J Mol Model (2001) 7:120–124
DOI 10.1007/s008940100022

ORIGINAL PAPER

Guang Wu · Shao-Min Yan

Frequency and Markov chain analysis of amino acid sequences of human connective tissue growth factor

Received: 12 November 1999 / Accepted: 15 January 2001 / Published online: 17 May 2001

Abstract The amino acid sequence of human connective tissue growth factor was measured according to two, three- and four-amino-acid sequences. The measured frequency and probability were compared with predicted frequency and probability. In human connective tissue growth factor, 81 (23.276%) and 21 (6.034%) of 348 two-amino-acid sequences can be explained by the predicted frequency and probability according to a purely random mechanism, 113 (55.122%) and 50 (24.390%) of 205 non-appearing two-amino-acid sequences can be explained by the predicted frequency and probability according to a purely random mechanism; no measured Markov transition probability for the second amino acid in two-amino-acid sequences matches the predicted conditional probability.

Keywords Amino acid sequence · Human connective tissue growth factor · Markov chain · Probability

Introduction

The human connective tissue growth factor has been the subject of numerous studies, among which mathematical approaches such as multiple sequence comparison and alignments have been used on many occasions for the analysis of amino acid sequences. However, to the best of the authors' knowledge, a detailed frequency and probability analysis regarding the amino acid sequence of human connective tissue growth factor has not yet been conducted. Thus, in this study we attempt to use frequency and probability to analyse the amino acid sequence of human connective tissue growth factor.

The human connective tissue growth factor is composed of 349 amino acids [1, 2]. Any two amino acids in order can construct a two-amino-acid sequence, so that a total of 348 two-amino-acid sequences can be constructed, i.e. the first and second, the second and third, etc. Furthermore, any three amino acids in order can also construct a three-amino-acid sequence, thus a total of 347 three-amino-acid sequences can be constructed, i.e. the first, second and third; the second, third and fourth; etc. Similar considerations can be deduced for sequences of more than three amino acids.

The use of such a determination of amino acid sequences is useful because (i) we know that a single amino acid "word" is not constructed of three amino acids as a "word" of DNA is constructed by three codons, but we do not know how many amino acids construct a single amino acid "word", and (ii) we do not know whether there are "punctuation" and "spaces" in an amino acid sequence of a protein, 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.

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 human connective tissue growth factor should be one of these 400 possible sequences and any two-amino-acid sequence that does not appear in human connective tissue growth factor should also be one of these 400 possible sequences. If each two-amino-acid sequence had the same probability to appear in the human connective tissue growth factor, each two-amino-acid sequence would be expected to appear about 0.870 times (348/400). 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 se-

G. Wu (✉)
Laboratoire de Toxicocinétique et Pharmacocinétique,
Faculté de Pharmacie,
Université de la Méditerranée Aix-Marseille II,
27 Boulevard Jean Moulin, F-13385 Marseille Cedex 05, France

S.-M. Yan
Department of Pathology, Medical School,University of Udine,
Udine, Italy

Present address: G. Wu
Novartis Pharma AG, WKL-135.1.16, 4002 Basel, Switzerland
e-mail: guang.wu@pharma.Novartis.com
Fax: +41 61 696 6992

quence in human connective tissue growth factor should be one of these 8000 possible sequences and any three-amino-acid sequence that does not appear in human connective tissue growth factor should also be one of these 8000 possible sequences. If each three-amino-acid sequence had the same probability to appear in the human connective tissue growth factor, each three-amino-acid sequence would be expected to appear about 0.043 times (347/8000). Similar deductions can be made for sequences of more than three amino acids.

Not surprisingly, some kinds of amino acid sequences do not appear at all in human connective tissue growth factor, not only because the human connective tissue growth factor does not have a long enough amino acid structure to hold all possible combinations, but also more importantly because the evolution process determines the preference of some particular amino acid sequences, some of which would appear more frequently.

There are 22 arginines (R) and 28 glycines (G) in human connective tissue growth factor. If a two-amino-acid sequence of "RG" were constructed by a purely random mechanism, an "RG" would be expected to occur with a frequency of 1.765 (22/349×28/348×348), i.e. "RG" would be expected to appear twice, but "RG" does not appear in the real situation, so the construction of "RG" cannot be explained by a purely random mechanism. By contrast, there are 22 alanines (A) and 28 prolines (P) in human connective tissue growth factor, the frequency of a random construction of "AP" would be expected to be 1.765 (22/349×28/348×348), i.e. "AP" would be expected to appear twice, which is true in the real situation, so the construction of "AP" can be explained by a purely random mechanism.

