In Silico Biology 7, 0028 (2007); ©2007, Bioinformation Systems e.V.

BTXpred: Prediction of bacterial toxins

Sudipto Saha and Gajendra P. S. Raghava*

Institute of Microbial Technology Sector-39A, Chandigarh, India

* Corresponding author

Email: raghava@imtech.res.in
Phone: +91-172-2690557; Fax: +91-172-2690632
Web: http://www.imtech.res.in/raghava/

Edited by H. Michael; received January 29, 2007; revised April 19, 2007; accepted April 25, 2007; published June 11, 2007

Abstract

This paper describes a method developed for predicting bacterial toxins from their amino acid sequences. All the modules, developed in this study, were trained and tested on a non-redundant dataset of 150 bacterial toxins that included 77 exotoxins and 73 endotoxins. Firstly, support vector machines (SVM) based modules were developed for predicting the bacterial toxins using amino acids and dipeptides composition and achieved an accuracy of 96.07% and 92.50%, respectively. Secondly, SVM based modules were developed for discriminating entotoxins and exotoxins, using amino acids and dipeptides composition and achieved an accuracy of 95.71% and 92.86%, respectively. In addition, modules have been developed for classifying the exotoxins (e. g. activate adenylate cyclase, activate guanylate cyclase, neurotoxins) using hidden Markov models (HMM), PSI-BLAST and a combination of the two and achieved overall accuracy of 95.75%, 97.87% and 100%, respectively. Based on the above study, a web server called 'BTXpred' has been developed, which is available at http://www.imtech.res.in/raghava/btxpred/.

Supplementary information is available at http://www.imtech.res.in/raghava/btxpred/supplementary.html.

Keywords: bacterial toxins, exotoxins, endotoxins, BTXpred, prediction server

Introduction

The bacterial toxins are a major cause of diseases since they are responsible for the majority of symptoms and lesions during infection [Böhnel and Gessler, 2005; Blackall and Marques, 2004]. They can be classified into two categories; i) exotoxins, a soluble substance secreted by bacteria in the host tissues, and ii) endotoxins, generally residing within the cell wall and released into host tissues upon cell death [Prescott *et al.*, 1993]. The exotoxins act at a distance from the site of infection and can diffuse through the organism. The elucidation of the cellular mechanism of action of the bacterial exotoxins remains a complex problem, but they appear to share a common mechanism of action such as (i) binding to specific receptors on the plasma membranes of the sensitive cells, (ii) pore-formation, (iii) internalization or translocation across the membrane barrier and iv) direct secretion

[Middlebrook and Dorland, 1984; Popoff, 2005].

The exotoxins have a special affinity for particular tissues and may be divided into three categories on the basis of the site affected: i) neurotoxins act on nervous system, ii) enterotoxins on intestinal mucosa, and iii) cytotoxins on general tissue [Prescott *et al.*, 1993]. The neurotoxins recognize specific receptors on the unmyelinated areas of the presynaptic membrane and inhibit acetylcholine release. The enterotoxins act by activating adenylate cyclase or guanylate cyclase [Fishman, 1990]. Some staphylococcal enterotoxins cause the food poisoning syndrome [Dinges *et al.*, 2000]. The cytotoxins act on general tissues; for example, vacuolating cytotoxin is one of the most important virulence factors produced by *Helicobactor pylori*, a causative agent of severe gastric diseases such as ulcers and cancers [Montecucco *et al.*, 1999].

