	6
				F	FFF	FFF	0, 0, 0, 0, 1, 3, 3	(4!x 1!x 0!x 2!x 0!x 0!x 0!x 0!x)x(0!x 0!x 0!x 0!x 1!x 3!x 3!x)	0.017850	7
			F	F	F	FFFF	0, 0, 0, 1, 1, 1, 4	(3!x 3!x 0!x 0!x 1!x 0!x 0!x 0!x)x(0!x 0!x 0!x 1!x 1!x 1!x 4!x)	0.035699	5
				F	FF	FFFF	0, 0, 0, 0, 1, 2, 4	(4!x 1!x 1!x 0!x 1!x 0!x 0!x 0!x)x(0!x 0!x 0!x 0!x 1!x 2!x 4!x)	0.026775	6
					FFF	FFFF	0, 0, 0, 0, 0, 3, 4	(5!x 0!x 0!x 1!x 1!x 0!x 0!x 0!x)x(0!x 0!x 0!x 0!x 0!x 3!x 4!x)	0.001785	10
				F	F	FFFFF	0, 0, 0, 0, 1, 1, 5	(4!x 2!x 0!x 0!x 0!x 1!x 0!x 0!x)x(0!x 0!x 0!x 0!x 1!x 1!x 5!x)	0.005355	9
					FF	FFFFF	0, 0, 0, 0, 0, 2, 5	(0!x 0!x 1!x 0!x 0!x 1!x 0!x 0!x)x(0!x 0!x 0!x 0!x 0!x 2!x 5!x)	0.001071	11
					F	FFFFFF	0, 0, 0, 0, 0, 1, 6	(5!x 1!x 0!x 0!x 0!x 0!x 1!x 0!x)x(0!x 0!x 0!x 0!x 0!x 1!x 6!x)	0.000357	12
						FFFFFFF	0, 0, 0, 0, 0, 0, 7	(6!x 0!x 0!x 0!x 0!x 0!x 0!x 1!x)x(0!x 0!x 0!x 0!x 0!x 0!x 7!x)	0.000008	13

for all 20 types of amino acids is 192, thus the rank per amino acid is 192/236 = 0.81. Similarly, there are 21 prolines ("P"s) in bovine bcl-2 and the rank for "P"s is 129, thus the rank per each "P" is 129/21 = 6.14. In these manners, we can compare the impact of chance on amino-acid distributions in primary structure across a protein family.

Statistical inference regarding the randomness in protein primary structure

Statistical inference with respect to the construction of amino-acid sequences

The measured and predicted frequencies can be compared using the rounded integral value. The measured frequencies of two- and three-amino-acid sequences can be checked by the Kolmogorov-Smirnov normality test. In fact, a non-normal distribution would be expected; thus this test is used only for safety. Then 400 and 8000 random integral data (white noise) are produced ranging from the minimal to maximal frequency, the Mann-Whitney Rank sum test is used to compare the random data with measured frequencies. No measured frequencies for multi-amino-acid sequences are used in the above tests, because of the limitation in software and personal computer. $P<0.05$ is considered to be of statistical significance, and SigmaStat for Windows (version 2.0, 1992–1995 Jandel Corporation) is used to run all the tests.

The measured probability and predicted probability are compared using four decimal rounded values. Generally each type of 20 amino acids has a chance of 1/20 (P = 0.05) to repeat once, a two-amino-acid sequence has the chance of 1/400 (P = 0.0025) to repeat once, and a three-amino-acid sequence has the chance of 1/8000 (P = 0.000125) to repeat once in the protein primary structure. In case of human nitric-oxide synthase type I form, there are 127 leucines ("L"s), which is the most abundant amino acid, and 21 tryptophans ("W"s), which is the least amino acid. If the first amino acid is "L", then the chance of the second amino acid to be "L" is 126/1433 (P = 0.0879 > 0.05), if the first amino acid is "W", then the chance of the second amino acid to be "W" is 20/1433 (P = 0.0140 < 0.05). With respect to multi-amino-acid sequences, the chance of appearance of two-amino-acid sequence of "LL" is 127/1434 x 126/1433 (P = 0.0078 < 0.01); the chance of the first repetition of "LL" is 125/1432 x 124/1431 (P = 0.0076 < 0.01); the chance of appearance of three-amino-acid sequence of "LLL" is 127/1434 x 126/1433 x 125/1432 (P = 0.0007 < 0.001) and the chance of the first repetition of "LLL" is 124/1431 x 123/1430 x 122/1429 (P = 0.0006 < 0.001).

Statistical inference with respect to the randomness of follow-up amino acid

The Markov chain transition probability and conditional probability are compared using four decimal rounded values. Generally the statistical inference in this random calculation is similar to the previous one, i.e. each type of 20 amino acids has a chance of 1/20 (P = 0.05) to follow a preceding amino acid. For example, there are 58 "N"s in human nitric-oxide synthase type I form, the probabilities of first appearance of "N" in a two- and three-amino-acid sequence are 58/1433 (P = 0.0405) and 58/1432 (P = 0.0405), respectively. Similarly the probabilities of second appearance of "N" in a two- and three-amino-acid sequence are 57/1431 (P = 0.0398) and 57/1429 (P = 0.0399), respectively.

Statistical inference with respect to the distributions of amino acids along a protein

As each type of amino acids 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", "F", "FF" and "FF" is 52 times (0.321295 /0.006120) more likely to occur than the distribution of "F", "F", "F", "F", "F", "F" and "F" (Table 2).

General meanings of random approach

Present and absent amino-acid sequences

With respect to the amino-acid sequence being present in a protein, if the measured probability/frequency matches the predicted probability/frequency, the occurrence of amino-acid sequences can be considered to be random, otherwise the presence of amino-acid sequences cannot be explained by purely random mechanism.

With respect to the amino-acid sequence being absent from a protein, if the measured probability/frequency matches the predicted probability/frequency, the absence of amino-acid sequences can be considered to be random; if not, the absence of amino-acid sequences cannot be explained by purely random mechanism.

Follow-up amino acid

If the Markov chain transition probability matches the conditional probability, the amino acid following certain preceding amino acid/amino acids can be explained by a purely random mechanism. By the contrary, if the Markov chain transition probability differs from the conditional probability, the follow-up amino acid cannot be explained by a purely random mechanism.

Amino-acid distribution along a protein

If a type of amino acids has a highest distribution probability or lowest distribution rank, it means that these amino acids adopt the most-likely-to-occur distribution, or they distribute in the probabilistically simplest way along a protein. Otherwise, the amino acids has lower distribution probability or higher distribution rank indicating that the chance has less impact on their distribution in a protein.

Application for analysis on protein primary structure by random principle

Amino-acid sequences and protein construction

It is interesting that some kinds of amino-acid sequences are predictable for their presence or absence according to a purely random mechanism. About 47% absent and 23% present two-amino-acid sequences show that their frequencies can be explained by purely random mechanism (Table 3). This means that a protein sequence can be divided into "random" and "non-random" subsequences, i.e. randomly predictable subsequence and randomly unpredictable subsequence.

Each repetition has the probability of 1/400 in two-amino-acid sequence, of 1/8000 in three-amino-acid sequence, and of 1/160 000 in four-amino-acid sequence, thus the amino-acid sequences, which appear more than once, are particularly interesting, because these repetitions would not be considered to occur by chance, they may have some non-random reason underlying.

