

# Prediction of Two- and Three-Amino-Acid Sequences of *Citrobacter Freundii* $\beta$ -Lactamase from its Amino Acid Composition

Guang Wu<sup>1\*</sup>, Shao-Min Yan<sup>2</sup>

<sup>1</sup>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

<sup>2</sup>Department of Pathology, Medical School, University of Udine, Piazzale Santa Maria della Misericordia, I-33100 Udine, Italy

## Abstract

The repeated amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase may be indispensable for its function, because such repetitions cannot be simply attributed to a chance. In order to fully explore the functional units in *Citrobacter Freundii*  $\beta$ -lactamase, it may need to analyse all the amino acid pairs, triplets, etc. along *Citrobacter Freundii*  $\beta$ -lactamase from one terminal to the other terminal, to count their frequencies and calculate their probabilities. The amino-acid sequence of *Citrobacter Freundii*  $\beta$ -lactamase was counted according to two-, three- and four-amino-acid sequences. The counted frequency and probability were compared with the predicted frequency and probability. The amino acid sequences, which appear in *Citrobacter Freundii*  $\beta$ -lactamase and can be predicted from its amino acid composition according to a purely random mechanism, should not be deliberately evolved and conserved. By contrast, the amino acid sequences, which appear in *Citrobacter Freundii*  $\beta$ -lactamase but cannot be predicted from its amino acid composition according to a purely random mechanism, should be deliberately evolved and conserved. Accordingly 99 (26.053%) and 33 (8.684%) of 380 two-amino-acid sequences can be predicted by the frequency and probability according to a purely random mechanism. Some kinds of amino acid sequences, which absent in *Citrobacter Freundii*  $\beta$ -lactamase and can be predicted from its amino acid composition according to a purely random mechanism, should not be deliberately excluded from *Citrobacter Freundii*  $\beta$ -lactamase. By contrast, some kinds of amino acid sequences, which absent in *Citrobacter Freundii*  $\beta$ -lactamase and cannot be predicted from its amino acid composition according to a purely random mechanism, should be deliberately excluded from *Citrobacter Freundii*  $\beta$ -lactamase. Accordingly 89 (48.370%) and 41 (22.283%) of 184 kinds

of absent two-amino-acid sequences can be predicted by the frequency and probability according to a purely random mechanism, and 7236 (99.848%) of 7247 kinds of absent three-amino-acid sequences can be predicted by the frequency according to a purely random mechanism. The amino acids, whose probabilities in following certain preceding amino acids can be predicted from *Citrobacter Freundii*  $\beta$ -lactamase amino acid composition according to a purely random mechanism, should not be deliberately evolved and conserved, accordingly 2 (0.526%) of 380 counted first order Markov transition probabilities for the second amino acid in two-amino-acid sequences match the predicted conditional probabilities

## Introduction

The relationship between  $\beta$ -lactamases and their functions has been the objective of numerous experimental and theoretical studies, with respect to the theoretical approach, so far we know many functional units in  $\beta$ -lactamases by multiple sequence comparisons and alignments. However, much effort is still needed to explore the functional units in  $\beta$ -lactamases.

In order to fully explore the functional units in *Citrobacter Freundii*  $\beta$ -lactamase, we may need to analyse all the amino acid pairs, triplets, etc. along *Citrobacter Freundii*  $\beta$ -lactamase from one terminal to the other terminal, to count their frequencies and calculate their probabilities, because we still do not fully understand that (1) each of various functional units needs how many amino acids, (2) from where a functional unit begins and stops, and (3) how non-functional units are mixed with functional units. From evolutionary and probabilistic viewpoints, the functional units should be deliberately evolved and conserved, thus the occurrences of amino acids in them should not be explained by a purely random mechanism; by contrast the non-functional units should not be deliberately evolved and conserved, thus the occurrences of amino acids in them could possibly be explained by a purely random mechanism.