The bacterial toxins have a wide range of applications today. For example the cholera toxin (CT) and the heat-labile toxin from *E. coli* (LT) have been used as strong mucosal adjuvants in experimental models [Bagley *et al.*, 2002]; bacterial pore-forming hemolysins, such as listeriolysin O, have the potential for use in cytosolic drug delivery systems [Provoda and Lee, 2000]; botulinum toxin has been employed in orthopedics, physiatrics, gastroenterology, gynecology, neurology, pediatrics, general surgery, plastic surgery and several other specialties and also to treat hyperhidrosis and wrinkles in dermatology [Tamura and Chang, 2003]. The ability of the toxoid vaccine to induce toxin-neutralizing antibodies has provided the basis for the use of therapeutic antitoxins and immunoglobulins for the prophylaxis and treatment of diseases caused by bacterial toxins. The discovery of an effective method to detoxify tetanus and diphtheria toxins by formaldehyde treatment, allowed the introduction of mass immunization that led to almost complete elimination of both diseases from the developed countries [Glenny and Hopkins, 1923]. *Bacillus thuringiensis* subsp. *israelensis* and *B. sphaericus* are Gram-positive sporulating bacteria, which produce potent endotoxin crystals during sporulation and are highly toxic to susceptible mosquito larvae when they ingest them. These bacterial agents are environmentally safe as they are active at very low dosage and exhibit extreme host specificity [Thanabalu *et al.*, 1992; Bhattacharya, 1998]. Moreover, they can be produced economically in large-scale. All these have made bacterial toxins powerful experimental and clinical tools. Thus prediction of bacterial toxins would be very useful for the researchers working in the field of toxicology. The identification of any new potential bacterial toxins could be exploited for economic means, as bacteria are easy to culture in large scale.

In this study, a systematic attempt has been made to develop a method for predicting bacterial toxins, their class (exotoxin or endotoxin) and sub classes of exotoxins.

Methods

Collection and compilation of bacterial toxins

We obtained 185 bacterial toxins from Swiss-Prot database [Boeckmann *et al.*, 2003] which include 99 exotoxins and 86 endotoxins. Similarly, non-toxin protein sequences were obtained from Swiss-Prot by combined search using SRS (http://www.expasy.org/srs5/). The query was performed by searching for the term "function" in the "Comment" field but excluding entries with the term "toxin" in the same field by the 'BUTNOT' option of SRS. The retrieved protein sequences were checked in order to eliminate toxin proteins [Saha and Raghava, 2007]. The non-toxin protein sequences were from prokaryotic and eukaryotic origin and their broad function is listed in the Supplementary material. (For details see Supplementary material, http://www.imtech.res.in/raghava/btxpred/supplementary.html).

Data set

The PROSET software [Brendel, 1992] was used to remove sequences that have more than 90% sequence identity. The final non-redundant dataset consists of 150 bacterial toxins that have 77 exotoxins and 73

endotoxins. In the final dataset no two proteins have more than 90% sequence identity. The data set is available at http://www.imtech.res.in/raghava/btxpred/supplementary.html. The exotoxins were further classified based on their molecular targets, i) activate adenylate cyclase, ii) activate guanylate cyclase, iii) food poisoning, iv) neurotoxins, v) macrophage cytotoxin, vi) vacuolating cytotoxin and vii) thiol activated cytotoxin; viii) hemolysin.

Support vector machine

The SVM was implemented using the freely downloadable software package SVM_light [Joachims, 1999]. The software enables the user to define a number of parameters as well as to select from a choice of inbuilt kernel functions, including a radial basis function (RBF) and a polynomial kernel. There are different kernel and learning options available in SVM_light. In our study, we used "g" parameter which is gamma in RBF kernel. In learning options we used parameters like "c" , which is a trade-off between training error and margin and "j" which is a cost-factor, by which training errors on positive examples outweight errors on negative examples. Preliminary tests show that the radial basis function (RBF) kernel gives results better than other kernels. Therefore, in this work the RBF kernel was used for all the experiments. The input vectors used are amino acid composition (20 vectors) and dipeptide composition (400 vectors) of each protein sequence.

Protein features

Amino acids composition. The amino acid composition is the fraction of each amino acid in a protein. The fraction of all 20 natural amino acids was calculated using the following equations:

$$Fraction\ of\ amino\ acid\ i = \frac{total\ number\ of\ amino\ acids\ (i)}{total\ number\ of\ amino\ acids\ in\ protein}$$

where i can be any amino acid.

