African Journal of Microbiology Research Vol. 5(12), pp. 1479-1483, 18 June, 2011
Available online http://www.academicjournals.org/ajmr
DOI: 10.5897/AJMR10.876
ISSN 1996-0808 ©2011 Academic Journals

Full Length Research Paper

Prediction of bacterial toxins by an improved feature extraction and IB1 algorithm fusion

Chaohong Song

College of Science, Huazhong Agricultural University, Wuhan 430070, China. E-mail: chh_song@mail.hzau.edu.cn.
Tel: +86 189 71212816. Fax: +86 027 87282133.

Accepted 10 March, 2011

Correctly identifying bacterial toxin is of great benefit to cell biology and medical research. In order to improve predictive accuracy, based on the concept of pseudo amino acid composition, combined with the methods of approximate entropy and IB1 algorithm, a new method is proposed to predict bacterial toxins in this paper. The improved method gives comprehensive consideration of amino acid composition, side-chain mass of the amino acid, hydrophilic, and hydrophobic characteristics of a protein sequence. The total prediction accuracy of our method was 97.52% for bacterial toxin and non-toxin, and 97.33% for discriminating endotoxins from exotoxins, which were much higher than that of the previous methods.

Key words: Approximate entropy, IB1 algorithm, bacterial toxin, exotoxins, endotoxins, pseudo amino acid composition.

INTRODUCTION

The whole world can be considered to be consisted of various systems such as economic system, biological system, etc., (Backlund, 2000) and all the systems in the existent physical reality, from the cosmologic phenomena to immunology and comportments of the subatomic particles, seem to be characterized with the presence of various *patterns* (Steen, 1988) in their structure (Pullan and Bhadeshia, 2000) and behavior (Dusenbery, 2009). The existent physical reality is not so complicated, in that, mathematically it is possible, but contrarily, the diversity of the systems can be restricted to a very little number of mathematically possible variations, in biology too. Various research methods and mathematical instruments, such as, “thinking machines” (Turing, 1950) are developed for understanding this “reduced complexity”.

The recognition of patterns (Gibson, 2003) is a major problem today, when the text recognition and transformation from the scanner’s image that is directly in an editable text is commonly used. The other problems (that is, from the screening of the human biometrical characteristics in mass-accessed places for security considerations, to the recognition of biological molecules, such as proteins, DNA and RNA) from the untreatable great quantity of data to the instruments of measurement (like the chips for DNA recognition) are in dynamic development. For physicians today, they use new and efficient methods for patients, like the selection and determination of tumor markers directly on the mRNA. An otherwise untreatable great quantity of data from the instruments of measurement, equipped with molecule-recognition chips, are used for the direct examination of biopsy material from tumors and metastases (or tumor cell lines developed from this), via software based on mathematical algorithms (Ochs et al., 2009).

Thus, it is possible, for example, to have a quick, INDIVIDUALISED (tendency of the medicine in future) and repeated change in the cancer therapy drug combination schema, weekly if necessary, and anytime using the most efficient combination and preventing, in this way, the development of the drug resisting the tumor. Finally, the success of the therapy, or long-time survival of the patient, in comparison with the others used, empirically proposed combinations. The methods of recognition of patterns have important mathematical implications, like the Bayesian algorithms (Howson and Urbach, 1989), Support Vector Machines (Cortes and Vapnik, 1995), IB1 algorithm (Aha et al., 1991), and others, like Artificial Learning Systems (Akerkar and Sajja, 2009), Artificial

Intelligence and Neural Networks (Russell and Norvig, 2003). However, the development of this area influences positively the last mentioned areas, which are important for other applications too, outside the pattern recognition (for example: neurobiology and physics).

The bacterial toxins are a major cause of diseases during infection (B¨ohnel and Gessler, 2005), and can be classified into exotoxins and endotoxins. These two types of toxins have different role and mechanism in the body and correctly identifying bacterial toxin, is of great benefit to mankind. In fact, some of these powerful disease-causing toxins have been exploited to further basic knowledge of cell biology or for medical purposes. For example, cholera toxin and the related labile-toxin of *E. coli*, as well as *B. pertussis* toxin, have been used as biologic tools to understand the mechanism of adenylate cyclase activation (Harnett, 1994; Bokoch et al., 1983; Neer, 1995), and the strong mucosal adjuvants have been used in experimental models (Bagley et al., 2002). Though bacterial toxins can be identified by experimental methods, it is costly and time-consuming. So, how to economically, rapidly and accurately identify bacterial toxins becomes a very important problem.