The  $\beta$ -lactamase (EC 3.5.2.6, cephalosporinase) in *Citrobacter Freundii* is composed of 381 amino acids (Lindberg and Normark, 1986; Oefner *et al.*, 1990; Tsukamoto *et al.*, 1990; Yamaguchi *et al.*, 1987), we may count the first and second amino acids as a two-amino-acid sequence, the second and third as another two-amino-acid sequence, the third and fourth, until the 380th and 381st, thus there is a total of 380 two-amino-acid sequences. Furthermore, we may count the first, second and third amino acids as a three-amino-acid sequence, the second, third and fourth as another three-amino-acid sequence, until the 379th, 380th and 381st, thus there is a total of 379 three-amino-acid sequences. Similar

Received February 16, 2000; accepted February 24, 2000. \*For correspondence. Email Guang.Wu@pharmacie.univ-mrs.fr; Tel. +33 4 91 83 56 45; Fax. +33 4 91 80 26 12.

consideration can be deduced for more-than-three-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, thus there are 400 ( $20^2$ ) kinds of possible two-amino-acid sequences (combinations). There are 380 two-amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase, if there was no repetition in them, there would be 380 kinds of two-amino-acid sequences, therefore one of 400 kinds of possible two-amino-acid sequences would be expected to appear about 0.950 times (380/400). Clearly some of 400 kinds of possible two-amino-acid sequences cannot appear in *Citrobacter Freundii*  $\beta$ -lactamase, a kind of two-amino-acid sequence which either appears or absents in 380 two-amino-acid sequences of *Citrobacter Freundii*  $\beta$ -lactamase should be one of 400 kinds of possible two-amino-acid sequences. Similarly, three amino acids in a three-amino-acid sequence can be constructed from any one of 20 amino acids, there are 8000 ( $20^3$ ) kinds of possible three-amino-acid sequences. There are 379 three-amino-acids sequences in *Citrobacter Freundii*  $\beta$ -lactamase, if there was no repetition in them, there would be 379 kinds of three-amino-acid sequences, therefore one of 8000 kinds of possible three-amino-acid sequences would be expected to appear about 0.047 times (379/8000). Naturally, some of 8000 kinds of possible three-amino-acid sequences cannot appear in *Citrobacter Freundii*  $\beta$ -lactamase, a kind of three-amino-acid sequence which either appears or absents in 379 three-amino-acid sequences of *Citrobacter Freundii*  $\beta$ -lactamase should be one of 8000 kinds of possible sequences. Similar consideration can be applied to more-than-three-amino-acid sequences.

The reason that some kinds of amino-acid sequences absent in *Citrobacter Freundii*  $\beta$ -lactamase is not only because *Citrobacter Freundii*  $\beta$ -lactamase does not have such a long amino acid structure to hold all possible combinations, but also more importantly because the evolutionary process determines the preference of some particular amino-acid sequences, some of which would appear more frequently. Some kinds of absent amino-acid sequences may deliberately be eliminated from evolutionary process for the construction of *Citrobacter Freundii*  $\beta$ -lactamase, thus the lack of them should not be predicted by a purely random mechanism. But some kinds of absent amino-acid sequences may not deliberately be eliminated, thus the lack of them could be predicted by a purely random mechanism.

For example, there are 41 alanines (A) in *Citrobacter Freundii*  $\beta$ -lactamase. If a two-amino-acid sequence of 'AA' was constructed by a purely random mechanism, the 'AA' would be expected to occur by the frequency of 4.304 ( $41/381 \times 40/380 \times 380$ ), i.e. the 'AA' would be expected to appear four times, which is true in the real situation, so the construction of 'AA' is predictable by a purely random mechanism. By contrast, there are 13 arginines (R) in *Citrobacter Freundii*  $\beta$ -lactamase, the frequency of random construction of 'AR' is 1.399 ( $41/381 \times 13/380 \times 380$ ), i.e. the 'AR' would be expected to appear once, but the 'AR' appear three times in the real situation, so the construction of 'AR' does not follow a purely random mechanism, but follows a functional and evolutionary propose.

Thus the first question this study answers is what a percentage of amino-acid sequences can be predicted by

a purely random mechanism and what not in *Citrobacter Freundii*  $\beta$ -lactamase by comparing predicted probability and frequency with the counted probability and frequency. Following this, the second question this study answers is what a percentage of absent kinds of amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase is predictable by a purely random mechanism and what not by comparing predicted probability and frequency with the counted probability and frequency.

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 in following a preceding amino acid is 1/20. There are 9 phenylalanines (F) in *Citrobacter Freundii*  $\beta$ -lactamase, an 'F' would have the probability of 0.024 (9/380) in following a preceding 'A', which is true in the real situation, therefore an 'F' follows a purely random mechanism in following a preceding 'A'. By contrast, a 'R' would have the probability of 0.034 (13/380) in following a preceding 'A' according to a purely random mechanism, but a 'R' has the probability of 0.073 in following a preceding 'A' in the real situation, this real probability is what the Markov chain is interested (the first order Markov chain transition probability). Thus the third question this study answers is what a percentage of the Markov transition probability can be predicted by a purely random mechanism and what not in *Citrobacter Freundii*  $\beta$ -lactamase by comparing predicted conditional probability with the counted Markov transition probability.