Dipeptide composition

The dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20 × 20). This representation encompassed the information about amino acid composition along local order of amino acid. The fraction of each dipeptide was calculated using following equation:

$$Fraction\ of\ dipep\ (i) = \frac{total\ number\ of\ dipep(i)}{total\ number\ of\ all\ possible\ dipeptides}$$

where $dipep(i)$ is one out of 400 dipeptides.

Hidden Markov model profile

The HMM profiles were generated for seven sub-classes of exotoxins using HMMER [Eddy, 1998]. The multiple alignment of protein sequences was obtained using CLUSTAL-W. The program hmmbuild of HMMER has been used to build profile HMM and then calibrated using hmmcalibrate. The program Hmmpfam was used for searching a query sequence against the created profile HMM database. The E-value 0.01 has been set as a cut off for searching the sequences against these profiles.

PSI-BLAST

A PSI-BLAST module has been designed for searching a query sequence against test dataset using PSI-BLAST [Altschul *et al.*, 1997]. In this study, PSI-BLAST has been run using three iterations and with cut-off E-value of

0.01. The module allows to predict bacterial toxin and sub-classes of exotoxins (adenylate cyclase, activate guanylate cyclase, food poisioning, neurotoxins, macrophage cytotoxin, vacuolating cytotoxin and thiol activated cytotoxin).

Evaluation of modules

Five-fold cross-validation: The performance of all the modules, developed in this study, has been evaluated using a 5-fold cross-validation technique. In 5-fold cross-validation, the dataset has been randomly divided into five sub-sets and four sub-sets used for training and the remaining sub-set for testing. This process is repeated five times so that each set is used for testing once. The performance of a method is an average performance of method on five sub sets. Five threshold-dependent parameters - sensitivity, specificity, accuracy, positive predictive value (PPV) and Matthew's correlation coefficient (MCC) [Baldi *et al.*, 2000] were used for evaluating performance of methods.

Leave-one-out cross-validation: The leave-one-out cross-validation (LOOCV) technique has been used for evaluating the modules developed for predicting sub class of exotoxins. In LOOCV, all the sequences except one are used for training and are tested on this remaining sequence. This process is repeated N times, where N is the total number of sequences in such a way that each sequence is used for testing once.

Results

Prediction and classification of bacterial toxins

Prediction of bacterial toxins: The performance of various modules developed for discriminating the bacterial toxins from non-toxins has been shown in Tab. 1. It is clear that simple SVM based modules using amino acid composition were able to predict toxins with high accuracy 96.07% with MCC 0.93. This is because the toxins have unique types of composition, which is different from non-toxins. Different parameters of RBF kernel were tried to optimize the SVM based method and the best results were obtained using RBF kernel with g = 1, c = 5000, j = 1 as shown in Table S1 (see Supplementary material). The performance of SVM module using dipeptides composition performs poorer than amino acid composition (see Table S2, Supplementary material). It may be due to the short sequence length of toxins as it is difficult to obtain significant number of dipeptides for small proteins. The performance of SVM based modules developed in this study was significantly higher than PSI-BLAST, a similarity search method used commonly.

Table 1: The performance of PSI-BLAST and SVM based modules (using amino acids and dipeptides composition) in prediction of bacterial toxins.

Performance/approach	Sensitivity	Specificity	PPV	Accuracy	MCC
Amino Acids	92.14%	100%	100%	96.07%	0.9293
Dipeptides	86.43%	98.57%	98.49%	92.50%	0.8612
PSI-BLAST (toxin)	67.14%				

PPV: positive predictive value; MCC: Matthew's correlation coefficient

Prediction of Class of Bacterial Toxins: The SVM modules were developed for predicting whether a bacterial toxin is an exotoxin or an endotoxin. The performance of various modules has been shown in Tab. 2. The SVM modules based on amino acid composition and dipeptide composition classify exotoxins and endotoxins with an accuracy of 95.71% and 92.86%, respectively. Different parameters of the RBF kernel were tried to optimize the best performance as shown in Tables S3 and S4 (see Supplementary material). It shows that the SVM modules,

based on amino acid composition, perform better than the dipeptides composition. The overall accuracy of PSI-BLAST was 45.71% and 90% for exotoxins and endotoxins respectively, which is lower than SVM based modules.