It is easy to say that no amino acid sequences of a protein are constructed by a purely random mechanism or no amino acid sequences of a protein can be explained by a purely random mechanism. However, it is not easy to answer, for example, what percentage of two-amino-acid sequences in a protein can be explained by a purely random mechanism and what percentage cannot. Thus the first question we would like to answer in this study is what percentage of amino acid sequences in human connective tissue growth factor can be explained by a purely random mechanism and what percentage cannot by the comparing predicted probability and frequency with the measured probability and frequency. Following this, the second question is what percentage of non-appearing amino acid sequences in human connective tissue growth factor can be explained by a purely random mechanism by comparing the predicted probability and the frequency with measured probability and frequency and what percentage cannot.

In an amino acid sequence, the issue of which amino acid is more likely to follow a preceding amino acid is also interesting. In an ideally random situation, each amino acid could be possible, thus the probability of following a preceding amino acid is 1/20. In the human connective tissue growth factor, there are 22 alanines (A). Thus an "A" would have a probability of 0.060 (21/348) of following a preceding amino acid "A", but the "A" has a probability of 0.095 of following a preceding "A" in the real situation. This real probability cannot be explained by a purely random mechanism, which is the concern of the Markov chain (the first order Markov chain transition probability). Thus the third question we address in this study is what percentage of the Markov transition probability in human connective tissue growth factor can be explained by a purely random mechanism and what percentage cannot. This question can be answered by comparing the predicted conditional probability with the measured Markov transition probability.

Materials and methods

The amino acid sequence of the human connective tissue growth factor was obtained from the Swiss-Protein database, access number P29279 [3].

Measuring two-, three- and four-amino-acid sequences

The measurement of two-, three- and four-amino-acid sequences in human connective tissue growth factor was conducted as stated in the Introduction. For two-amino-acid sequences, the first and second amino acids, the second and third, the third and fourth, until the 348th and 349th were recorded, and their frequencies and probabilities were calculated. For three-amino-acid sequences, the first, second and third amino acids, the second, third and fourth, until the 347th, 348th and 349th were recorded and their frequencies and probabilities were calculated, and similar for sequences of more than three amino acids. No measurement on sequences of more than four amino acids was conducted, because no repetition regarding sequences of more than four amino acids was found, thus each sequence of more than four amino acids is unique.

Calculating possible two-, three- and four-amino-acid sequences

All 20 kinds of amino acids exist in human connective tissue growth factor and the number of each kind of amino acid is at least one. Thus there are 400 (20^2) possible two-amino-acid sequences, but 7,600 (20^2×19) and 152,000 (20^3×19) for three- and four-amino-acid sequences because there are only two tryptophans (W) in human connective tissue growth factor.

Calculating predicted probability and frequency

The predicted probability was calculated according to the random mechanism as stated in the Introduction. For example, there are 22 alanines and 22 arginines in human connective tissue growth factor; for two-amino-acid sequences at any position, the predicted probabilities for "AA", "AR", "RR" and "RA" are 22/349× 21/348, 22/349×22/348, 22/349×21/348 and 22/349×22/348. For three-amino-acid sequences, the predicted probability for "AAA" is 22/349×21/348×20/347. The numbers of predicted probabilities are identical to the numbers of possible two-, three- and four-amino-acid sequences, e.g. 400 (20^2) for two-amino-acid sequences.

The predicted frequency is the rounded integral value of the production of predicted probability and total number of amino acid sequences. Thus the predicted frequency for "AA" is one (22/349×21/348×348). Naturally, the predicted frequency is less accurate than the predicted probability; however, the predicted frequency is easier to use for the sequences of more than two amino acids, because the predicted probability is extremely low.

Calculating predicted conditional probability

The predicted conditional probability for an amino acid to follow a preceding amino acid is calculated according to the random mechanism as stated in the Introduction. For example, there are 22 alanines and 22 arginines in human connective tissue growth factor. The predicted conditional probabilities for "AA" and "RA" are 21/348 and 22/348 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 22/348 and 21/348 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 20/347. The numbers of predicted conditional probabilities is identical to the numbers of possible two-, three- and four-amino-acid sequences, e.g. 400 (20^2) for two-amino-acid sequences.

Calculating Markov transition probability

The Markov chain is used to calculate the transition probability from one state to another state [4, 5, 6, 7]. For a two-amino-acid sequence, how large is the probability for an amino acid to follow a certain preceding amino acid? This constructs a conditional probability (the first order Markov chain), i.e. the probability that an amino acid occurs in a two-amino-acid sequence given a certain kind of 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 that an amino acid occurs in a three-amino-acid sequence given certain kinds of the first two amino acid [*P*(third amino acid|first and second amino acids)].