## Results

### Two-Amino-Acid Sequences and Their First Order Markov Chain Transition Probabilities

In *Citrobacter Freundii*  $\beta$ -lactamase, 184 of 400 (46.000%) possible kinds of two-amino-acid sequences do not exist, followed by that 111 (27.750%) kinds appear once, 65 (16.250%) kinds twice, 23 (5.750%) kinds three times, 16 (4.000%) kinds four times and 1 (0.250%) kind six times.

Of 380 two-amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase, 99 (26.053%) and 33 (8.684%) sequences are predictable by the frequency and probability according to a purely random mechanism.

Of 184 kinds of absent two-amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase, 89 (48.370%) and 41 (22.283 %) kinds are predictable by the frequency and probability according to a purely random mechanism.

The two-amino-acid sequences which do not match the predicted frequencies are particularly important, because these mismatches should be deliberately obtained, especially when the difference between counted and predicted frequencies is equal to 2 or larger than 2, because the predicted frequency is the rounded value of predicted probability and the difference being equal to 1 may be due to the rounded error. For example, the predicted frequency of 'AN' is 2, whereas the counted frequency of 'AN' is 4, this difference should not be due to the chance. Table 1 shows these two-amino-acid sequences, for example, the 'AR' has the counted probability of 0.008 (3/380) and the 'LA' has the highest counted probability of 0.016 (6/380).

Of 380 counted first order Markov transition probabilities for the second amino acid in two-amino-acid sequences, 2 (0.53%) counted first order Markov transition



## Discussion

In this study we answered three questions to explore the functional units in *Citrobacter Freundii*  $\beta$ -lactamase, the methods used in this study are somewhat similar to the methods used our other studies (Wu 1999, 2000a, b, c).

By comparing predicted frequency/probability with the counted frequency/probability, one can know (i) which kind of amino acid sequence appears more in *Citrobacter Freundii*  $\beta$ -lactamase; (ii) which kind of amino acid sequence deliberately appears or absents in *Citrobacter Freundii*  $\beta$ -lactamase; and (iii) which amino acid deliberately follows a certain amino acid.

Apparently, no statistical significance is given in this study, in fact, the results are too highly statistically significant to mention. In general case, an amino acid has the 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 case of *Citrobacter Freundii*  $\beta$ -lactamase, there are 41 'A', which is the most abundant amino acid, and 2 'C', which is the least amino acid. If the first amino acid is 'A', then the chance of the second amino acid to be 'A' is 40/380 ( $P = 0.105$ ), if the first amino acid is 'C', then the chance of the second amino acid to be 'C' is 1/380 ( $P = 0.0026$ ). However we mainly deal with the repetition of two- and three-amino-acid sequence in this study, the chance of appearance of two-amino-acid sequence of 'AA' is  $41/381 \times 40/380$  ( $P = 0.011$ ), the chance of the first repetition of 'AA' is  $39/379 \times 38/378$  ( $P = 0.010$ ), the chance of appearance of three-amino-acid sequence of 'AAA' is  $41/381 \times (40/380 \times 39/379)$  ( $P = 0.001$ ) and the chance of the first repetition of 'AAA' is  $38/378 \times 37/377 \times 36/376$  ( $P = 0.0009$ ). If we consider the two- and three-amino-acid sequences constructed by other less abundant amino acids in *Citrobacter Freundii*  $\beta$ -lactamase, the chance of appearance and repetition would be further smaller, therefore all the probabilities of two- and three-amino-acid sequences are less than the conventional  $P < 0.05$ . In such a case we have no need to increase the sample size to detect the statistical difference, also we use the factor of '2' (two repetitions) in Tables for the statistical difference which is much more statistically significant than  $P < 0.05$ .

Our study is different from the current multiple sequence comparisons and alignments, which require as many as possible proteins for comparison. However, each amino acid sequence is separated into many parts in multiple sequence comparisons and alignments, which cannot construct two-, three-, four-amino-acid sequences as what we have done. It could be possible to use our method to analyse all  $\beta$ -lactamases from many organisms or many proteins from a given organism or many proteins from different organisms, however this aim can be achieved only after the acceptance of this method by the scientific community and the complete computerisation of this method. On the other hand, each protein should have its own particular function, otherwise it can be replaced by other proteins, thus the detailed analysis of one protein is still needed.

