Available online at www.sciencedirect.com

ScienceDirect

Journal of Theoretical Biology 243 (2006) 252–260

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

# Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification

Sukanta Mondal[a], Rajasekaran Bhavna[b], Rajasekaran Mohan Babu[b], Suryanarayanarao Ramakumar[a,b,*]

[a]*Department of Physics, Indian Institute of Science, Bangalore 560 012, India*
[b]*Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India*

Received 13 April 2006; received in revised form 6 June 2006; accepted 8 June 2006
Available online 27 June 2006

**Abstract**

Conotoxins are disulfide rich small peptides that target a broad spectrum of ion-channels and neuronal receptors. They offer promising avenues in the treatment of chronic pain, epilepsy and cardiovascular diseases. Assignment of newly sequenced mature conotoxins into appropriate superfamilies using a computational approach could provide valuable preliminary information on the biological and pharmacological functions of the toxins. However, creation of protein sequence patterns for the reliable identification and classification of new conotoxin sequences may not be effective due to the hypervariability of mature toxins. With the aim of formulating an in silico approach for the classification of conotoxins into superfamilies, we have incorporated the concept of pseudo-amino acid composition to represent a peptide in a mathematical framework that includes the sequence-order effect along with conventional amino acid composition. The polarity index attribute, which encodes information such as residue surface buriability, polarity, and hydropathy, was used to store the sequence-order effect. Several methods like BLAST, ISort (Intimate Sorting) predictor, least Hamming distance algorithm, least Euclidean distance algorithm and multi-class support vector machines (SVMs), were explored for superfamily identification. The SVMs outperform other methods providing an overall accuracy of 88.1% for all correct predictions with generalized squared correlation of 0.75 using jackknife cross-validation test for A, M, O and T superfamilies and a negative set consisting of short cysteine rich sequences from different eukaryotes having diverse functions. The computed sensitivity and specificity for the superfamilies were found to be in the range of 84.0–94.1% and 80.0–95.5%, respectively, attesting to the efficacy of multi-class SVMs for the successful in silico classification of the conotoxins into their superfamilies.

*Keywords:* Hypermutable mature conotoxin; Superfamily classification; Pseudo-amino acid composition; Polarity index; Support vector machines (SVMs)

## 1. Introduction

Predatory cone snails have perhaps been popular for their attractive shells and their deadly weapon, the venom. In general, conotoxins are biosynthesized as precursors, consisting of the N-terminal signal sequence, a propeptide region and the hypermutable mature toxin that has multiple disulfide connectivity (Jones and Bulaj, 2000; Terlau and Olivera, 2004). The intra- and inter-specific pharmacological diversity in them has evolved due to their unprecedented inventory of unique post-translational modifications (Jones and Bulaj, 2000; Craig et al., 1999). Conotoxins have been classified into several pharmacologically distinct superfamilies, namely A, I, M, O, T, P and S based on different characteristics such as highly conserved N-terminal precursor sequence, disulfide connectivity and similar mode of actions. Each superfamily is further classified into various families based on the cysteine arrangement, for example, A-superfamily includes $\alpha$, $\alpha$A, $\kappa$A families whereas M-superfamily possess $\mu$, $\psi$ families and O-superfamily possess $\delta$, $\mu$O, $\omega$, $\kappa$, $\gamma$ families (Olivera, 1997; McIntosh and Jones, 2001). The small size, specific cysteine pattern, hypermutation and the interactions between the conotoxins and the ion channels or receptors

*Corresponding author. Department of Physics, Indian Institute of Science, Bangalore 560 012, India. Tel.: +91 80 2293 2718, +91 80 2293 2469; fax: +91 80 2360 2602, +91 80 2360 0551.
*E-mail address:* ramak@physics.iisc.ernet.in (S. Ramakumar).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.06.014

make them a tremendous combinatorial pharmacological library for development of drugs (Terlau and Olivera, 2004; Olivera, 1997). Recently, the popularity of cone snail toxins has increased due to their potential to treat disease states such as chronic pain, epilepsy, cardiovascular disease, psychiatric disorders, cancer and stroke (McIntosh and Jones, 2001). Ziconotide's ($\omega$-conotoxin MVII) approval for chronic pain paves the way for the next generation of compounds for pain relief (Garber, 2005; Miljanich, 2004).

