Vol.2017-BIO-51 No.9
2017/9/26

A short review of methods for Caspase cleavage site prediction

Yu Bao[1] Simone Marini[2] Takeyuki Tamura[1] Mayumi Kamada[1] Shingo Maegawa[1]
Hiroshi Hosokawa[1] Jiangning Song[3] Tatsuya Akutsu[1]

Abstract: Caspases/granzyme B cleavage is a fundamental part of proteolytic cleavage, which is involved into most aspects of cellular activities, including but not limited to gene regulation and cell life-cycle regulation. In this preliminary report we review and compare 7 state-of-the-art sequence-based bioinformatics approaches and tools for caspases/granzyme B cleavage prediction. We also conduct an independent dataset consisting of caspases/granzyme B substrates from various species and perform prediction using various predictors.

Keywords: Caspase substrate prediction, machine learning, tool analysis

Introduction

Proteases are a kind of enzymes that could hydrolyze peptide bonds to catalyze the breakdown of protein or peptide substrates. Of all the gene products in human about 2% (about 500 to 600) gene products belong to the category of proteases and these proteases are involved into most aspects of cellular activities, including but not limited to cell cycle, cell proliferation, programmed cell death, DNA replication, tissue remodeling and immune response [1].

Apoptosis, or programmed cell death, is an important mechanism that could be found in all tissues during development, homeostasis, and disease. Caspases are a family of proteases that have been reported to play a key role in driving the process of apoptosis or inflammation [2], [3]. The first two members of the caspase family was reported in 1993 [4] and evidences are found that these proteinases might play essential role in apoptosis. Subsequent studies of these proteinases lead to the discovery of several other caspase family members which also contribute greatly to apoptosis and/or inflammation. Thus it is important to identify native substrates of caspases and granzyme B in order to understand their physiological roles that have been implicated in the pathological processes.

To date, more than 15 mammalian caspases have been found [5] and they can be divided into three groups based on the substrates they specify: group I caspases, including caspase-1, 4, 5 and 13, this group prefer bulky hydrophobic amino acids at the P4 site and cleave the peptide sequence (W/L)EHD, group II caspases (capspase-2, 3 and 7) has a preference to cleave the sequence motif DEXD, whereas group III (caspase-6, 8, 9 and 10) cleaves the motif (I/V/L)E(H/T)D. In contrast to the caspases, GrB prefers to cleave the sequence motif IEXD.

Sequence and structural analysis of substrates of caspases and granzyme B have enabled the development of computational approaches for the prediction of caspases/granzyme B cleavage from sequence alone. However, the rapid growth in prediction approaches since the last comprehensive comparison, which was reported almost half a decade ago, creates an urgent need to critically assess and compare the now-large and diverse prediction methods.

In this preliminary report, we present a comprehensive analysis of 11 sequence-based methods for caspases/granzyme B cleavage prediction, contributing to the elucidate of the nature of different predictors and facilitating potential improvement of caspases/granzyme B cleavage prediction.

Tool analysis

In this preliminary report we make a comprehensive summarization to the details of the tools caspase/granzyme B cleavage prediction. These are GrabCas, CaSPredictor, PoPS, SitePrediction, Cascleave, Cascleave2, Pripper, PCSS, CASVM, CAT3, Blast.

We introduce and evaluate these tools according to their input types, ways of Models construction and development as well as their Prediction utility. More comprehensive parameters evaluated in this preliminary report could be found in Table 1.

Independent data test

In this section, to assess the performance of the review tools in an objective and fair manner, we constructed independent test datasets corresponding to Caspase 1, 3 substrates for Homo sapiens,in order to make a better evaluation of the prediction performance of these tools on other species, we also construct another two independent datasets corresponding to Caspase 1, 3 substrates for Escherichia coli and Mus musculus species, we per-

[1] Kyoto University, Kyoto, 6110011, Japan
[2] University of Pavia, Pavia, 27100, Italy.
[3] Monash University, Victoria, 3800, Australia.

Table 1 My caption

Evaluated parameters	Predictable species	Evaluated tools	PoPS
	Webservers		SitePrediction
	Algorithms		Cascleave
	Option of batch prediction		Pripper
	Adjustment of prediction thresholds		PCSS
	Standalone software		CAT3
	Language implemented		Blast
	Training dataset		GrabCas
	Ratio of positive to negative samples		CaSPredictor
	Shift window size		Cascleave2
	Time for processing a sequence		CASVM
	Whether,solvent accessibility(SA) and secondary structure (SS) is considered		

form our evaluation using these three kinds of datasets.

Since at the time when this preliminary report is written several tools are already not available, we will only perform independent data test on a part of the evaluated tools including PoPS, SitePrediction, Cascleave, Pripper, PCSS, CAT3, Blast.

Data preparation

We carefully prepared three kinds of datasets corresponding to the Caspase 1, 3 substrates for Homo sapiens, Escherichia coli and Mus musculus species, we extract all the fasta sequences of substrates of Caspase 1 and 3 according to species from Merops. To decrease the effect of overlap, we eliminate the sequences that are overlapped with the available training dataset of prediction tools including Cascleave and Cascleave 2.0, both of these tools are state-of-art which makes their training set cover most of the sequences in the training set of the tools developed before them. Out analysis shows that more than half of the sequences in the whole dataset are eliminated due to this action, leaving 70 Caspase 1 substrates as well as 121 Caspase 3 substrates in total for three kinds of species. For negative dataset, we randomly select proteins excluding proteins confirmed as substrates of Caspase 1, 3 of each species. To prevent biased performance, the sizes of negative datasets are the same as those of positive datasets. These datasets are named as Cas1-all and Cas3-all.

Performance evaluation for each tools

In this preliminary report we evaluate performance of each tools via ROC curve (receiver operating characteristic curve), Details of evaluation results will be reported elsewhere.

References

[1] Jasti S Rao. Molecular mechanisms of glioma invasiveness: the role of proteases. *Nature reviews. Cancer*, 3(7):489, 2003.

[2] Ting-Jun Fan, Li-Hui Han, Ri-Shan Cong, and Jin Liang. Caspase family proteases and apoptosis. *Acta biochimica et biophysica Sinica*, 37(11):719–727, 2005.

[3] Alan G Porter and Reiner U Jänicke. Emerging roles of caspase-3 in apoptosis. *Cell death & differentiation*, 6(2), 1999.

[4] Junying Yuan, Shai Shaham, Stephane Ledoux, Hilary M Ellis, and H Robert Horvitz. The c. elegans cell death gene ced-3 encodes a protein similar to mammalian interleukin-1β-converting enzyme. *Cell*, 75(4):641–652, 1993.

[5] Indrajit Chowdhury, Binu Tharakan, and Ganapathy K Bhat. Caspases an update. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 151(1):10–27, 2008.