Recently, some researches have been made in this field and achieved inspiring results, using support vector machines (SVM) and dipeptides composition. Saha and Raghava (2007) achieved an accuracy of 96.07 and 92.50% for bacterial toxins and non toxins, respectively, and an accuracy of 95.71 and 92.86% for discriminating endotoxins and exotoxins, respectively. Yang and Li (2009) achieved higher MCC in the same dataset by using increment of diversity and support vector machines. Encouraged by their research, in this study, we attempted to develop a new method to predict bacterial toxins and their class (exotoxin or endotoxin).

MATERIALS AND METHODS

The software used for working the data

MATLAB (Gilat, 2004) is a high-level language and interactive environment that enables computational tasks to be performed faster than that of traditional programming languages such as C, C++ and FORTRAN. It has been widely used in various application areas, such as computational biology and pattern recognition. All calculations done in this paper were realized by programming, under MATLAB 2007.

Dataset

The data that we used in this paper were collected from Swiss-Prot database (Boeckmann et al., 2003) and from the dataset used by Saha and Raghava (2007). We freely downloaded them from http://www.imtech.res.in/raghava/btxpred/supplementary.html.

Using the cd-hit soft (Li, 2006) to remove sequences with more than 90% sequence identity, and using it to delete the sequences whose length is ≤100, we obtained two datasets. One contained 141 bacterial toxins and 303 non-toxins, while the other contained 73 exotoxins and 77 endotoxins.

Schemes of sequence feature

The pseudo amino acid composition of a sequence includes a lot of information about the sequence, such as the main feature of amino acid composition, and the sequence order correlation (Chou, 2001). So, in this paper, we constructed the feature vectors of a protein sequence with the concept of Chou's pseudo amino acid composition.

Suppose a protein chain X with length l amino acid residues is: $R_1 R_2 \cdots R_l$, we denoted the protein sequence as a vector in $20+s+\lambda$ dimension space. That is:

$$X=(x_1,x_2\cdots x_{20},x_{21}\cdots x_{20+s},x_{20+s+1}\cdots x_{20+s+\lambda})$$

Here, X is a normalized vector of $u=(u_1,u_2\cdots u_{20},u_{21}\cdots u_{20+s+\lambda})$

And $u_t=\begin{cases} f_t & 1\le t\le 20 \\ \varpi_1 ApEn_{t-20} & 21\le t\le 20+s \\ \varpi_2\theta_{t-20-s} & 21+s\le t\le 20+s+\lambda \end{cases}$

Where, f_t is the frequency of the 20 amino acids in protein X, ϖ_1,ϖ_2 is the weight factor for sequence order effect, $ApEn_i$ is the approximate entropy of protein sequences (Pincus, 1991), which describes the complexity of protein sequences, and θ_j is the j-tier sequence correlation factor, which reflects the sequence order correlation among the most contiguous residues of the jth.

$ApEn_i$ could be computed by the following equations:

$$ApEn_i=ApEn(m,r)=\varphi^m(r)-\varphi^{m+1}(r)$$

$$\varphi^m(r)=\sum_{i=1}^{N-m+1}\ln C_i^m(r)/(N-m+1)$$

$C_i^m(r)=\sum_j \operatorname{sgn}(r-d(x(i),x(j)))$, $1\le i\le N-m+1$ and

$$d(x(i),\ \ x(j))=\max_{k=1,2\ \cdots m}|u\ \ (i+k-1)-u\ \ (\ j+k-1\ \ )|$$

Here, $x(i)=(u(i),u(i+1)\cdots u(i+m-1)),\quad i=1,2\cdots N-m+1$ are the protein subsequences that begin at component i within X. N is the component number of the given X, while r and m are the filter parameter and mode dimension, respectively. In computing, we select $m=2,3,4$ and $r=0.1,0.15,0.2,0.25$, and then we obtain 12 approximate entropies, that is, $s=12$ θ_j was computed by the following formulae:

$$\theta_j=\sum_{i=1}^{l-j}\varphi(R_i,R_{i+j})/(l-j)\quad (j<l,j=1,2\cdots\lambda\quad)$$