Table 2: The performance of PSI-BLAST and SVM based modules (using amino acid and dipeptides composition) in discrimination of exotoxins and endotoxins.

Performance/Approach	Sensitivity	Specificity	PPV	Accuracy	MCC
Amino acids	100%	91.43%	92.64%	95.71%	0.9203
Dipeptides	94.29%	91.43%	92.06%	92.86%	0.8596
PSIBLAST(exotoxin)	45.71%				
PSIBLAST(endotoxin)	90.00%				

PPV: positive predictive value; MCC: Matthew's correlation coefficient

Prediction of sub class of exotoxins: The functional annotation is an important feature and of major interest to the experimental biologists. An attempt was made to predict the functions of bacterial toxins. It has been observed that the target and mechanism of action of all exotoxins are not the same, thus exotoxins were grouped based on their target of action or function in seven categories. The endotoxins were not considered in this study, since all of them have insecticidal properties. The HMM and PSI-BLAST profiles were created for seven sub class of exotoxins. In the sub-classification of exotoxins, the SVM based method was not used due to the limited number of data in each class. The performance of various HMM, PSI-BLAST, and hybrid approach has been summarized in Tab. 3. The overall accuracy of the HMM-based approach was 95.75%, and that of the PSI-BLAST approach was 97.87%, but the hybrid approach was 100%. The result showed that one toxin that activates adenylate cyclase cannot be predicted by HMM based profile and by PSI-BLAST based profile. The toxin that was not predicted by HMM could be predicted by PSI-BLAST and *vice versa*.

Table 3: The accuracy of HMM and PSI-BLAST at e = 0.01 in classification of exotoxins based on function.

Class / Approach	Activate adenylate	Activate guanylate cyclase	Neurotoxin	Macrophage Cytotoxin	Vaculation	Thiol-activated cytotoxin	Poisioning	Overall
HMM (A)	100	87.50	100	100	100	100	90	95.75
PSI-BLAST (B)	100	87.50	100	100	100	100	100	97.87
Hybrid (A+B)	100	100	100	100	100	100	100	100

Discussion

The bacterial toxins such as *Diphtheria* toxin (DT), *Pseudomonas* exotoxin A (PE) and *Clostridium perfringens* enterotoxin (CPE) are currently used as therapeutic agents for solid tumors [Michl and Gress, 2004]. Some bacterial toxins are very potent and relatively easy to produce and classified as bio-threat agents. These include the botulinum neurotoxins, ricin, staphylococcal enterotoxin B, and the *Clostridium perfringens* epsilon toxin [Marks, 2004]. Thus, there is a need to assign the function of bacterial proteins in order to exploit the full potential of these proteins. Due to the advancement in sequencing technology, genomes of 296 bacteria are either completely sequenced or are in the advanced stage of sequencing (http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl/). However, the function of the limited number of bacterial proteins have been

assigned so far. Thus, there is an urgent need to develop computation tools for assigning the function of protein particularly of toxins.

In order to assist the bacteriologist in assigning the function of bacterial proteins, a systematic attempt has been made for predicting the bacterial toxins and their types. One of the major problems was how to create datasets of bacterial toxins, endotoxin, exotoxins and different type of exotoxins, as there is no database available on bacterial toxins. It was beyond the scope of this study to classify the available bacterial toxins because the classification of bacteria is a complex problem. In order to overcome this problem we extracted the bacterial toxins from Swiss-Prot and classified them as endotoxins, exotoxins and type of exotoxins based on Swiss-Prot annotation. We have examined the full text entries of Swiss-Prot manually (specifically comments) and extracted only those proteins whose function is assigned based on experimental observations. We have only considered the full-length proteins and filtered out all the partial sequences. The creation of a negative dataset (non-toxins) is as important as a positive dataset (toxins) for developing any classification method. Thus we have extracted from Swiss-Prot non-toxins of either prokaryotic or eukaryotic origin and used them as negative dataset in this study (http://www.imtech.res.in/raghava/btxpred/supplementary.html).