Generally speaking, no occurrence of multi-amino-acid sequences can be predicted by a purely random mechanism, because of the low probability to occur. But on the other hand, all the non-occurrence of multi-amino-acid sequences can be predicted by a purely random mechanism as well as because of the low probability to occur, which leaves an interesting question of whether the reason that most multi-amino-acid sequences are not selected for the construction of a protein is due to a purely random mechanism.

The amino-acid sequences that do not match the predicted frequency are important, especially when the difference between measured and predicted frequencies is equal to or larger than two, because the predicted frequency is the rounded value of the predicted probability and the difference, being equal to one, may be due to the rounding error.

Table 3 Two-amino-acid sequences in different proteins

Protein	absent pairs			present pairs		times of appearance										
	total	state A	state B	total	state A	state B	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Human AML-1[a]	182 (45.50)	115 [63.19]	52 [28.57]	452	77 (19.25)	41 (10.25)	100 (25.00)	63 (15.75)	27 (6.75)	15 (3.75)	1 (0.25)	7 (1.75)	4 (1.00)	0	0	1 (0.25)	0	0
Human ADH1[b]	191 (47.75)	119 [62.30]	52 [27.23]	373	92 (23.00)	32 (8.00)	111 (27.75)	60 (15.00)	20 (5.00)	9 (2.25)	8 (2.00)	1 (0.25)	0	0	0	0	0	0
Human CTGF[c]	205 (51.25)	113 [55.12]	50 [24.39]	348	81 (20.25)	21 (5.25)	98 (24.50)	60 (15.00)	23 (5.75)	10 (2.50)	3 (0.75)	1 (0.25)	0	0	0	0	0	0
Human GR[d]	147 (36.75)	61 [41.50]	29 [19.73]	521	107 (26.75)	46 (11.50)	114 (28.50)	70 (17.50)	40 (10.00)	13 (3.50)	7 (1.75)	5 (1.25)	3 (0.75)	0	1 (0.25)	0	0	0
Human MAO-B[e]	142 (35.5)	37 [26.06]	9 [6.34]	518	97 (24.25)	52 (13.00)	124 (31)	69 (17.25)	32 (8)	15 (3.75)	12 (3)	3 (0.75)	2 (0.5)	1 (0.25)	0	0	0	0
Human PAF-AH- α[f]	178 (44.50)	72 [40.45]	0 [0]	408	95 (23.75)	45 (11.25)	117 (29.25)	57 (14.25)	27 (6.75)	13 (3.25)	4 (1.00)	4 (1.00)	0	0	0	0	0	0
Human TNF- α[g]	243 (60.75)	176 [72.43]	42 [17.28]	232	64 (16.00)	24 (6.00)	105 (26.25)	34 (8.50)	14 (3.50)	3 (0.75)	1 (0.25)	0	0	0	0	0	0	0
Human tyrosinase[h]	114 (28.50)	15 [13.16]	0 [0.00]	528	156 (39.00)	75 (18.75)	142 (35.50)	82 (20.50)	42 (10.50)	13 (3.25)	4 (1.00)	2 (0.50)	0	0	0	0	0	1 (0.25)
Human TAT[i]	147 (36.75)	56 [38.10]	18 [12.24]	453	119 (29.75)	65 (16.25)	138 (34.50)	65 (16.25)	33 (8.25)	9 (2.25)	3 (0.75)	3 (0.75)	0	1 (0.25)	0	1 (0.25)	0	0
Bovine p53[j]	187 (46.75)	98 [52.41]	25 [13.37]	385	80 (20.00)	32 (8.00)	120 (30.00)	51 (12.75)	21 (5.25)	13 (3.25)	4 (1.00)	2 (0.50)	1 (0.25)	0	1 (0.25)	0	0	0
mouse p53[k]	188 (47.00)	86 [45.75]	32 [17.02]	389	85 (21.25)	27 (6.75)	112 (28.00)	52 (13.00)	29 (7.25)	14 (3.5)	2 (0.50)	1 (0.25)	2 (0.50)	0	0	0	0	0
Sheep p53[l]	194 (48.50)	102 [52.58]	19 [9.79]	381	82 (20.50)	35 (8.75)	108 (27.00)	56 (14.00)	22 (5.50)	14 (3.50)	2 (0.50)	2 (0.50)	0	1 (0.25)	1 (0.25)	0	0	0
AMPC-CITFR[m]	184 (46.00)	89 [48.37]	41 [22.28]	380	99 (24.75)	33 (8.25)	111 (27.75)	65 (16.25)	23 (5.75)	16 (4.00)	0	1 (0.25)	0	0	0	0	0	0

The states A and B indicate that two-amino-acid sequences can be explained by the predicted frequency and the predicted probability according to a purely random mechanism, respectively.
() and [] are the percentages calculated from 400 possible two-amino-acid sequences and from total number of corresponding two-amino-acid sequences, respectively.
[a] Wu and Yan, 2000c; [b] Wu, 2000e; [c] Wu and Yan, 2001b; [d] Wu, 2000a; [e] Wu and Yan, 2001a; [f] Wu and Yan, 2000d; [g] Wu, 2000b; [h] Wu and Yan, 2002a; [i] Wu, 2000c; [j] Wu and Yan, 2002b; [k] Wu, 2000g; [l] Wu, 2000d; [m] Wu and Yan, 2000a.

The amino-acid sequences are functionally and evolutionary biased, i.e. a protein favours certain repeated sequences, although the current knowledge does not provide much correlation of repeated sequences with knowing sites in the protein. For example, most of repeated three-amino-acid sequences of human tyrosinase are located in luminal domain (from 19 to 476), one of "LLG" is located in the transmembrane region (from 477 to 497), one of "LME" and one of "MEK" are located in cytoplasmic domain (from 498 to 529) (Wu and Yan 2002a).

If a possible amino-acid subsequence does not appear in a protein, naturally it is not needed for its function. Of these "useless" sequences, some can be explained by a purely random mechanism and some cannot, which also differs from the traditional view. It appears likely from such analyses that the protein may not be as highly structured as was previously thought. This leaves an intriguing question of whether the inclusion and exclusion of non-random sequences in a protein may lead to functional changes, which, however, requires scanning a total family of proteins to draw a firm conclusion.

Follow-up amino acid and protein construction

In human glutathione reductase, of 477 measured first-order Markov chain transition probabilities for the second amino acid in two-amino-acid sequences, one (0.210%) measured first order Markov chain transition probability matches the predicted conditional probability and can be explained by a purely random mechanism (Wu, 2000a). However, the vast majority of the measured Markov chain transition probabilities do not match the predicted conditional probabilities implicating that follow-up amino acid is generally not arbitrary. The results from Tables 4, 5 and 6 show that the Markov chain transition probability increases from two-amino-acid sequences to multi-amino-acid sequences, therefore the random chance for an amino acid to follow an arbitrary amino acid/amino acids decreases with the increase in the length of amino-acid sequence.

Amino-acid distribution and protein construction

In fact, the number of possibly theoretical distributions increases dramatically as the number of amino acids increases. For example, there are three types of possibly theoretical distributions for three amino acids distributing in three parts, 11 types for six amino acids distributing in six parts, 54 types for 11 amino acids distributing in 11 parts. Although there are many possible distributions for a type of amino acids in a protein (such as 15 possible distributions for seven amino acids in Table 2), a protein adopts only one possible distribution during its evolutionary process. Therefore there is solely one distribution probability/rank for each type of amino acids and a

Table 4 Measured frequency, predicted frequency and first-order Markov chain transition probability (MCTP) of two-amino-acid sequences, which have a difference ≥2 between counted and predicted frequencies in human alcohol dehydrogenase α-chain (Wu, 2000e).

