Title	Data Fusion for cleavage target prediction
All Authors	Marini S(1), Demartini A(2), Vitali F(2), Bellazzi R(2), Akutsu T(1)
Affiliation	

(1) Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan.
(2) Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy.

Motivation

Protein cleavage is a pivotal process in cell metabolism. It is involved, among other processes, in cell differentiation and cycle control, stress and immune response, removal of abnormally folded proteins and cell death. Proteases (i.e. protein responsible for cleavage) account for ~2% of all gene products. As consequence, wrongly regulated proteolytic activity may result in diseases. The problem of predicting cleavage targets have been addressed by a number of algorithms [1]. Traditional prediction models tackle the cleavage target machinery encoding directly related information to the outcome class (e.g. by extracting sequence patterns or frequency matrices). We are aware, however, that a huge amount of indirectly-related information is available in public data sets. Peptidases and targets are both proteins, and share similarities as well as non-cleavage interactions in knowledge bases; they are both encoded by genes, and gene interactions are also in databases. Our proposed Data Fusion algorithm leverages on these secondary information sources to infer novel peptidase targets.

Methods

Our approach is based on tri-factorization [2]. The multiplicity of data are fused by inferring a joint model, and without altering their original structure, i.e. data are explicitly represented in the form of a relational block matrix *R*. Diagonal blocks of *R* are set to 0, while other blocks are sparse matrices, populated with the relations harvested from the various data sources. *R* elements are constrained into the range [0, 1], where 0s represent negative or unknown relationships, while 1s are interpreted as certain relationships. We considered three elements in our matrix, namely peptidases, targets and genes. From MEROPS we obtained 657 human peptidases affecting 3460 targets and forming 8931 pairs. From their mapping on Uniprot, 3833 genes coding for peptidases or targets were retained. This information was used to populate the peptidase-target, peptidase-gene and target-gene *R* blocks.

During the data fusion process, each *R* block is decomposed into three sub-matrices, characterized by low dimensions (if compared to the original *R* block size). There is no clear consensus about a technique to define these dimensions [2], and we proceeded by choosing a rank for a given block based on the number of known interactions. Once the dimensions are set, the three sub-matrices are used to reconstruct a user-defined target block *Rt*. *R* decomposition is obtained through an iterative process, where constraint matrices play an important role. Constraint matrices are populated with the associations relating objects of the same type. In our application we utilized five constraints: one gene-gene interaction matrix from BIOGrid; two target-target and protease-protease interaction matrices from STRING (0.7 as combined score threshold); two target-target and protease-protease BLAST similarity matrices (10^{-10} as e-value threshold). Once the convergence is reached, the target block *Rt* is examined to infer novel relationships. Our objective is to find protease-target putative interactions, therefore our target block is the protease-target one. To detect a new interaction, we applied the row-centric rule [2]. Note that since the iterative process starts from a random initialization, we repeated the whole data fusion process 15 times with different random initializations and retained only the interactions that satisfied the row-centric rule in all the runs.

Results

1787 new protease-targets were predicted by our approach, involving 139 proteases and 716 targets. To validate our results, we utilized an independent algorithm, CasCleave [1]. CasCleave is based on traditional

Machine Learning, therefore it is complementary to our data fusion approach. Though our approach pinpointed targets for all possible peptidases, we could validate only the 73 Caspase-interacting subset of our targets, since CasCleave has a limited scope. By comparing the cleavage CasCleave probability distributions of our predicted targets with the ones over the whole human proteome, we found all predictions are in accordance with CasCleave (likelihood >0.5; mean 0.82); 6 new targets predicted for Caspase-1 (p-value 1.23×10^{-5}), 37 for Caspase-3 (p-value 2.2×10^{-16}); 4 for Caspase-6 (p-value 6.8×10^{-4}); 5 for Caspase-7 (p-value 9.14×10^{-3}); 4 for Caspase-8 (p-value 1.1×10^{-3}); and 17 for Granzyme B (p-value 6.84×10^{-3}). P-values were computed with KS test. In future research we will expand the list of considered objects in the data fusion (e.g. domains) and validate our results with wet lab experiments.

[1] Wang, Mingjun, et al. "Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets." Bioinformatics (2013): btt603.
[2] Zitnik, Marinka, and Blaz Zupan. "Data fusion by matrix factorization." Pattern Analysis and Machine Intelligence, IEEE Transactions on 37.1 (2015): 41-53.