Sequence analysis

CASVM: web server for SVM-based prediction of caspase substrates cleavage sites

Lawrence J.K. Wee¹, Tin Wee Tan¹ and Shoba Ranganathan²,¹,¹

¹Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore and
²Department of Chemistry and Biomolecular Sciences & Biotechnology Research Institute, Macquarie University, Sydney, Australia

Received on April 27, 2007; revised on June 12, 2007; accepted on June 17, 2007

Advance Access publication June 28, 2007

Associate Editor: Alex Bateman

ABSTRACT
Summary: Caspases belong to a unique class of cysteine proteases which function as critical effectors of apoptosis, inflammation and other important cellular processes. Caspases cleave substrates at specific tetrapeptide sites after a highly conserved aspartic acid residue. Prediction of such cleavage sites will complement structural and functional studies on substrates cleavage as well as discovery of new substrates. We have recently developed a support vector machines (SVM) method to address this issue. Our algorithm achieved an accuracy ranging from 81.25 to 97.92%, making it one of the best methods currently available. CASVM is the web server implementation of our SVM algorithms, written in Perl and hosted on a Linux platform. The server can be used for predicting non-canonical caspase substrate cleavage sites. We have also included a relational database containing experimentally verified caspase substrates retrievable using accession IDs, keywords or sequence similarity.

Availability: http://www.casbase.org/casvm/index.html
Contact: shoba.ranganathan@mq.edu.au
Supplementary information: http://www.casbase.org/casvm/help/index.html

1 INTRODUCTION

Caspases belong to a unique class of cysteine proteases which function as critical effectors of apoptosis, inflammation and other important cellular processes, such as cell proliferation, cell differentiation, cell migration and receptor internalization (Algeciras-Schimnich et al., 2002; Launay et al., 2005; Los et al., 2002). Caspases contain a cysteine residue at the active site and cleave substrates at specific tetrapeptide sites (denoted P₀-P₁-P₂-P₃) with a highly conserved aspartate (D) at the P₁ position (Talanian et al., 1997). To date, at least 14 mammalian caspases have been discovered and they can be grouped into three classes based on their preferential tetrapeptide specificities (Thornberry et al., 1997). Group I caspases (-1, -4 and -5) recognize the sequence (W/L)EHHD; Group II caspases (-2, -3 and -7) prefer the sequence DEXD; while Group III caspases (-6, -8, -9 and -10) cleave proteins with the sequence (L/V)E(T/H)D.

As reviewed in Earnshaw et al. (1999) and Fischer et al. (2003), caspase substrates belong to a myriad of protein classes, such as structural elements of cytoplasm and nucleus, components of the DNA repair machinery, protein kinases, GTPases and viral structural proteins. Although more than 280 caspase substrates have been discovered to date, it is possible that several more remain undetected. The identification and characterization of caspase substrates will be critical for deepening our understanding of the role of these enzymes in the various cellular pathways. However, the accurate detection of caspase cleavage sites in target proteins requires complex and time consuming in vivo and in vitro experiments. Given the readily available sequence data in public databases, a useful alternative is to conduct in silico screening for potential cleavage sites among proteins.

At present, a number of tools are available for computational prediction of caspase substrate cleavage sites. PeptideCutter (Gasteiger et al., 2005), a general protease cleavage prediction server, employs the preferential caspase cleavage specificities of various caspases for prediction. Lohmuller et al. (2003) conceived the peptidase substrate prediction tool (PEPS) based on position-specific scoring matrices (PSSM) for caspase-3 substrate cleavage. Garay-Malpartida et al. (2005) developed the CasPredictor software which utilizes an algorithm which analyzes the cleavage sites for amino acid substitution, amino acid frequency and the presence of ‘PEST’ sequences, short for sequences containing proline (P), glutamate (E), aspartate (D), serine (S) and threonine (T) residues (Rechsteiner and Rogers, 1996; Rogers et al., 1986), in the vicinity of the cleavage site. The GraBCas software, created by Backes et al. (2005) provided an improved PSSM-based method which further accounted for the P₀₋₋₋ residues. Yang (2005) had previously applied different neural networks for predicting cleavage sites such as, single-layer perceptrons, multi-layer perceptrons and the Bayesian bio-basis function neural networks.

We have applied the support vector machine (SVM) algorithm to this domain and have shown that the SVM method achieved an accuracy ranging from 81.25 to 97.92%.
making it one of the best methods currently available (Wee et al., 2006). Here, we report the web server implementation of our SVM method, named CASVM, for the automated prediction of caspase cleavage sites on protein sequences. We have also constructed a relational database for easy retrieval of experimentally verified caspase substrates and provided the relevant datasets used in the development of our method on the website.