Amino-acid sequence	Measured frequency	Predicted frequency	MCTP
AA	5	2	0.161
AE	0	2	0.000
AK	5	3	0.161
AV	6	3	0.194
CG	4	2	0.250
CK	3	1	0.187
EV	5	2	0.250
FS	5	1	0.357
GF	3	1	0.083
GP	0	2	0.000
GT	5	2	0.139
HE	3	0	0.375
HG	1	3	0.031
IN	3	1	0.115
KP	4	2	0.125
KT	0	2	0.000
LG	5	3	0.172
LL	4	2	0.138
LV	1	3	0.034
NP	4	1	0.364
PA	0	2	0.000
PQ	3	0	0.150
QD	2	0	0.286
RI	3	1	0.333
SG	2	0	0.087
ST	4	1	0.174
TK	0	2	0.000
VA	5	3	0.139
VI	5	3	0.139
VK	1	3	0.028

Table 5 Measured frequency, predicted frequency and second-order Markov chain transition probability (MCTP) of three-amino-acid sequences, which have a difference ≥2 between counted and predicted frequencies in sheep p53 protein (Wu, 2000d).

Amino-acid sequence	Measured frequency	Predicted frequency	MCTP
AQA	2	0	1.000
EPP	3	0	0.750
EYF	2	0	1.000
GNL	2	0	0.667
KKG	2	0	0.333
KKP	2	0	0.333
LAP	2	0	0.667
LLP	2	0	0.500
LSS	2	0	0.500
NEA	2	0	1.000
NLL	3	0	0.750
PEP	2	0	0.500
PLS	2	0	0.667
PPG	2	0	0.250
PPP	3	0	0.375
RSS	2	0	0.667
SPS	2	0	0.333
SSF	2	0	0.222

Table 6 Appeared-more-than-once four-, five-, six-, seven-, eight-, and nine-amino-acid sequences, their measured frequency, predicted frequency and Markov chain transition probability (MCTP) in human platelet-activating factor acetylhydrolase α-subunit (Wu and Yan, 2000d).

Amino-acid sequence	Measured frequency	Predicted frequency	MCTP
ASCS	2	0	1.000
DKTI	2	0	0.667
DQTV	2	0	1.000
IKMW	2	0	1.000
KTIK	2	0	1.000
RDKT	2	0	1.000
SRDK	2	0	1.000
TIKM	2	0	0.500
VSAS	2	0	1.000
VWDY	2	0	1.000
DKTIK	2	0	1.000
KTIKM	2	0	1.000
RDKTI	2	0	1.000
SRDKT	2	0	1.000
TIKMW	2	0	1.000
DKTIKM	2	0	1.000
KTIKMW	2	0	1.000
RDKTIK	2	0	1.000
SRDKTI	2	0	1.000
DKTIKMW	2	0	1.000
RDKTIKM	2	0	1.000
SRDKTIK	2	0	1.000
RDKTIKMW	2	0	1.000
SRDKTIKM	2	0	1.000
SRDKTIKMW	2	0	1.000

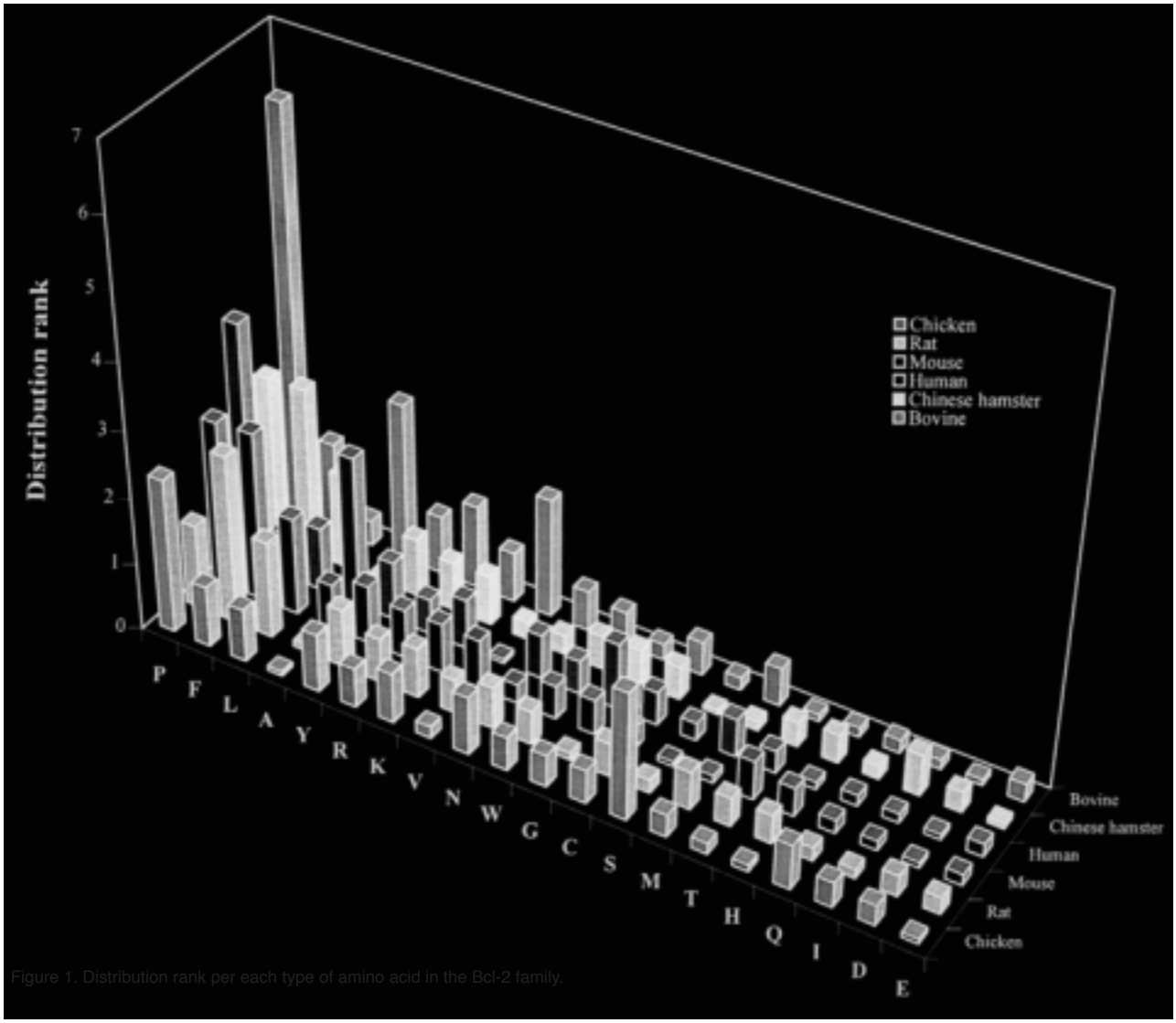


Figure 1. Distribution rank per each type of amino acid in the Bcl-2 family.

maximum of 20 types of distribution probability/rank in a protein. Using the distribution probability/rank as a measure, one can analyse the effects of chance on amino-acid distribution in proteins.

It is found that the distribution probabilities for several types of amino acids are at or very near to the probabilistically simplest distribution in our studies. For example, about one fourth types of amino acids distribute in the highest probabilistic way in the Bcl-2 family (Figure 1). These phenomena are striking in different functional regions (four BH regions and transmembrane region), where vast majority of amino acids occur with the probabilistically simplest distribution.