Statistical comparison

The measured frequency and predicted frequency are compared using the rounded integral value, and the measured probability/Markov transition probability and predicted probability/conditional probability are compared using three decimal rounded values. The measured frequencies of two- and three-amino-acid sequences are 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 7,600 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 four-amino-acid sequences are used in the above tests, because of the limitation in software and personal computer, e.g. 152,000 measured frequencies against 152,000 random data on a worksheet for four-amino-acid sequences. $P<0.05$ 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.

Results

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

In the human connective tissue growth factor, 205 of 400 (51.250%) possible two-amino-acid sequences do not exist, followed by 98 (24.500%) sequences that appear once, 60 (15.000%) sequences twice, 23 (5.750%) sequences three times, 10 (2.500%) sequences four times, three (0.750%) sequences five times and one (0.250%) sequence six times. These sequences are not randomly distributed.

Of 348 two-amino-acid sequences in human connective tissue growth factor, 81 (23.276%) and 21 (3.034%) sequences can be explained by the predicted frequency and probability according to a purely random mechanism.

Of 205 non-appearing two-amino-acid sequences in human connective tissue growth factor, 113 (55.122%) and 50 (24.390%) sequences can be explained by the predicted frequency and probability according to a purely random mechanism.

The two-amino-acid sequences that do not match the predicted frequency are particularly interesting, especially when the difference between measured and predicted frequencies is equal to or greater 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. If, for example, the predicted frequency of "RG" is two, whereas the measured frequency of "RG" is zero, this difference should have some underlying non-random reason. Table 1 shows these two-amino-acid sequences; for example, "RL" has the measured probability of 0.011 (4/348).

Of 392 measured first order Markov transition probabilities for the second amino acid in two-amino-acid se-

Table 1 The measured frequency (MF), the predicted frequency (PF) and the first order Markov chain transition probability (MP) of two-amino-acid sequences that have a difference equal to or greater than two between measured and predicted frequencies

Sequence[a]	MF	PF	MP	Sequence[a]	MF	PF	MP	Sequence[a]	MF	PF	MP
RG	0	2	0.000	RL	4	1	0.182	RV	4	1	0.182
NC	3	1	0.333	DG	5	2	0.250	CK	1	3	0.026
CP	5	3	0.128	CS	4	2	0.103	CT	6	2	0.154
CV	0	2	0.000	EC	0	2	0.000	EW	2	0	0.118
QS	2	0	0.286	QT	2	0	0.286	GE	3	1	0.107
GP	4	2	0.143	IK	3	1	0.300	IF	2	0	0.200
LC	4	2	0.190	LE	3	1	0.143	KA	0	2	0.000
KR	0	2	0.000	KQ	2	0	0.083	KV	0	2	0.000
FG	3	1	0.250	PA	5	2	0.179	PD	4	2	0.143
PH	2	0	0.071	PP	0	2	0.000	SM	3	1	0.150
TK	0	2	0.000	TT	3	1	0.136	YR	4	1	0.500
VR	3	1	0.136	VC	4	2	0.182	VP	0	2	0.000

[a] [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]

Table 2 The measured frequency (MF), the predicted frequency (PF) and the second order Markov chain transition probability (MP) of three-amino-acid sequences that have a difference equal to or greater than two between measured and predicted frequencies

Sequence	MF	PF	MP	Sequence	MF	PF	MP	Sequence	MF	PF	MP
AVG	2	0	1.000	RCP	2	0	0.667	RLE	2	0	0.500
DEP	2	0	1.000	DGA	2	0	0.400	CPD	2	0	0.400
GVC	2	0	0.667	LDG	2	0	1.000	LCS	2	0	0.500
KTC	2	0	0.667	VCT	2	0	0.500				

quences, no measured Markov transition probability matches the predicted conditional probability. The measured first order Markov chain transition probabilities for which the difference between measured and predicted frequencies is equal to or greater than two are also presented in Table 1, e.g., if the first amino acid in a two-amino-acid sequence is "R", then the probability that the second amino acid is "L" is 0.182 (Table 1).

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

In the human connective tissue growth factor, 7,264 of 7,600 (95.579%) possible three-amino-acid sequences do not exist, followed by 325 (4.276%) sequences that appear once and 11 (0.145%) sequences twice. These sequences are not randomly distributed.

The maximal predicted probability and frequency of three-amino-acid sequence are 0.001 (39/349×38/348×37/347, for example, "CCC") and 0 (39/349×38/348×37/347×347); thus no sequences of more than two amino acids in human connective tissue growth factor can be explained by a purely random mechanism.

The three-amino-acid sequences that have a difference equal to or greater than two between predicted and measured frequencies are presented in Table 2. For example, "AVG" has the measured probability of 0.006 (2/347).

No predicted conditional probability matches the measured second order Markov transition probability, e.g. no other amino acids but the "G" can be the third amino acid in the three-amino-acid sequence when the first two amino acids are "AV".