Here $\varphi(R_i,R_{i+j})$ is correlation function,

$$\varphi(R_i,R_{i\ +j})=\frac{1}{3}\sum_{k=1}^{3}(H_k(R_{\ i+j})-H_k(R_i))^{\ 2}$$

and

$$H_k(R_i) = (H_k^0(i) - \frac{1}{20}\sum_{i=1}^{20} H_k^0(i)) / \sqrt{\frac{1}{20}\sum_{i=1}^{20}(H_k^0(i) - \frac{1}{20}\sum_{i=1}^{20} H_k^0(i))^2}$$

Where $H_k(R_i), k = 1,2,3$ are the value of hydrophobicity, hydrophilicity and side-chain mass of the amino acid R_i, respectively, $H_1^0(i)$ and $H_2^0(i)$ are the corresponding original hydrophobicity and hydrophilicity values of the *i*th amino acid (Argos et al., 1982; Hopp-Woods, 1981), respectively, and $H_3^0(i)$ is the side chain mass of the *i*th amino acid that can be obtained easily from any biochemistry text book. Generally, we used number to represent the 20 native amino acids from 1 to 20, according to their alphabetical order.

IB1 algorithm

IB1 algorithm is a classification algorithm characterized by incremental, supervised learning (Aha, 1990). It achieves effective results usually by the steps such as normalization, similarity and prediction. For some given numeric protein sequences, we first normalize them by the following formulae:

$$nomr(x_a \; a) = (x_a - \min a) / (\max a - \min a)$$

Where $\min a$ and $\max a$ are the lowest and highest values of attribute a, respectively, while x_a is the attribute's a value of sequence x.

Then, the similarity between a new sequence and the entire test sequences is calculated according to the similarity function. Using the similarity, we can describe the degree that a new sequence is similar to all sequences. Usually, we select the following function as the study's similarity function:

$$sim(x \; y) = \sqrt{\sum_{i=1}^{n} f(x_i, y_i)}$$

Where x, y are two protein sequences, and $f(x_i, y_i) = (x_i - y_i)^2$. If $x_i \neq y_i$, else $f(x_i, y_i) = 1$.

If $sim(x,z) = \underset{y}{\mathrm{m}\,i\,\mathrm{x}}\, sim(x,y)$, we believe that the sequence, x, belongs to the same class of y.

Evaluation of the performance

In order to easily compare the performance with other methods, we also use sensitivity (*Sn*), specificity (*Sp*), Matthew's correlation coefficient (*MCC*) and the overall prediction accuracy (*Ac*) as indicators (Baldi et al., 2000; Carugo, 2007) for evaluating the correct prediction rate and reliability of the study's method. Here:

$$Sn = TP/(TP + FN)$$

$$Sp = TP/(TP + FP)$$

$$MCC = (TP \times TN - FP \times FN) / \sqrt{(TP+FP)(TP+FN)(FN+TN)(TN+FP)}$$

$$Ac = (TP + TN)/M$$

Where *M* is the total number of protein sequences. *TP* denotes the number of the correctly recognized positives, *FN* denotes the number of the positives recognized as negatives, *FP* denotes the number of the negatives recognized as positives, and *TN* denotes the number of correctly recognized negatives.
In order to explain the study's method, here is an example:

2ABA_DROME: MGRWGRQSPVLEPPDPQ……AATNNLFIFQDKF is a protein sequence.

However, an introduction will be done first on how to extract the feature of the protein.

Step 1

Calculate f_t, the frequency of the 20 amino acids, in the aforementioned protein sequence.

(0.0501, 0.0581, 0.0681, 0.0160, 0.0701, 0.0481, 0.0561, 0.0180, 0.0842, 0.0601, 0.0220, 0.0301, 0.0401, 0.0240, 0.0681, 0.0541, 0.0741, 0.0641, 0.0681, 0.0261)

Step 2

Calculate $ApEn_i$ approximate entropy of protein sequences. First, represent the protein sequence as a time series X by replacing every amino acid of protein sequences by the relevant value of its hydrophobic amino acids; then, calculate the number of similar subsequences which begin at component i within X. As such, with length m, $C_i^m(r)$ can be obtained. At last, we can calculate $ApEn_i$, following the formula in "schemes of sequence feature". Consequently, the entire $ApEn$ sequences can be used to construct the following vector:

(1.3942, 1.4781, 1.4530, 1.4349, 0.6445, 0.8158, 0.9543, 0.9704, 0.1157, 0.1966, 0.3658, 0.4216)

Step 3

Calculate θ_j and the j-tier sequence correlation factor, but first calculate $H_k(R_i), k = 1,2,3$ (the value of hydrophobicity, hydrophilicity and side-chain mass of each amino acid), then through the correlation function, we could obtain θ_j. All sequence correlation factors can be used to construct the following vector:

(0.0040, 0.0043, 0.0040, 0.0043, 0.0040, 0.0041, 0.0040, 0.0042, 0.0037, 0.0040, 0.0040 0.0038, 0.0040, 0.0041, 0.0039, 0.0040, 0.0040, 0.0040, 0.0039, 0.0037)

Step 4

Merge the aforementioned three vectors into a vector as the formula in "schemes of sequence feature", and standardize it. In this research, where ϖ_1, ϖ_2 change in a certain range, they are 0.022 and 0.34, corresponding to the best prediction result, and then we can obtain the feature vector of this protein sequence.

Table 1. Performances of various methods in the prediction of bacterial toxins.

Method	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
Our method	94.52	97.87	0.9437	97.52
Amino Acids[a]	92.14	100	0.9293	96.07
Dipeptides[a]	86.43	98.57	0.8612	92.50

[a] comes from Saha and Raghava (2007).

Table 2. Performances of various methods in discriminating exotoxins and endotoxins.

Method	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
Our method	98.59	95.89	0.9469	97.33
Increment of diversity[b]	92.91	99.24	0.9428	
Amino Acids[a]	100	91.43	0.9293	95.71
Dipeptides[a]	94.29	91.43	0.8612	92.86

[a] comes from Saha and Raghava (2007), [b] comes from Yang and Li (2009).

Table 3. Comparison of two kinds of feature extraction methods for IB1 algorithm.

Method		Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
Bacterial toxins	Improved feature extraction	94.52	97.87	0.9437	97.52
	Amino acids alone	97.17	96.70	0.9339	97.07
Exotoxins and endotoxins	Improved feature extraction	98.59	95.89	0.9469	97.33
	Amino acids alone	93.65	80.82	0.7659	88

The second problem is how to predict the sequence. For the dataset, we first extract the feature of all protein sequences by using the aforementioned steps, before using the IB1 method classification to calculate them. A specific calculation process is that first, we select one sequence as the tested object and the others as the test set, and then use the IB1 algorithm to find the minimum similarity between the tested object and the others. We believe that the tested object is in the same type as that of the sequence which has the minimum similarity with the tested object. According to these steps, each sequence in the dataset is forecasted, in turn, after which we obtain the value of *TP, FN, FP* and *TN,* before we could calculate *Sn, Sp* and *MCC*.

RESULTS AND DISCUSSION

For the uniformity of comparison, in this paper, Jackknife test was used on the dataset. By programming and calculating, the performance of our method proposed for discriminating the bacterial toxins from non-toxins was shown in Table 1. The performances of other previous methods were also shown in Table 1. It was clear that our method with improved feature extraction and IB1 algorithm fusion was able to predict toxins with the total accuracy of 97.52% and 0.9437 *MCC*, which were higher than that of the previous results (Table 1).

The study's method was also used to predict whether a bacterial toxin was an exotoxin or an endotoxin. The total accuracy and *MCC* of this method achieved 97.33% and 0.9469*,* respectively, which were also higher than that of any other existed results (Table 2).

In order to further analyze the effectiveness of the algorithm and the effectiveness of feature extraction of the method proposed in this paper, we used IB1 algorithm to predict bacterial toxins based on the amino acid composition alone, and the results are listed in Table 3. From Table 3, we could see the difference between two performances of two feature extraction methods. It is obvious that the improved feature extraction, proposed in this paper, was indeed better than amino acids alone, which showed that our feature extraction method much effectively reflect the characteristics of bacterial toxins, and was more suitable for predicting bacterial toxins. Comparing Tables 1 and 2, we could see that the performance of IB1 algorithm is much better than that of SVM for bacterial toxins and non toxins with amino acids composition alone. Although IB1 algorithm was poor for discriminating exotoxins and endotoxins with amino acids composition alone, it was perfect when it was connected with the improved feature extraction. It showed that the combination of IB1 algorithm and the improved feature extraction method proposed in this paper could significantly improve the prediction accuracy of bacterial toxins.