Protein sequence motifs are signatures of protein families and are often used as tools for the prediction of protein function and family classification. Systematically derived motif databases, such as PROSITE (Hulo et al., 2004), PRINTS (Attwood et al., 2003), are therefore feasible, allowing the classification of many of the newly appearing protein sequences into known (super) families. We have defined sequence patterns for $\alpha$, $\delta$, $\mu$ and $\omega$ conotoxin families, which are presently available in PROSITE with PS60014, PS60005, PS60013 and PS60004 accession numbers, respectively, under PDOC60004 documentation. There is only limited scope for encoding all the available information into protein sequence patterns due to the hypervariability of mature conotoxin sequences since their genes are evolving fast (Conticello et al., 2001). Moreover, the available three-dimensional structural information about the conotoxins in most cases is limited. Additionally, signal peptide and pro-peptide sequences for all the toxins are not often available and this makes it important to be able to solve the classification problem by using the sequences of the mature toxins. Hence we have investigated here, the possibility of using theoretical approaches for classifying conotoxins into their respective superfamilies based on the sequence of the mature toxin.

A number of computational methods have been suggested for protein classification based on the sequence alignment, consensus patterns using motifs, profiles, hidden Markov models and machine learning approaches (for more details see Cheng et al., 2005). The application of machine learning techniques to bioinformatics problems has become increasingly popular in recent years. The statistical prediction algorithms have proved to be effective in particular, for protein identification or classification. At such instances, the sequence or structural attributes used in classification schemes must represent the relevant information reflecting form and function for that class of protein. The computation of amino acid composition is a common approach but it does not take the sequence-order effect into account. The pseudo-amino acid composition enables the calculation of sequence-order correlation between the residues in a protein chain (Chou, 2001). We have used the discrete sequence correlation factors in conjunction with the 20 components of amino acid composition for building the features. The features generated by this method could be an useful input for classification system. A popular algorithm for classification is the support vector machines (SVMs). The SVMs, proposed by Vapnik (2000) is a remarkable statistical method among the machine learning algorithms. This is because SVMs are designed to maximize the margin to separate two classes so that the trained model generalizes well on unseen data (Yang, 2004). SVMs have gained popularity in the last few years due to their success in medicine, bioinformatics, computational biology and structure-activity relationships, and find applications in drug design (Burbidge et al., 2000), G-protein coupled receptors classification (Karchin et al., 2002), membrane protein types recognition (Cai et al., 2003), protein subcellular localization prediction (Garg et al., 2005), protein–protein interaction sites recognition (Bradford and Westhead, 2005) and phosphorylation sites recognition (Kim et al., 2004).

From the point of view of protein sequence and structure analysis, conotoxins can serve as attractive systems for the studies in sequence comparison, pattern extraction, structure–function correlations, protein–protein interactions and evolutionary analysis. Despite their importance and extensive experimental investigations on them, they have been hardly explored through in silico methods. In this report, we have investigated several methods such as the BLAST algorithm (Altschul et al., 1997), ISort (Intimate Sorting) predictor (Chou and Cai, 2006), least distance algorithms (Nakashima et al., 1986; Chou, 1980, 1989), and the multi-class SVMs to classify A, M, O and T superfamily of peptides. Generally, the dominating approach for solving multi-class problem is based on reducing a single multi-class problem into multiple binary problems (Cramer and Singer, 2001), known as the *one-versus-rest* (1-*v*-*r*) SVMs, though the problem of false positives is associated with this method (Ding and Dubchak, 2001). However, the optimization procedure described by Cramer and Singer (2001) eliminates this problem and has been implemented in this paper using a different algorithm (Tsochantaridis et al., 2004). We have examined both the approaches for solving the multi-class SVMs problem and have reported the benefits and shortcomings in the light of our results. The present case study is focussed on building an in silico method, based on the features derived from the sequence data to assign the conotoxins into their respective superfamilies, in order to provide beneficial preliminary information about the biological functions of the toxins.

## 2. Materials and methods

### 2.1. Data sets

The sequences for conotoxins were obtained from the Swiss-Prot release 47.1 (Bairoch et al., 2004). Since the number of entries in some superfamilies like P (five entries) and S (one entry) were very less, these superfamilies were not included for the analysis. The I-conotoxin superfamily was not included in view of the discussion in the literature