In our studies, we observed that the distribution probabilities for most amino acids and amino-acid sequences are at or near the probabilistically simplest distribution. This is not a homogenous distribution along a protein, but rather a heterogeneous distribution, which with a high distribution probability may provide the base for the protein function. On the other hand, the distribution probability for the homogeneous distribution is very low (Table 7), so it is not easy to adapt the homogeneous

Table 7. Homogenous distribution probability with respect to the different numbers of amino acids

Number of amino acid	Probability	Number of amino acid	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}		

distribution from the probabilistic viewpoint during the evolutionary process. Thus the tendency for amino acids clustering in the protein primary structure results from the necessity of both function and randomness.

Application for mutation analysis by random principle

Amino-acid sequences and mutation

As it has been realised for many years that most mutations causing changes in amino-acid sequence are of no consequence for protein function, the results from our studies raises an intriguing issue of whether or not these unharmed mutations occur in the "random" subsequences. Logically, mutations in randomly predicted subsequences are most likely to occur spontaneously. After determined if a mutation occurs in predictable or unpredictable subsequences, a mutation in predictable subsequences is unlikely to lead to a substantial change in protein function, otherwise a mutation in unpredictable subsequences is likely to result in a substantial change in protein function. For example, there are two mutations in rat monoamine oxidase B form. One mutation occurs at position 139 changing "L" to "H" which leads to four two-amino-acid sequences changed, i.e. "PL" -> "PH" and "LA" -> "HA" because amino acids at positions 138 and 140 are "P" and "A", respectively. Our results show that all these four changes belong to predictable subsequences, accordingly this mutation may not lead to a substantial change in enzymatic function. This is true because the mutation does not change the substrate affinity. Another mutation occurs at position 199 changing "I" to "F" which leads to four two-amino-acid sequences changed, i.e. "II" -> "IF" and "IS" -> "FS" because amino acids at positions 198 and 200 are "I" and "S", respectively. Our results show that the "IS" is in an unpredictable subsequence, accordingly the mutation may lead to a substantial change in enzymatic function, which is true because the mutation increases the affinity for serotonin and tyramine (Wu and Yan, 2001a)

Table 8 Changes in first-order Markov transition probability of appeared-more-than-once two-amino-acid sequences in human tyrosine aminotransferase and its variant (Wu, 2000c)

Sequence	Normal	Variant
GA	0.100	0.103
GR	0.067	0.069
GN	0.100	0.103
GD	0.067	0.069
GG	0.067	0.067
GI	0.067	0.069
GL	0.100	0.069
GK	0.067	0.069
GF	0.067	0.069
GS	0.067	0.069
GW	0.067	0.069
PG	0.188	0.156
PV		0.063
VA	0.065	0.063
VR	0.065	0.063
VN	0.065	0.063
VG	0.065	0.063
VH	0.065	0.063
VI	0.065	0.063
VL		0.063
VK	0.097	0.094
VF	0.097	0.094
VP	0.194	0.188

Follow-up amino acid and mutation

Mutations can lead to the increase or decrease in Markov chain transition probability, for example, although a single amino acid is changed in the variant of human tyrosine aminotransferase (G -> V at position 362, Natt *et al*., 1992), two appeared-more-than-once sequences (PV and VL) appear in the variant, the measured probability is only changed in four appeared-more-than-once sequences (GL, PG, PV and VL), however 23 changes are found in the first-order Markov chain transition probability (Table 8), so it seems that the measure of Markov chain transition probability is more sensible affected by mutation.

Mutations can change the value of Markov chain transition probability, and may consequently affect on protein function, such as the above variant causing tyrosinemia type II. Taking human haemoglobin α-chain as another example, there are three variants leading to low O_2 affinity (Table 9).

Amino-acid distribution and mutation

Mutations can increase or decrease in the distribution probability/rank of composed amino acids, consequently changing in protein function, which phenomena are revealed by our approaches. For example, there are seven variants in human haemoglobin α-chain causing α-thalassemia (Table 10). Variants 1 and 3 target the distribution probability increased in both affected amino acids, while the others change the distribution probability of affected amino acids to opposite directions. In variants 4, 5 and 7, the distribution probability is lower in replaced

Table 9 First-order Markov chain transition probability of two-amino-acid sequences which appear more than once in human haemoglobin α-chain and its three variants (Wu, 1999).

Sequence	Normal	Variant 1	Variant 2	Variant 3
AA	0.095	0.095	0.095	0.100[b]
AD	0.095	0.095	0.095	0.100[b]
AE	0.095	0.095	0.095	0.100[b]
AH	0.190	0.190	0.190	0.200[b]
AL	0.190	0.190	0.190	0.200[b]
AS	0.095	0.095	0.095	[a]
AV	0.095	0.095	0.095	0.100[b]
NA	0.500	0.500	0.500	0.500
DL	0.250	0.250	0.250	0.220[b]
DK	0.250	0.250	0.250	0.220[b]
GA	0.286	0.286	0.286	0.286
GK	0.286	0.286	0.286	0.286
HA	0.300	0.300	0.330[b]	0.300
HG	0.200	0.200	[a]	0.200
LA	0.111	0.111	0.111	[a]
LD				0.111[a, b]
LL	0.111	0.111	0.111	0.111
LS	0.333	0.333	0.333	0.333
LT	0.111	0.111	0.111	0.111
KL	0.182	0.182	0.182	0.182
KT	0.182	0.182	0.182	0.182
KV	0.182	0.182	0.182	0.182
FL	0.286	0.286	0.286	0.286
FP	0.286	0.286	0.286	0.286
PA	0.429	0.429	0.429	0.429
SA	0.182	0.182	0.182	0.182
SH	0.182	0.182	0.182	0.182
TN	0.222	0.222	0.222	0.222
YG			0.500[a, b]	
VA	0.154	0.167[b]	0.154	0.154
VD	0.154	0.167[b]	0.154	0.154
VL	0.154	[a]	0.154	0.154
VK	0.154	0.167[b]	0.154	0.154

[a] and [b], there are differences in sequence and probability between variant and normal haemoglobins, respectively. Variant 1: V -> E in 1 (Vasseur *et al*., 1992); Variant 2: H -> Y in 58 (Bairoch and Apweiler, 1999); Variant 3: A -> D in 130 (Fujisawa *et al*., 1992).