The SVM module based on amino acid composition achieved a slightly higher accuracy than the dipeptide composition, both in discriminating bacterial toxins and type of toxins (exotoxins or endotoxins). This observation was quite surprising when compared to the other SVM modules used for the prediction of families and subfamilies of G-protein coupled receptors and subcellular localization of human proteins [Bhasin and Raghava, 2004; Garg *et al.*, 2005]. It was observed that most of the toxins used in this study were short in length, therefore the frequency of occurrence of most of dipeptides in these sequences was very low or zero. In 'leave-one-out cross-validation' of the functions of exotoxins, the PSI-BLAST performed better than the HMM and the combined method predicted 100% accuracy, and it also showed that the functional toxin that could not be predicted by HMM could be predicted by PSI-BLAST and *vice versa*. So it is better to use the combined approach for functional analysis where there are unique profiles. In the sub-classification of exotoxins, SVM based method was not used, since the machine learning technique required adequate number of data for proper training and testing. In the functional prediction of exotoxins, hemolysin exotoxin was not considered, since there are more than 12 different functional domains performing the same function but the mechanism can be different as no significant alignment could be obtained. The web server developed, based on this study, will be very useful for researchers working in vaccine or drug development. This server allows the user to predict toxins and its type using SVM based module and PSI-BLAST method, where as HMM is used only in functional prediction of exotoxins.

Conclusion

The BTXpred server was developed for prediction of bacterial toxins and classifying them based on their release and function. This method, in association with the PSI-BLAST, can be used for automated annotation of genomic data. The study also proves that there is a direct correlation between the features of the proteins (amino acid, dipeptide composition) and the bacterial toxins. This server will also assist the preliminary analysis of possible functions of new exotoxins and in designing experiments for functional characterization of newly identified bacterial toxin sequences thereby reducing the number of essential experiments.

Based on our study, the BTXpred server has been developed that allows prediction of bacterial toxins and further classification of these toxins based on release and function. The server accepts the protein sequences in any standard format like EMBL, GCG, and FASTA or in plain text format. The server uses the readseq program (http://iubio.bio.indiana.edu/soft/molbio/readseq/) to read the input sequences. The server allows users to predict the bacterial toxins, its release, and further classification of exotoxins. The server provides the option of predicting toxins either on the basis of amino acid or dipeptide composition or PSI-BLAST. It also allows users to predict functions of exotoxins using HMM and PSI-BLAST methods. The server links to Bcepred server [Saha and Raghava, 2004] and ABCpred server [Saha and Raghava, 2006] for prediction of B-cell epitope in the bacterial toxins protein. This will help the users who are interested in generating antibodies against the toxin. The results provide summarized information about the query sequence and the prediction. The server and

related information is available from www.imtech.res.in/raghava/btxpred. A mirror site of BTXpred server is accessible from http://bioinformatics.uams.edu/mirror/btxpred/. The supplementary information is available at http://www.imtech.res.in/raghava/btxpred/supplementary.html.

Acknowledgements

We are grateful to Dr. Alok Mondal for reading manuscript critically. We acknowledge the financial support from the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Govt. of India.