Although the protein function is related to its 3-dimensional structure, the primary structure is the basis for the 3-dimensional structure, one could not fully analyse the 3-dimensional structure without the detailed knowledge

of a protein primary structure. Also it is impracticable to analyse the several kinds of structures of a protein in a single study.

It is interesting that some of amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase and some absent kinds of amino-acid sequences from *Citrobacter Freundii*  $\beta$ -lactamase are predictable according to a purely random mechanism. This means that we can divide *Citrobacter Freundii*  $\beta$ -lactamase into 'random' and 'non-random' regions. 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 in this study raise an interesting issue of whether or not these unharmed mutations occur in the 'random' region.

For more-than-two-amino-acid sequences, no occurrence of any kind of sequences in *Citrobacter Freundii*  $\beta$ -lactamase can be predicted by a purely random mechanism, because of the low probability to occur. But on the other hand, all non-occurrence of any kind of the sequences in *Citrobacter Freundii*  $\beta$ -lactamase can be predicted by a purely random mechanism also because of the low probability to occur. This leaves an interesting question of whether the reason that most more-than-two-amino-acid sequences are not selected for the construction of *Citrobacter Freundii*  $\beta$ -lactamase is due to a purely random mechanism.

The amino-acid sequences are functionally and evolutionally biased, it would be interesting to know why the *Citrobacter Freundii*  $\beta$ -lactamase favours these repeated three-amino-acid sequences. In general, the repeated three-amino-acid sequences are located outside of *Citrobacter Freundii*  $\beta$ -lactamase signal region (1 to 20), thus they might have no signal function. Also these repeated three-amino-acid sequences have no location in *Citrobacter Freundii*  $\beta$ -lactamase binding region (335 to 337), so they are unlikely to have the binding function. However, all these repeated three-amino acids sequences are located inside *Citrobacter Freundii*  $\beta$ -lactamase chain (21 to 381).

Although we have used numerous mathematical methods in our previous studies, we hope that this study can provide some insight into *Citrobacter Freundii*  $\beta$ -lactamase.

## Experimental Procedures

The amino acid sequence of the *Citrobacter Freundii*  $\beta$ -lactamase was obtained from the Swiss-Protein, access number P05193 (Bairoch and Apweiler, 1999).

### Counting Two-, Three- and Four-Amino-Acid Sequences

The two-, three- and four-amino-acid sequences in *Citrobacter Freundii*  $\beta$ -lactamase were counted 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 380th and 381st were counted, 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 379th, 380th and 381st were counted and their frequencies and probabilities were calculated. No more-than-four-amino-acid sequences were counted, because no repetition regarding more-than-four-amino-acid sequences was found, thus each more-than-four-amino-acid sequence is unique.

### Calculating Possible Two-, Three- and Four-Amino-Acid Sequences

Because all 20 kinds of amino acids exist in *Citrobacter Freundii*  $\beta$ -lactamase and the number of each kind of amino acid is at least larger than 2, thus there are 400 ( $20^2$ ) possible kinds of two-amino-acid sequences, but there are 7600 ( $20^2 \times 19$ ) and 152000 ( $20^3 \times 19$ ) possible kinds of three- and four-amino-acid sequences because of presence of only 2 cysteines (C) in *Citrobacter Freundii*  $\beta$ -lactamase.

### Calculating Predicted Probability and Frequency

The predicted probability was calculated according to the random mechanism as stated in the introduction. For example, there 41 alanines and 13 arginines in *Citrobacter Freundii*  $\beta$ -lactamase, for two-amino-acid sequences at any position, the predicted probabilities for 'AA', 'AR', 'RR' and 'RA' are  $41/381 \times 40/380$ ,  $41/381 \times 13/380$ ,  $13/381 \times 12/380$  and  $13/381 \times 41/380$ . For three-amino-acid sequences, the predicted probability for 'AAA' is  $41/381 \times 40/380 \times 39/379$ . The numbers of predicted probabilities are identical to the numbers of possible kinds of 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 frequencies for 'AA' is 4 ( $41/381 \times 40/380 \times 380$ ). Naturally the predicted frequency is less accurate than the predicted probability, however, the predicted frequency is more easy to use for the more-than-two-amino-acid sequences, because the predicted probability is extremely low.