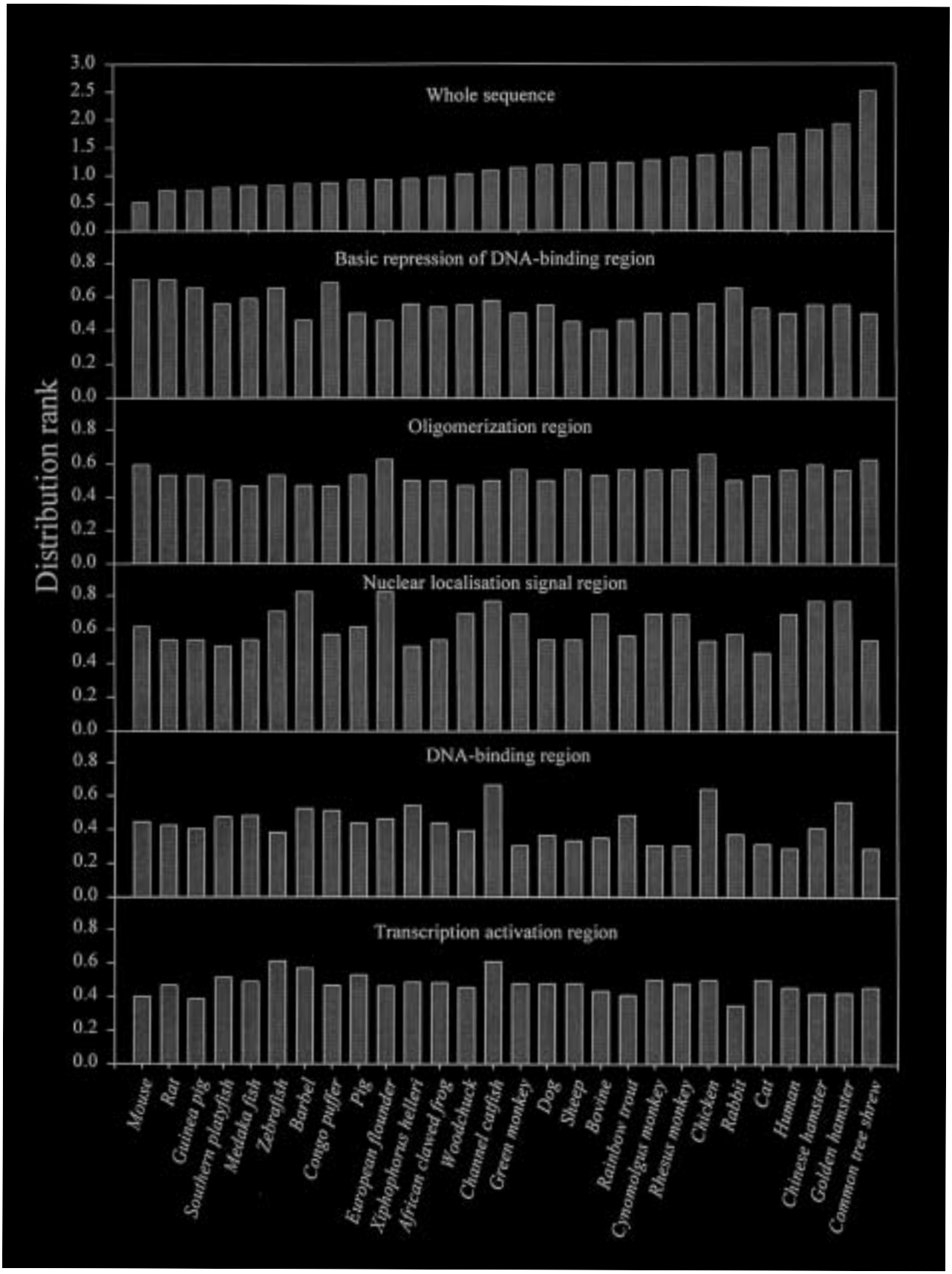


Figure 2. Distribution rank per amino acid in the whole sequence and in different functional regions (the transcription activation region, the DNA-binding region, the nuclear localisation signal region, the oligomerization region and the repression of DNA-binging region) across the p53 family.

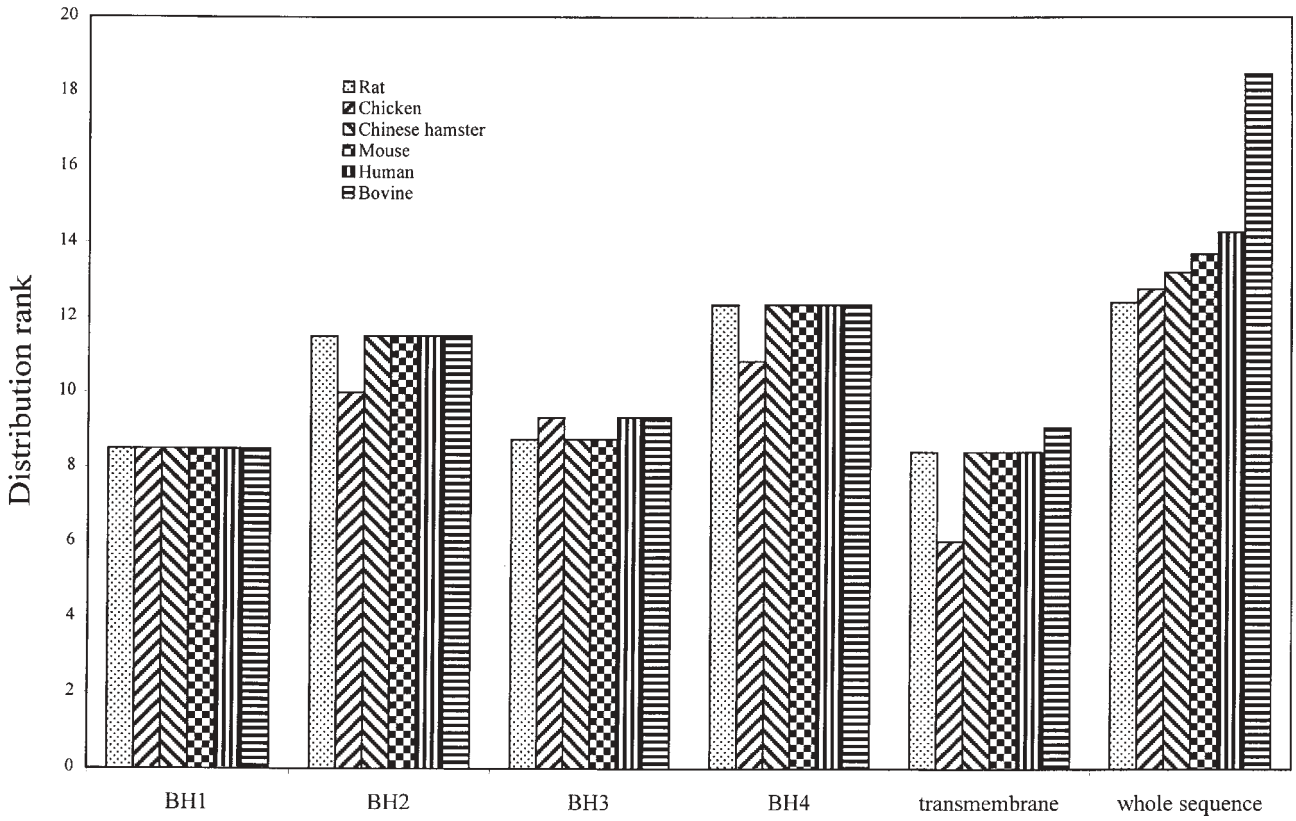


Figure 3. Distribution rank per amino acid in the whole sequence and different functional regions (transmembrane region, BH1–4 regions) across the Bcl-2 family.