2 DESCRIPTION OF CASVM SERVER

Details on the SVM algorithms implemented in the server has been described previously in our earlier work (Wee et al., 2006). Briefly, an SVM classifier (termed P4P1-trained classifier) was trained with the sequences from a dataset containing unique caspase cleavage sites obtained from experimentally verified caspase substrates and an equal number of ‘non-cleavage’ sites (random tetrapeptide sequences extracted elsewhere on the same substrate). To account for the influence of adjacent residues on substrate cleavage, we constructed two other SVM classifiers trained with sequence segments containing the tetrapeptide cleavage sites with upstream two residues (P4P2-trained classifier) and tetrapeptide sequences with upstream 10 residues up to P14 and downstream ten residues up to P10 (P4P10-trained classifier) from the substrates. To minimize overtraining due to the high occurrence of aspartic acid at P1 position and enable the prediction of cleavage sites with residues other than aspartic acid at P1, the SVM classifiers were trained on sequences with the P1 residue removed. These SVM classifiers were used in CASVM and were shown to achieve an accuracy of 81.25, 89.58 and 93.75%, respectively when tested on datasets not used in training.

The server is written in Perl and is hosted on a Linux platform. The LIBSVM software package is used for implementing the SVM algorithm (Chang and Lin, 2001). The server homepage presents an intuitive interface for user input and processing. Users can submit (through copy and paste) a raw or FASTA-formatted protein sequence and select a number of processing. Upon form submission, the input sequence will be scanned over the entire length of the sequence when tested, sequence segments containing the specified P1 residue in the input sequence (optional), sequence length, an abbreviated version of the cleavage site with the selected SVM classifier. The output of the CASVM prediction displays the name of input sequence (optional), sequence length, an abbreviated version of the sequence, a list of potential cleavage sites (all tetrapeptide sequences with the specified P1 residue in the input sequence) and the CASVM-predicted cleavage sites. All cleavage sites are labeled with the P1 residue position.

3 DISCUSSION

CASVM is a web server for the SVM-based prediction of caspase substrates cleavage sites. As the substrates used in the method are derived from a variety of organisms (human, mouse, rat, fruit fly, cow, chicken, frog, worm and viruses) and are cleaved by various caspases (caspase -1, -3, -6, -7, -8, -9, -12, -13 and -14), the server is applicable to the detection of cleavage sites in substrates from various organisms and is not caspase-specific. In our analyses, we observed that the inclusion of residues in the immediate vicinity of the cleavage site increased the prediction accuracy (Wee et al., 2006). These findings are consistent with Backes et al. (2005) and Garay-Malpartida et al. (2005) reporting the role of the P1'-P1' residues and neighboring PEST sequences on substrate cleavage, respectively. Specifically, we have included SVM classifiers accounting for limited and extended sequences flanking the tetrapeptide cleavage site as represented by the P4P10-trained and P4P100-trained classifiers.

With the range of caspase substrates now available, a large variety of cleavage sites are presented, many of which differed markedly from the consensus tetrapeptide specificities reported in Thornberry et al. (1997). Although caspases are thought to be selective for aspartic acid at the P1 position, a notable number of substrates were cleaved at tetrapeptide sites bearing glutamic acid at the P1 position. Existing methods for caspase cleavage sites prediction (excepting ours) are largely limited to the discovery of cleavage sites with aspartic acid at P1 as they assume the consensus XXXD motif as the basis for their algorithms. The strict adherence to the inclusion of aspartic acid may be limiting the sensitivity of these tools since it is intuitively likely that many more substrates will have P1 residues as glutamic acid. Therefore, we have included the option to screen and predict for tetrapeptide sequences containing either aspartic acid or glutamic acid at the P1 position.

CASVM also contains a relational database of over 270 experimentally verified caspase substrates. The database, accessible at http://www.casbase.org/casvm/squery/index.html, can be queried using UniProt accession IDs (UniProt Consortium, 2007), keywords or the integrated BLASTP program (Altschul et al., 1990). The dataset of 219 unique caspase cleavage sites which was used for developing our SVM method can be downloaded from the website (Supplementary Material).

ACKNOWLEDGEMENTS

LJKW gratefully acknowledges the award of a research scholarship from the National University of Singapore.

Conflict of Interest: none declared.

REFERENCES