### Calculating Predicted Conditional Probability

The predicted conditional probability for an amino acid in following a preceding amino acid is calculated according to the random mechanism as stated in the introduction. For example, there 41 alanines and 13 arginines in *Citrobacter Freundii*  $\beta$ -lactamase, the predicted conditional probabilities for 'AA' and 'RA' are  $40/380$  and  $41/380$  for the second amino acid of 'A' in two-amino-acid sequences in following an 'A' and a 'R', the predicted conditional probabilities for 'AR' and 'RR' are  $13/380$  and  $12/380$  for the second amino acid of 'R' in following an 'A' and a 'R'. The predicted conditional probability of the third amino acid of 'A' in a three-amino-acid sequence in following 'AA' is  $39/379$ . The numbers of predicted conditional probabilities are identical to the numbers of possible kinds of 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 to calculate the transition probability from one state to another state. (Ash, 1965; Csiszár and Körner, 1981; Feller, 1968; van der Lubbe, 1997). For a two-amino-acid sequence, an amino acid has a certain 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 first amino acid [ $P(\text{second amino acid}|\text{first amino acid})$ ]. The calculation of this probability is the transition from the state of one-amino-acid sequence (if it can be called as a sequence) to the state of two-amino-acid sequence, and the state of two-amino-acid sequence is only dependent on the state of one-amino-acid sequence. For a three-amino-acid sequence, the second order Markov chain can be defined, i.e. the probability of an amino acid occurs in a three-amino-acid sequence given a certain first two amino acid [ $P(\text{third amino acid}|\text{first and second amino acids})$ ]. The calculation of this probability is the transition from the state of two-amino-acid sequence to the state of three-amino-acid sequence, and the state of three-amino-acid sequence is only dependent on the state of two-amino-acid sequence.

### Statistical Comparison

The counted frequency and predicted frequency are compared using the integral rounded value, and the counted probability/Markov transition probability and predicted probability/conditional probability are compared using three decimal rounded value (the detailed statistical significance is addressed in the discussion).

### Acknowledgements

The Electronic Engineer P. Cossetini at the Center for Advanced Research in Space Optics, Trieste, Italy is kindly acknowledged. Special thanks go to anonymous Referees for their valuable comments and Ms Milda Simonaitis for her editorial assistance.

### References

- 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.  
 Csiszár, I., and Körner J. 1981. Information theory. Academic Press, New York.  
 Feller, W. 1968. An introduction to probability theory and its applications. 3rd edn, John Wiley and Sons, Vol I, New York.  
 Lindberg, F., and Normark, S. 1986. Sequence of the *Citrobacter freundii* OS60 chromosomal ampC beta-lactamase gene. *Eur. J. Biochem.* 156: 441-445.

- Oefner, C., D'arcy, A.A., Daly, J.J., Gubernator, K., Charnas, R.L., Heinze, I., Hubschwerlen, C., and Winkler, F.K. 1990. Refined crystal structure of beta-lactamase from *Citrobacter freundii* indicates a mechanism for beta-lactam hydrolysis. *Nature* 343: 284-288.  
 Tsukamoto, K., Tachibana, K., Yamazaki, N., Ishii, Y., Ujii, K., Nishida, N., and Sawai T. 1990. Role of lysine-67 in the active site of class C beta-lactamase from *Citrobacter freundii* GN346. *Eur. J. Biochem.* 188: 15-22.  
 van der Lubbe, J.C.A. 1997. Information theory. Cambridge University Press, Cambridge.  
 Wu, G. 1999. The first and second order Markov chain analysis on amino acids sequence of human haemoglobin  $\alpha$ -chain and its three variants with low  $O_2$  affinity. *Comp. Haematol. Int.* 9: 148-151  
 Wu, G. 2000a. 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 Related Topics)* (in press)  
 Wu, G. 2000b. The first, second, third and fourth order Markov chain analysis on amino acids sequence of human dopamine  $\alpha$ -hydroxylase. *Mol. Psychiatry* (in press)  
 Wu, G. 2000c. Frequency and Markov chain analysis of the amino-acid sequence of human alcohol dehydrogenase  $\alpha$ -chain. *Alcohol Alcohol.* (in press)  
 Yamaguchi, A., Adachi, H., and Sawai, T. 1987. Identification of the active site of *Citrobacter freundii* beta-lactamase using dansyl-penicillin. *FEBS. Lett.* 218: 126-130.

