CaMPDB: a resource for calpain and modulatory proteolysis

David A. duVerle[1] dave@kuicr.kyoto-u.ac.jp
Ichigaku Takigawa[1] takigawa@kuicr.kyoto-u.ac.jp
Yasuko Ono[2] ono-ys@igakuken.or.jp
Hiroyuki Sorimachi[2] sorimachi-hr@igakuken.or.jp
Hiroshi Mamitsuka[1] mami@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Kyoto University, Gokasho, Uji 611-0011, Japan
[2] Calpain Project, Rinshoken, Tokyo 156-8506, Japan

Keywords: calpain, database, modulatory proteolysis, cleavage prediction

1 Introduction

CaMPDB is built as a resource for modulatory proteolysis, focusing on calpain, a Ca^{2+}-dependent Cys protease. Calpains constitute a major protease family distributed over a wide range of organisms. More importantly, calpains regulate substrate functions by limited proteolysis, *i.e.* proteolytic processing, resulting in the modulation of a wide variety of biological phenomena. Malfunction of calpain has been observed in several serious diseases, including muscular dystrophies and diabetes [1]. Activity of calpain in vivo is regulated by its specific endogenous inhibitor protein: calpastatin [2]. Although calpastatin binds to calpain around the catalytic site cleft like regular substrates, it is resistant to proteolysis by calpain.

In order to help consolidate the knowledge of molecular entities involved in modulatory proteolysis and improve general understanding of this process, we have built CaMPDB with the following unique features: 1) Large collection of calpain sequences gathered from a variety of organisms, along with information on substrates (cleavage sites) and calpastatin. 2) Confirmation of collected cleavage sites through biochemical experiments. 3) Statistics based on collected data. 4) Built-in cleavage prediction tool.

2 Database Framework

CaMPDB has a total of 3,723 entries over 363 organisms, separated into three categories: 'Calpain', 'Substrates' and 'Calpastatin'. Each category is broken down into four standard subsections: 'Overview', 'Browse', 'Search' and 'Statistics', with an additional 'Predict' subsection for 'Substrates'.

Each entry includes basic nomenclature data, cross-references to six major databases (Entrez, Swiss-Prot, OMIM, HGNC, HPRD and KEGG), structure information consisting of: domain information, amino acid sequence with computationally predicted secondary structures, 3D structure and references linking to PubMed. Each entry in 'Substrates' has some additional fields regarding cleavage sites (see section 3). Entries in each category can be retrieved in two ways: through browsing (as a flat list or a clickable taxonomy tree) or by query search (using name, query ID or any other text annotation).

'Statistics' shows a summary of statistics for the corresponding category along with some data analysis results such as amino acid preferences of cleavage sites or philogenetic analysis of calpastatin's functional domains across species.

For Calpain, we collected 1,496 entries from Entrez that were marked with *CysPc*, indicating Ca^{2+}-dependent cysteine proteinase superfamily.

For 'Substrates', we collected 104 known calpain substrates (SB) from literature, resulting in 267 cleavage sites being stored in CaMPDB. We biochemically evaluated cleavage efficiency of 56 sites, using purified calpains and iTRAQ™ labeling. Each iTRAQ™ signal value gives a coarse estimate of proteolysis efficiency of the corresponding peptides by calpains.

In addition, we computationally expanded the set of SB sequences by an additional 1,914 putative substrates (XSB) using BLAST and a defined set of rules, to select sequences in Entrez that have satisfying similarity, both overall and around the cleavage site.

For 'Calpastatin', we collected 209 sequences in Entrez that included *Calpain_inhib* anywhere in the entry and *calpastatin* in the definition line.

3 Cleavage Site Prediction

The 'Predict' tab of the 'Substrates' section provides a prediction tool that will output position-based scores reflecting the probability of cleavage by calpain for any given sequence. This tool offers three separate prediction model options based on PSSM and Support Vector Machines (SVM) algorithms with either linear or Radial Basis Function (RBF) kernels. For each model, a sliding window of length L is ran along the sequence, with scores calculated for each subsequence.

During training, all models used the same set of 267 positive instances. SVM models additionally used a random sample of negative instances (selected out of all non-cleaved positions in SB sequences).

Top performances for each method were measured using Area Under the ROC Curve (AUC) with 10x10-fold cross-validation and reached maximal values of 69.1% (SEM: 0.75%) for PSSM (L = 2x30), 77.3% (SEM: 0.04%) for SVM linear and 80.1% (SEM: 0.04%) for SVM RBF (using respectively L = 2x7 and L = 2x10 amino acids on each side). With best prediction scores obtained within 10 amino acids of the cleavage site for SVM-based predictors, it seems possible to infer some early hypotheses regarding amino acid position specificity. For smaller window sizes (between 5 and 10 amino acids on each side), RBF performs significantly better than linear kernel. Finally, the study of results for varying lengths of extension around cleavage sites reveals some interesting asymmetry between left and right side: for both RBF and linear kernel SVM predictors, statistically better performances are achieved when focusing on the right (primed) side of the cleavage site.

Based on empirical evidence and performances during cross-validation testing, the value of L for the online tool was set to 40 amino acids (20 on each side of the cleavage site) for PSSM and 20 (2x10) for SVM. These values are justified by the 3D structure of a calpain/calpastatin complex [3]: calpastatin firmly binds to the calpain protease domain by approximately 20 residues.

4 Discussion

We are currently focussing on obtaining more interpretable results on cleavage prediction, through the use of rule-extraction methods conjugated with boosting meta-algorithms.

References

[1] L. Bertipaglia and E. Carafoli. Calpains and human disease. *Subcell Biochem.*, 45:29–53, 2007.

[2] D. E. Goll et al. The calpain system. *Physiol Rev.*, 83:731–801, 2003.

[3] T. Moldoveanu, K. Gehring, and D.R. Green. Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains. *Nature*, 456(7220):404–408, 2008.