ACKNOWLEDGEMENTS

The work was partly supported by the Fundamental Research Funds for the Central Universities (2010JC003).

REFERENCES

Aha DW (1990). A Study of Instance-based Algorithms for Supervised Learning Tasks: Mathematical, Empirical, and Psychological Evaluations, PhD dissertation, Dept. of Information and Computer Science, Univ. of California, Irvine.

Aha DW, Kibler D, Albert MK (1991). Instance-Based Learning Algorithms. Machine Learning, 6: 37-66.

Akerkar RA, Sajja PS (2010). Knowledge-based systems. Jones and Bartlett Publishers, Sudbury, MA, USA.

Argos P, Rao JK, Hargrave PA (1982). Structural prediction of membrane-bound proteins, Eur. J. Biochem., 128: 565-575.

Backlund A (2000). The definition of system, Kybernetes, 29: 444–451.

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16: 412–424.

Bagley KC, Abdelwahab SF, Tuskan RG, Fouts TR, Lewis GK (2002). Cholera toxin and heat-labile enterotoxin activate human monocyte-derived dendritic cells and dominantly inhibit cytokine production through a cyclic AMP-dependent pathway, Infect. Immun., 70: 5533-5539.

Boeckmann BA, Bairoch R, Apweiler MC, Blatter A, Estreicher E, Gasteiger MJ, Martin K, Michoud C, O'Donovan PI, Pilbout S, Schneider M (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res., 31: 365-370.

B¨ohnel H, Gessler F (2005). Botulinum toxins – cause of botulism and systemic diseases? Vet. Res. Commun., 29: 313-345.

Bokoch GM, Katada T, Northup JK, Hewlett EL, Gilman AG (1983). Identification of the predominant substrate Identification of the predominant substrate for ADP-ribosylation by islet activating protein, J. Biol. Chem., 258: 2072-2075.

Carugo O (2007). Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots. BMC Bioinformatics, 8: 380.

Chou KC (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 43: 246–255.

Cortes C, Vapnik V (1995). Support-Vector Networks. Machine Learn., 20: 273-297.

Dusenbery DB (2009). Living at Micro Scale, Harvard University Press, Cambridge, p. 124.

Gibson W (2003). Pattern recognition, G. P. Putnam's Sons, New York

Gilat A (2004). MATLAB: An Introduction with Applications. John Wiley and Sons, Hoboken, New Jersey.

Harnett MM (1994). Analysis of G-proteins regulating signal transduction pathways, Methods Mol. Biol., 27: 199-211.

Hopp TP, Woods KR (1981). Prediction of protein antigenic determinants from amino acid sequences, Proc. Natl. Acad. Sci. USA, 78: 3824-3828.

Howson C, Urbach P (1989). Scientific Reasoning: the Bayesian Approach. Open Court Pub Co, Peru.

Neer EJ (1995). Heterotrimeric G proteins: organizers of transmembrane signals, Cell, 80: 249-257.

Ochs MF, Rink L, Tarn C, Mburu S, Taguchi T, Eisenberg B, and Godwin AK (2009). Detection of Treatment-Induced Changes in Signaling Pathways in Gastrointestinal Stromal Tumors Using Transcriptomic Data, Cancer Res., 69: 9125-9132.

Pincus SM (1991). Approximate entropy as a measure of system complexity, Proc. Natl. Acad. Sci. USA, 88: 2297-2301.

Pullan W, Bhadeshia H (2000). Structure. Cambridge University Press, Cambridge, pp. 9-23.

Saha SG, Raghava GP (2007). BTXpred: Prediction of bacterial toxins, In Silico Biol., 7: 405-412.

Russell SJ, Norvig P (2003). Artificial Intelligence: A Modern Approach, Upper Saddle River, New Jersey.

Steen LA (1988). The Science of Patterns. Science, 240: 611-616.

Tamura BM, Chang B (2003). Botulinum toxin: application into acupuncture points for migraine. Dermatol. Surg., 29: 749-754.

Turing AM (1950). Computing machinery and intelligence, Mind, 59: 433-460.

Wei ZL, Dam GA (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, 22: 1658-1659.

Yang L, Li QZ, Zuo YC, Li T (2009). Prediction of Animal Toxins Using Amino Acid Composition and Support Vector Machine, J. Inner Mongolia University, 40: 443-448.