Table 10 Amino acids and their numbers, measured distribution probability (MDP), and rank in theoretical distribution probability (RTDP) in seven variants causing α-thalassemia compared with those in normal human haemoglobin α-chain in parentheses (Wu and Yan, 2000b)

Type of variant	Amino acid	Number		MDP		RTDP	
Variant 1[a]	D	9	(8)	0.1967	(0.0673)	1	(5)
G -> D in 59	G	6	(7)	0.1543	(0.0268)	3	(7)
Variant 2[b]	R	4	(3)	0.5625	(0.2222)	1	(2)
L -> R in 109	L	17	(18)	0.0183	(0.0831)	16	(2)
Variant 3[c]	A	20	(21)	0.0965	(0.0273)	1	(11)
A -> D in 110	D	9	(8)	0.1770	(0.0673)	3	(5)
Variant 4[d]	L	17	(18)	0.0366	(0.0831)	9	(2)
L -> P in 125	P	8	(7)	0.2243	(0.1585)	2	(3)
Variant 5[e]	L	17	(18)	0.0366	(0.0831)	9	(2)
L -> P in 129	P	8	(7)	0.2523	(0.1585)	1	(3)
Variant 6[f]	P	8	(7)	0.2523	(0.1585)	1	(3)
S -> P in 131	S	10	(11)	0.1905	(0.2020)	1	(1)
Variant 7[g]	L	17	(18)	0.0366	(0.0831)	9	(2)
L -> P in 136	P	8	(7)	0.2523	(0.1583)	1	(3)

[a] Curuk *et al.*, 1993; [b] Sanguansermsri *et al.*, 1979; [c] Honig *et al.*, 1981; [d] Zeng *et al.*, 1988; Liang *et al.*, 1991; Ko *et al.*, 1993; [e] Darbellay *et al.*, 1995; [f] Wajcman *et al.*, 1993; [g] Harkness *et al.*, 1990.

Table 11. Distribution rank per amino acid in mutations compared with that of normal rat MAO-A

	Amino acid	Position	Distribution rank	Difference	
Mutation 1	F -> A	208	0.848	↓	3%
Mutation 2	F -> I	208	0.886	↑	1.26%
Mutation 3	F -> V	208	0.922	↑	5.37%
Mutation 4	F -> Y	208	0.875	->	0

↑, ↓ and -> present the rank increasing, decreasing and unchanged in comparison with normal rat MAO-A, respectively.

amino acids and higher in replacing ones. On the contrary, variants 2 and 6 lead to the distribution probability increased in substituted amino acids and decreased in substituting ones (Wu and Yan, 2000b). From probabilistic viewpoint, the formation of variants 1 and 3 is more likely spontaneous, as the amino-acid distributions in these variants shift to higher probabilistic direction, i.e. the amino acids distribute more randomly in these variants.

There are four mutations documented in rat MAO-A changing “F” at position 208 to “A”, “I”, “V” and “Y”, respectively (Tsugeno and Ito, 1997). The distribution rank decreases in mutation 1 and increases in mutations 2 and 3, leading to a lower affinity for serotonin and tyramine. It is interesting to note that the distribution rank is unchanged in mutation 4, consistently no change in substrate affinity is found in mutation 4 (see Table 11).

In tumor suppressor p53 protein there are five functional regions documented in the Swiss-Protein databank, among them the DNA-binding region is quite longer (about 190 amino acids) than others, mutations in these regions may result in the p53 dysfunction and induce cancers, which is true in humans as the majority of such mutations are found in the DNA-binding region (Hollstein *et al*., 1991; de Vries *et al*., 1996). It is natural to ask the question why there are so many mutations in human p53. The results from our study may give some information to this issue (see Figure 2). In mouse p53, the distribution of amino acids reveals in a probabilistically easy way, as the distribution rank is about the same in the whole sequence and in different functional regions as well, which may represents that the primary structure of mouse p53 is more stable among the p53 family. Unlikely, the distribution probability of amino acids shows dual characters in human p53: the distribution ranks are lower in the functional regions, on the other hand this value is higher in the whole sequence (more than two times higher than that in mouse p53). Thus, from the distribution probabilistic viewpoint, the composition of human p53 is relatively stable in the functional regions rather than in the whole sequence. This contradictory feature may contribute to one of the potential effects on the mutations.

Application for evolutionary process by random principle

Logically, the functional regions in a protein should be considered to be deliberately evolved and conserved, thus the presence of amino-acid sequences in these regions is unlikely to be predictable by a purely random mechanism; whereas the non-functional regions in a protein are unlikely to be deliberately evolved and conserved, thus the presence of amino-acid sequences in these regions is possibly predictable by a purely random mechanism.

The speed of evolution in a species would catch up with the environmental changes, otherwise this species would die. This requires the nature selection to be of efficiency, 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. Thus the more the predictable amino-acid sequences in a protein, the more easily the protein is formed, and consequently the formation of this protein costs less time and energy. From Table 3 it can be found that the construction of human tyrosinase is more random than human acute myeloid leukaemia 1 protein, as about 42% two-amino-acid sequences occur randomly in the former one while only 25% in the later one.

By means of analysing amino-acid distributions in a protein family, we can get some insights into the effect of chance on the evolutionary process. Taking the p53 family as an example (Figure 2), 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 by 3.8-fold compared to that of mouse p53. This difference implies that the effect of chance has more impact on the amino-acid distribution of mouse p53 and less impact on that of common tree shrew p53.

Furthermore, the distribution probabilities for several types of amino acids are at or very near to the probabilistically simplest distribution in the protein family, for instance, 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 should not be deliberately evolved and conserved whereas the distributions with a low distribution probability should be deliberately evolved and conserved, because nature should be clever enough to spend the only necessary energy and time during evolution. Taking prolines (“P”s) as an example, their distributions reveal highly unpredictable in the primary stricture of Bcl-2 family, which may links to the specific necessary for Bcl-2 function. Variant 2 from human Bcl-2 modifies the most evolved and conserved distribution configuration of “P” by supplying another “P”, changing the rank of “P”s from 69 to 38, consequently, the dysfunctional Bcl-2 induces non-Hodgkin’s-lymphoma (Tanaka *et al*., 1992).

Although the distribution ranks vary from species to species, our results demonstrate that the distribution ranks are relatively lower in the functional regions (about 0.5 in average) than that in the whole sequences (about 1.2 in average) except for those in mouse p53 (Figure 2). Also this measure shows less difference between different species in all of the functional regions opposed to the whole sequence. Therefore the functional regions (especially the DNA-binding region) are not only conserved through the evolutionary process of species but also ranged in a probabilistically easy way, which guarantee the functional base of the p53 protein. Similar phenomena can be found in other protein families such as the Bcl-2 family (Figure 3).

Application for developing new drugs by random principle

By comparing the predicted probability/frequency with the measured probability/frequency, one can know which amino-acid "word" a protein favours. The most frequently appeared amino-acid sequences may serve as the potential targets for new drugs, because the drugs would have more chance to interact with them (Wu and Yan, 2002c). Also, the amino-acid sequences, which have the biggest difference between predicted and measured probabilities/ frequencies, could be the potential targets for new drugs, as they are highly evolved for the difference between predicted and measured probabilities/frequencies. Finally, a mutation is unlikely to occur at the amino acid with a high value of Markov chain transition probability. Thus the amino acids with high Markov chain transition probability may serve as the potential targets for new drugs, because they are unlikely to change into other amino acids.

Limitations

One may argue whether or not the chance affects significantly on the composition of primary structure of proteins because the changes in the values of probabilities or distribution ranks are not big as presented in the studies. We would like to call attention to the fact that the evolution is a long continuous and accumulating process, it would not be surprising that the similarity exists across the protein family and that some mutations and variants vary little from their origin.

At present, it is far too early to establish the definitive relationship between the protein primary structure and its dysfunction linking to certain disease, which still needs a considerable of work. However, the studies by our approaches may serve as a platform for further analyses.