Four-amino-acid sequences and the third order Markov chain transition probability

In the human connective tissue growth factor, 151,655 of 152,000 (99.773%) possible four-amino-acid sequences do not exist, followed by 344 (0.226%) sequences that appear once and one (0.001%) sequence (GVCT) that appears twice with the third order Markov transition probability of 1.000. Also no measured Markov transition probability matches the predicted conditional probability.

No repetition was found regarding any sequence of more than four amino acids; thus the human connective tissue growth factor does not favor repetition at this level.

Given the background of uneven distribution of amino acids in all known proteins, a conclusion can be drawn from our results that a considerable number of two-amino-acid sequences can be constructed by a random mechanism from the uneven distribution of amino acids. However, no conclusion can be reached about longer sequences simply due to the small sample size.

Discussion

In this study we attempted to answer three questions; the methods used in this study are somewhat similar to the methods used in cryptology [7, 8, 9]. A perfect code developed by a cryptography can resist any statistical test for random distribution, whereas the human connective tissue growth factor cannot do so and offers the possibility to be deciphered. Hence, our analysis may serve as the first step for the understanding of the wording of human connective tissue growth factor.

By comparing the predicted frequency/probability with the measured frequency/probability, one can determine which amino acid "word" is favored by a protein. Also the most frequent amino acid sequences may serve as potential targets of new drugs, because the drug would have a greater chance of interacting with them. In fact, this study also suggests that the amino acid sequences that have the greatest difference between predicted and measured frequencies/probabilities and appear in the human connective tissue growth factor should serve as potential targets for new drugs, because these sequences are highly evolved for the difference between predicted and measured frequencies/probabilities. Moreover, as the Markov transition probability increases from lower order to higher order, the random chance for an amino acid following an arbitrary amino acid decreases as the length of amino acid sequence increases. Hence, a mutation is unlikely to occur at the amino acid with a high Markov transition probability, and the amino acids with a high Markov transition probability may serve as potential targets of new drugs, because they are unlikely to change into other amino acids.

It is interesting that 23.276% and 6.034% of two-amino-acid sequences in human connective tissue growth factor can be explained by the predicted frequency and probability according to a purely random mechanism. This raises an interesting issue of whether a mutation in

a randomly constructed amino acid sequence has no effect on a protein function.

If a possible amino acid sequence does not appear in a protein, it is physically obvious that it is useless for the protein function. For example, 205 two-amino-acid sequences do not exist in human connective tissue growth factor, of which some can be explained by a purely random mechanism and some cannot. For example, the predicted frequency of "RG" is two, whereas the measured frequency is zero, which cannot be explained by a purely random mechanism. By contrast, both predicted and measured frequencies of "AW" are zero, thus the lack of "AW" in human connective tissue growth factor can be explained by a purely random mechanism.

For sequences of more than two amino acids, no sequence in human connective tissue growth factor can be explained by a purely random mechanism, but all the sequences which do not appear in human connective tissue growth factor can be explained by a purely random mechanism. This also leaves an interesting question of whether the fact that most sequences of more than two amino acids are not selected for the construction of human connective tissue growth factor is due to a purely random mechanism.

Although we have used numerous mathematical methods in our previous studies and this study, we hope to use more sophisticated methods to analyse the human connective tissue growth factor in future to gain more insight into this issue.

Acknowledgements The Electronic Engineer P. Cossettini at the Center for Advanced Research in Space Optics, Trieste, Italy is kindly acknowledged. Special thanks go to Professor Paul Garrett at the University of Minnesota, USA for providing his lecture notes on cryptology. Special thanks go to the referees for their insightful comments, which sharpened up the current version of manuscript.

References

1. Bradham, D. M.; Igarashi, A.; Potter, R. L.; Grotendorst, G. R. *J. Cell Biol.* **1991**, *114*, 1285–1294.
2. Igarashi, A.; Bradham, D. M.; Okochi, H.; Grotendorst, G. R. *J. Dermatol.* **1992**, *19*, 642–643.
3. Bairoch, A.; Apweiler R. *Nucleic Acids Res.* **1999**, *27*, 49–54.
4. Ash, R. B. *Information theory*; Interscience: New York, 1965.
5. Csiszár, I.; Körner, J. *Information theory*; Academic Press: New York, 1981.
6. Feller, W. *An introduction to probability theory and its applications*; 3rd edn, Vol I, John Wiley and Sons: New York, 1968.
7. van der Lubbe, J. C. A. *Information theory*; Cambridge University Press: Cambridge, 1997.
8. Sinkov, A. *Elementary cryptanalysis – A mathematical approach*; Mathematical Association of American, Yale University Press: New Haven, 1966.
9. van Tilborg, H. C. A. *An introduction to cryptology*; Kluwer: Boston, 1989.