References

- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389-3402.
- Bagley, K. C., Abdelwahab, S. F., Tuskan, R. G., Fouts, T. R. and Lewis, G. K. (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.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics **16**, 412-424.
- Bhasin, M. and Raghava, G. P. S. (2004). GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res. **32**, W383-389.
- Bhattacharya, P. R. (1998). Microbial control of mosquitoes with special emphasis on bacterial control. Indian J. Malariol. **35**, 206-224.
- Blackall, D. P. and Marques, M. B. (2004). Hemolytic uremic syndrome revisited: Shiga toxin, factor H, and fibrin generation. Am. J. Clin. Pathol. **121**, S81-S88.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**, 365-370.
- Böhnel, H. and Gessler, F. (2005). Botulinum toxins – cause of botulism and systemic diseases? Vet. Res. Commun. **29**, 313-345.
- Brendel, V. (1992). PROSET – a fast procedure to create non-redundant sets of protein sequences. Math. Comput. Model. **16**, 37-43.
- Dinges, M. M., Orwin, P. M. and Schlievert, P. M. (2000). Exotoxins of *Staphylococcus aureus*. Clin. Microbiol. Rev. **13**, 16-34.
- Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics **14**, 755-763.
- Fishman, P. H. (1990). Mechanism of action of cholera toxin. *In*: ADP-Ribosylating Toxins and G proteins. Insight into signal transduction, Moss, J. and Vaughan, M. (eds.), American Society of Microbiology, Washington, DC, pp. 127-138.
- Garg, A., Bhasin, M. and Raghava, G. P. S. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search. J. Biol. Chem. **280**, 14427-14432.
- Glenny, A. T. and Hopkins, B. E. (1923). Diphtheria toxoid as an immunizing agent. Brit. J. Exp. Pathol. **4**, 283-288.
- Joachims, T. (1999). Making large-scale SVM learning particle. *In*: Advances in kernal Methods Support

Vector Learning, Scholkopf, B., Burges, C. and Smola, A. (eds.), MIT Press, Cambridge, MA and London, pp. 42-56.

- Marks, J. D. (2004). Medical aspects of biologic toxins. Anesthesiol. Clin. North America **22**, 509-532.
- Michl, P. and Gress, T. M. (2004). Bacteria and bacterial toxins as therapeutic agents for solid tumors. Curr. Cancer Drug Targets **4**, 689-702.
- Middlebrook, J. L. and Dorland, R. B. (1984). Bacterial toxins: cellular mechanisms of action. Microbiol Rev. **48**, 199-221.
- Montecucco, C., Papini, E., de Bernard, M., Telford, J. L. and Rappuoli, R. (1999). Helicobacter pylori vacuolating cytotoxin and associated pathogenic factors. *In*: The Comprehensive Sourcebook of Bacterial Proteins Toxins, Alouf, J. E. and Freer, J. H. (eds.), Academic Press, London, pp. 264-283.
- Popoff, M. R. (2005). Bacterial exotoxins. Contrib. Microbiol. **12**, 28-54.
- Prescott, L. M., Harley, J. P. and Klein, D. A. (1993). Symbiotic associations: parasitism,pathogenicity and resistance. *In*: Microbiology 2nd edition. Wm. C. Brown Publishers, pp. 584-590.
- Provoda, C. J. and Lee, K.-D. (2000). Bacterial pore-forming hemolysins and their use in the cytosolic delivery of macromolecules. Adv. Drug Deliv. Rev. **41**, 209-221.
- Saha, S. and Raghava, G. P. S. (2004). BcePred: Prediction of continuous B-Cell epitopes in antigenic sequences using physico-chemical properties. ICARIS, LNCS **3239**, 197-204.
- Saha, S. and Raghava, G. P. S. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins **65**, 40-48.
- Saha, S. and Raghava, G. P. S. (2007). Prediction of neurotoxins based on their function and source. In Silico Biology **7**, 0025.
- Tamura, B. M. and Chang, B. (2003). Botulinum toxin: application into acupuncture points for migraine. Dermatol. Surg. **29**, 749-754.
- Thanabalu, T., Hindley, J., Brenner, S., Oei, C., and Berry, C. (1992). Expression of the mosquitocidal toxins of *Bacillus sphaericu*s and *Bacillus thuringiensis* subsp. *israelensis* by recombinant *Caulobacter crescentus*, a vehicle for biological control of aquatic insect larvae. Appl. Environ. Microbiol. **58**, 905-910.