At this stage, we are still unable to apply those approaches to analyse the DNA and RNA sequences, although it will certainly be our aim. Technically, the DNA and RNA sequences are much longer than its corresponding protein sequence, so this needs another algorithm. However, the comparison in measured and predicted probabilities/frequencies between DNA/RNA and its corresponding protein sequences will doubtless give more meaningful insights.

Conclusions

According to random principles, we explore three approaches to analyse protein primary structure, including the randomness in the construction of amino-acid sequences, in the follow-up amino acid, and in the distribution of amino acid/amino acids. (i) By comparing the measured probability/frequency with the predicted probability/frequency, we can evaluate the effect of chance on the composition of amino-acid sequences. (ii) By comparing the Markov chain transition probability with the predicted conditional probability, we can evaluate the effect of chance on the follow-up amino acid. (iii) By comparing the real distribution probability with the theoretical distribution probability, we can evaluate the effect of chance on the distribution of amino acids. These approaches can be used to quantitatively analyse the primary structure of intra-protein as well as inter-proteins, thus we can get more insights into the mechanisms of protein construction, mutation, and evolutionary process. Also, these approaches may have some potential use for developing new drugs.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Ash, R.B. 1965. Information Theory. Interscience. New York.

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.

Barker, W.C., Garavelli, J.S., Hou, Z., Huang, H., Ledley, R.S., McGarvey, P.B., Mewes, H.W., Orcutt, B.C., Pfeiffer, F., Tsugita, A., Vinayaka, C.R., Xiao, C., Yeh, L.S., and Wu, C. 2001. Protein Information Resource: a community resource for expert annotation of protein data. Nucleic Acids Res. 29: 29-32.

Benson, D.C. 1990. Fourier methods for biosequence analysis. Nucleic Acids Res. 18: 6305-6310.

Bradham, D.M., Igarashi, A., Potter, R.L., and Grotendost, G.R. 1991. Connective tissue growth factor: a cysteine-rich mitogen secreted by human vascular endothelial cells is related to the SRC-induced immediate early gene product CEF-10. J. Cell Biol. 114: 1285-1294.

Chou, K.C. 1999. Using pair-coupled amino acid composition to predict protein secondary structure content. J. Protein Chem. 18: 473-480.

Csiszar, I., and Krner, J. 1981. Information Theory. Academic Press. New York.

Curuk, M.A., Dimovski, A.J., Baysal, E. Gu, L.H., Kutlar, F., Molchanova, T.P., Webber, B.B., Altay, C., Gurgey, A., and Huisman, T.H. 1993. Hb Adana or alpha 2(59)(E8)Gly Asp beta 2, a severely unstable alpha 1-globin variant, observed in combination with the -(alpha)20.5 Kb alpha-thal-1 deletion in two Turkish patients. Am. J. Hematol. 44: 270-275.

Darbellay, R., Mach-Pascual, S., Rose, K., Graf, J., and Beris, P. 1995. Haemoglobin Tunis-Bizerte: a new alpha 1 globin 129 Leu Pro unstable variant with thalassaemic phonotype. Br. J. Haematol. 90: 71-76.

de Vries, E.M.G., Ricke, D.O., de Vries, T.N., Hartmann, A., Blaszyk, H., Liao, D., Soussi, T., Kovach, J.S., and Sommer, S.S. 1996. Database of mutations in the p53 and APC tumor suppressor genes designed to facilitate molecular epidemiological analyses. Hum. Mutat. 7:202-213.

Dequiedt, F., Kettmann, R., Burny, A., and Wilems, L. 1995a. Nucleotide sequence of the bovine P53 tumour-suppressor cDNA. DNA Seq. 5: 261-264.

Dequiedt, F., Kettmann, R., Burny, A., and Wilems, L. 1995b. Nucleotide sequence of the ovine P53 tumour-suppressor cDNA and its genomic organization. DNA Seq. 5: 255-259.

Everitt, B.S. 1999. Chance rules: an informal guide to

probability, risk, and statistics. Springer, New York.
Feller, W. 1968. An introduction to probability theory and its applications. 3rd edn, John Wiley and Sons, New York, Vol I, 38-40.
Fujisawa, H., Ogura, T., Kurashima, Y., Yokoyama, T., Yamashita, J., and Esumi, H. 1994. Expression of two types of nitric oxide synthase mRNA in human neuroblastoma cell lines. J. Neurochem. 63: 140-145.
Fujisawa, K., Hattori, Y., Ohba, Y., and Ando, S. 1992. Hb Yuda or alpha 130(H13)AlaAsp; a new alpha chain variant with low oxygen affinity. Hemoglobin 16: 435-439.
Grimsby, J., Chen, K., Wang, L.J., Lan N. C., and Shih, J.C. 1991. Human monoamine oxidase A and B genes exhibit identical exon-intron organization. Proc. Natl. Acad. Sci. U.S.A. 88: 3637-3641.
Hall, A.V., Antoniou, H., Wang, Y., Cheung, A.H., Arbus, A.M., Olson, S.L., Lu, W.C., Kau, C.L., and Marsden, P.A. 1994. Structural organization of the human neuronal nitric oxide synthase gene (NOS1). J. Biol. Chem. 269: 33082-33090.
Harkness, M., Harkness, D.R., Kutlar, F., Kutlar, A., Wilson, J.B., Webber, B.B., Codrington, J.F., and Huisman, T.H.J. 1990. Hb Sun Prairie or alpha(2)130(H13)Ala Prp beta 2, a new unstable variant occurring in low quantities. Hemoglobin 14: 479-489.
Hayakawa, H., Koike, G., and Sekiguchi, M. 1990. Expression and cloning of complementary DNA for a human enzyme that repairs O6-methylguanine in DNA. J. Mol. Biol. 213: 739-747.
Honig, G.R., Shamsuddin, M., Zaizov, R., Steinherz, M., Solar, I., and Kirschmann, C. 1981. Hemoglobin Petah Tikva (alpha 110 ala replaced by asp): a new unstable variant with alpha-thalassemia-like expression. Blood 57: 705-711.
Hollstein, M., Sidransky, D., Vogelstein, B. and Harris, C.C. 1991. p53 mutations in human cancers. Science 253:49-53.
Jenkins, J.R., Rudge, K., Redmond, S., and Wade-Evans, A. 1984. Cloning and expression analysis of full-length mouse cDNA sequences encoding the transformation associated protein p53. Nucleic Acids Res. 12: 5609-5626.
Jonassen, I. 1997 Efficient discovery of conserved patterns using a pattern graph. Comput. Appl. Biosci. 13: 509-522.
Karlin, S., Bucher, P., Brendel, V., and Altschul, S.F. 1991. Statistical methods and insights for protein and DNA sequences. Annu. Rev. Biophys. Biophys. Chem. 20: 175-203
Ko, T.M., Tseng, L.H., Chuang, S.M., Hsieh, F.J., and Lee, T.Y. 1993. Rapid detection of human hemoglobin Quong Sze by polymerase chain reaction. J. Formos. Med. Assoc. 92: 88-90.
Koike, G., Maki, H., Takeya, H., Hayakawa, H., and Sekiguchi, M. 1990. Purification, structure, and biochemical properties of human O^6-methylguanine-DNA methyltransferase. J. Biol. Chem. 265: 14754-14762
Levanon D., Negreanu V., Bernstein Y., Bar-Am I., Avivi L., and Groner Y. 1994. AML1, AML2, and AML3, the human members of the runt domain gene- family: cDNA structure, expression, and chromosomal localization. Genomics 23: 425-432.
Liang, S., Wen, X.J., and Lin, W.X. 1991. Detection of the Hb Quong Sze mutation in a Chinese family by selective amplification of the alpha 2-globin gene and restriction map analysis with Msp I. Hemoglobin 15: 535-540.
Lindberg, F., and Normark, S. 1986. Sequence of the *Citrobacter Freundii* OS60 chromosomal ampC beta-lactamase gene. Eur. J. Biochem. 156: 441-445.
Liu, W.M, and Chou, K.C. 1999. Prediction of protein secondary structure content. Protein Eng. 12: 1040-1050.
Lo Nigro, C., Chong, S.S., Smith, A.C.M., Dobyns, W.B., Carrozzo, R., and Ledbetter, D.H. 1997. Point mutations and an intragenic deletion in LIS 1, the lissencephaly causative gene in isolated lissencephaly sequence and Miller-Dieker syndrome. Hum. Mol. Genet. 6: 157-164.
Matsuo, Y., and Yokoyama, S. 1989. Molecular structure of the human alcohol dehydrogenase 1 gene. FEBS letters 243: 57-60.
Natt E., Kida K., Odievre M., di Rocco M., and Scherer G. 1992. Point mutations in the tyrosine aminotransferase gene in tyrosinemia type II. Proc. Natl. Acad. Sci. U.S.A. 89: 9297-9301.
Pennica, D., Nedwin, G.E., Hayflick, J.S., Seeburg, P.H., Derynck, R., Palladino, M.A., Kohr, W.J., Aggarwal, B.B., and Goeddel, D.V. 1984. Human tumour necrosis factor: precursor structure, expression and homology to lymphotoxin. Nature 312: 724-729.
Popov, O., Segal, DM., and Trifonov, E.N. 1996. Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems 38: 65-74.
Rettenmeier, R., Natt, E., Zentgraf, H., and Scherer, G. 1990. Isolation and characterization of the human tyrosine aminotransferase gene. Nucleic Acids Res. 18: 3853-3861.
Sanguansermsri, T., Matragoon, S., Changgloah, L., and Flatz, G. 1979. Hemoglobin Suan-Dok (alpha 2 109 (G16) Leu replaced by Arg beta 2): an unstable variant associated with alpha-thalassemia. Hemoglobin 3: 161-174.
Tanaka, S., Louie, D.C., Kant, J.A., and Reed, J.C. 1992. Frequent incidence of somatic mutations in translocated BCL2 oncogenes of non-Hodgkin's lymphomas. Blood 79: 229-237.
Takeda, A., Tomita, Y., Okinaga, S., Tagami, H., and Shibahara, S. 1989. Functional analysis of the cDNA encoding human tyrosinase precusor. Biochem. Biophys. Res. Commun. 162: 984-990.
Tsugeno Y., and Ito A. 1997. A key amino acid responsible for substrate selectivity of monoamine RT oxidase A and B. J. Biol. Chem. 272:14033-14036.
Tutic, M., Lu, X.A., Schirmer, R.H., and Werner, D. 1990. Cloning and sequencing of mammalian glutathione reductase cDNA. Eur. J. Biochem. 188: 523-528.
van der Lubbe, J.C.A. 1997. Information theory. Cambridge University Press. Cambridge.
Vasseur, C., Blouquit, Y., Kister, J., Prome, D., Kavanaugh, J.S., Rogers, P.H., Guillemin, C., Arnone, A., Galacteros, F., Poyart, C. Rosa, J., and Wajcman, H. 1992. Haemoglobin Thionville. An alpha-chain variant with a substitution of a glutamate for valine at NA-1 and having an acetylated methionine NH2 terminus. J Biol Chem.

