A computational approach for detecting peptidases and their specific inhibitors at the genome level

Fariselli P (2), Bartoli L (2), Calabrese R (2), Mita DG (1), Casadio R (2)

(1) Department of Experimental medicine, University of Naples Federico II, Naples, Italy
(2) Laboratory of Biocomputing, CIRB/Department of Biology,
University of Bologna, Bologna, Italy.

Motivation

Peptidases are proteolytic enzymes essential for the life of all organisms and responsible for fundamental cellular activities, such as protein turn-over and defense against pathogenic organisms. It is known that in Eukaryots about 2-5% of the genes encode for peptidases and peptidase homologs irrespectively of the organism source. The basic protease function is however "protein digestion". This activity can be potentially dangerous in living organisms, if not strictly controlled and for this reason, in vivo several specific inhibitors are present. Four major classes of peptidases are identified by the catalytic group involved in the hydrolysis of the peptide bond: 1) Serine Peptidase; 2) Aspartic Peptidases; 3) Cysteine Peptidase; 4) Metallopeptidase. Here we give a solution to the following basic problems: 1) recognize in human and mouse genomes both protease and protease inhibitor sequences; 2) predict which inhibitor is specific for a given peptidase. More specifically we address the following questions: 1) Given a pair of sequences, are they a pair of protease and inhibitor that can interact? 2) Given a protease (or inhibitor), how can we compute the list of the proteins in a defined database that can inhibit (or be inhibited by) the query protein? 3) Given a proteome, how can we compute the lists of peptidases and their relative inhibitors for each of the protease class described above?

Methods

In the last years, an invaluable source of information about proteases and their inhibitors has been made available through the MEROPS database, so that it is possible to look for peptidase or peptidase-inhibitor sequences (or structures). We tested PROSITE and PFAM and we integrate them in a unique framework taking advantage of their different behavior in detecting proteases and their inhibitors. We analyze the four major protease classes (Serine, Cysteine, Aspartic and Metal) and we test PROSITE and PFAM accuracy in the detection of proteases and inhibitors with respect to a non redundant set of globular proteins. In order to predict whether pairs of peptidases and inhibitors belong to the same class, we developed a system that performs two consecutive tasks: 1) extracts protease and inhibitor sequences from a given data set and labels them as belonging to one of the 4 classes; 2) tests whether the inhibitor can interact with the protease (yes, when the two sequences belong to the same class, eg serine peptidase and serine peptidase inhibitor). In order to address this problem, we designe a decision-tree method that processes the information obtained from PROSITE and PFAM and detects whether a query sequence can be classified as peptidase or inhibitor. The decision tree first uses PROSITE scan the data base in order to extract a protease and an inhibitor. When it fails the decision tree switches to PFAM.

Results

We first test PROSITE in the task of detection proteases and protease-inhibitors against a non redundant set of globular proteins that does not contain them. The accuracies are 70 and 90% for the proteases and inhibitors, respectively. Interestingly the number of false positive is nearly 0. This indicates that PROSITE has a very high specificity. When PFAM is scored using the same data sets the accuracies become 95 and 97% for the proteases and for the inhibitors, respectively. When the decision tree is adopted accuracies further improve of one percentage point (96% and 98%), suggesting that PROSITE specificity increases PFAM accuracy. However, in order to be able to measure the real system accuracy in the task of selecting pairs of proteases and interacting inhibitors, we compute the score of detecting pairs of protease-inhibitor among all possible pairs, namely peptidase/inhibitor, peptidase/other, inhibitor/other, peptidase/peptidase, inhibitor/inhibitor, other/other, excluding the self-combinations (a sequence against itself). Proceeding in this way we

ended up with a number of 18.559.278 pairs. We divided peptidase sequences in the four classes according to their biological activity as detailed above. We labeled the inhibitors accordingly, with the exception of one more class (U) containing inhibitors that are reported to be able to inhibit to some extent all types of peptidases (Universal inhibitor). Out of all the possible 18.559.278 pairs only the ones that pertains to proteases and inhibitors of the same class are counted as members of the positive class (true positive in the confusion matrix), which amounts to less than 7% of the all possible pairs. Scored on this stringent data set, the decision tree method achieves a joint global accuracy of 94% with a coverage for the positive cases (protease-inhibitor) of 99%.

Availability: http://www.biocomp.unibo.it/

Contact email: remo@biocomp.unibo.it