267: 12682-12691.

Wajcman, H., Vasseur, C., Blouquit, Y., Rosa, J., Labie, D., Najman, A., Reman, O., Leporrier, M., and Galacteros, F. 1993. Unstable alpha-chain hemoglobin variants with factitious beta-thalassemia biosynthetic ratio; Hb Questembert (alpha 131[H14]Ser Pro) and Hb Caen (alpha 132[H15]Val Gly). Am. J. Hematol. 42: 367-374.

Wu, G. 1999. The first and second order Markov chain analysis on amino acids sequence of human haemoglobin α-chain and its three variants with low O_2 affinity. Comp. Haematol. Int. 9: 148-151.

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. 2000c. The first, second and third order Markov chain analysis on amino acids sequence of human tyrosine aminotransferase and its variant causing tyrosinemia type II. Pädiatr. Grenzgeb. (Pediatrics and related topics). 39: 37-47.

Wu, G. 2000d. Frequency and Markov chain analysis of the amino-acid sequence of sheep p53 protein. J. Biochem. Mol. Biol. Biophys. 4: 179-185.

Wu, G. 2000e. Frequency and Markov chain analysis of the amino acid sequence of human alcohol dehydrogenase -chain. Alcohol Alcohol. 35: 302-306.

Wu, G. 2000f. The first, second, third and fourth order Markov chain analysis on amino acids sequence of human dopamine -hydroxylase. Mol. Psychiatry. 5: 448-451.

Wu, G. 2000g. Frequency and Markov chain analysis of amino-acid sequences of mouse p53. Hum. Exp. Toxicol. 19: 535-539.

Wu, G, and Yan, S.-M. 2000a. 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, and Yan, S.-M. 2000b. Prediction of distributions of amino acids and amino acid pairs in human haemoglobin -chain and its seven variants causing α-thalassemia from their occurrences according to the random mechanism. Comp. Haematol. Int. 10: 80-84.

Wu, G, and Yan, S.-M. 2000c. Prediction of two- and three-amino acid sequence of human acute myeloid leukemia 1 protein from its amino acid composition. Comp. Haematol. Int. 10: 85-89.

Wu, G, and Yan, S.-M. 2000d. Frequency and Markov chain analysis of amino-acids sequence of human platelet-activating factor acetylhydrolase α-subunit and its variant causing the lissencephaly syndrome. Pädiatr Grenzgeb (Pediatrics and related topics) 39: 513-526.

Wu, G, and Yan, S.-M. 2001a. 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. Biomol. Eng. 18: 23-27.

Wu, G, and Yan, S.-M. 2001b. Frequency and Markov chain analysis of amino-acid sequences of human connective tissue growth factor. J. Mol. Model. 5: 120-124.

Wu, G, and Yan, S.-M. 2001c. Analysis of distributions of amino acids, amino acid pairs and triplets in human insulin precursor and four variants from their occurrences according to the random mechanism. J. Biochem. Mol. Biol. Biophys. 5: 293-300.

Wu, G, and Yan, S.-M. 2001d. Analysis of distributions of amino acids and amino acid pairs in human tumor necrosis factor precursor and its eight variants according to random mechanism. J. Mol. Model. 7: 318-323.

Wu, G, and Yan, S.-M. 2002a. Prediction of presence and absence of two- and three-amino-acid sequence of human tyrosinase from their amino acid composition and related changes in human tyrosinase variant causing oculocutaneous albinism. Pädiatr. Grenzgeb. (Pediatrics and related topics) (in press).

Wu, G, and Yan, S.-M. 2002b. 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. Mol. Biol. Today (in press).

Wu, G, and Yan, S.-M. 2002c. Mathematical model of time needed for the immune system to detect and kill cancer cells in blood. Comp. Clin. Pathol. 3: 31-37.

Zeng, Y.T., Huang, S.Z., and Chennnnnn, M.J. 1988. The types and distribution of -thalassemia-2 in China. Hemoglobin 12: 455-458.