Computational Prediction of Modular Domain-Peptide Interactions

Kousik Kundu[1], Fabrizio Costa[1], and Rolf Backofen[1,2]

{kousik, costa, backofen}@informatik.uni-freiburg.de

[1]Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, Freiburg D-79110
[2]Centre for Biological Signalling Studies, BIOSS, University of Freiburg, Albertstr. 19, Freiburg D-79104

Introduction

Protein-protein interactions are the most essential cellular process in eukaryotes that involve many important biological activities such as signal transduction, maintaining cell polarity etc. Many protein-protein interactions in cellular signaling are mediated by modular protein domains. Peptide recognition modules (PRMs) are an important subclass of modular protein domains that specifically recognize short linear peptides to mediate various post translation modifications. Computational identification of modular domain-peptide interactions is an open problem with high relevance. In this study we applied machine learning approaches to identify interacting and non-interacting pairs of modular protein domains (SH2, SH3, PDZ) and peptides. Additionally we developed an approach to tackle the data imbalance problem using a powerful semi-supervised strategy.

SH2PepInt

SH3PepInt

PDZPepInt

SH2PepInt

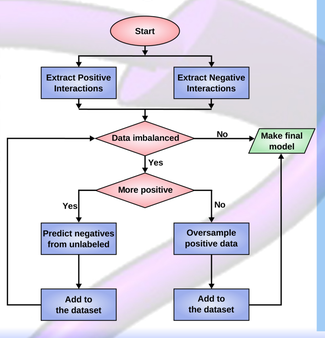
Approach:

- Models high-order correlations between amino acid position
- Tackles data imbalance problem using semi-supervised strategy
- Genome-wide analysis

Data: Microarray and peptide array data

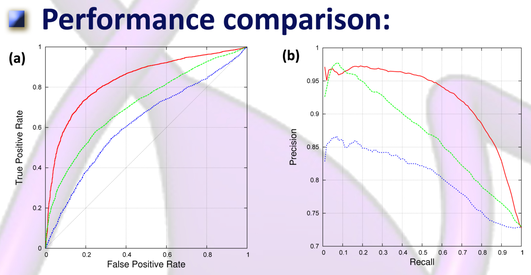
Flowchart for the iterative negative data filtering:

An initial high quality dataset is extracted from experimental evidence. If the negatives are in excess (right branch) then we simply duplicate the positive instances. If the positives are in excess (left branch) then we make an initial model using over-sampled negatives; this model is then used to score all the available peptides. Those that are more confidently predicted as negatives are added to the dataset. The procedure is iterated until a balanced dataset is reached. The final model is computed on the balanced dataset.



Results:

Performance comparison:



A 5-fold cross-validation performance

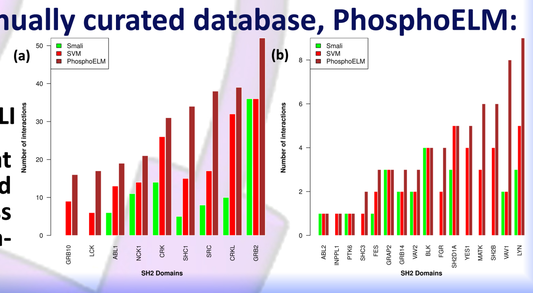
(a) AUC ROC comparison:
SVM - 0.83 (solid red line)
SMALI - 0.71 (dashed green line)
Energy model - 0.62 (dotted blue line)

(b) AUC PR comparison:
SVM - 0.93 (solid red line)
SMALI - 0.87 (dashed green line)
Energy model - 0.81 (dotted blue line)

Performance evaluation on manually curated database, PhosphoELM:

Brown bars: validated interactions
Red bars: predicted interactions by SVM
Green bars: predicted interactions by SMALI

(a) Showing those SH2 domains having at least 10 interactions in PhosphoELM 9.0 and (b) Showing the SH2 domains having less than 10 interactions in PhosphoELM 9.0 database.

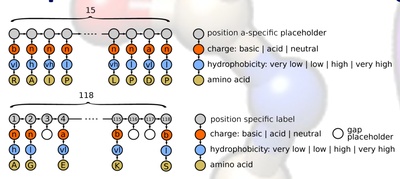


SH3PepInt

Approach:

- No need for an initial peptide alignment
- Models high-order correlations between amino acid position
- Integrates physico-chemical information
- Genome-wide analysis

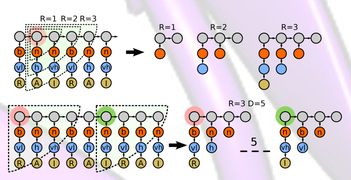
Graph kernel feature encoding:



Charge: Basic (R, K, H), Acidic (D, E), Neutral (Other amino acids)
Hydrophobicity: Very high (I, L, V), High (A, M, C, F), Low (G, T, S, W, Y, P) and Very low (Other amino acids)

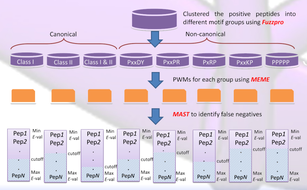
Data: Peptide array data

Graph kernel feature:



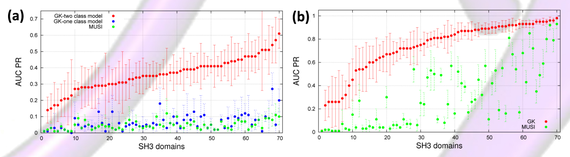
R= Radius, D= Distance

High quality negative filtering:



Results:

Performance comparison:

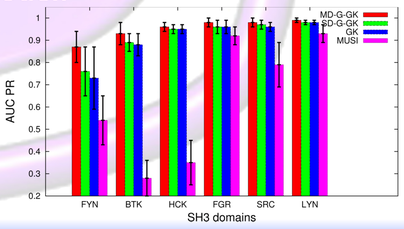


A 10-fold cross-validation performance

(a) Comparison with nonfiltered negative interactions for binary class Graph Kernel (GK), one-class Graph Kernel and MUSI. (b) Comparison when using filtered negative interactions for Graph Kernel (GK) and MUSI. The error bars represent respective standard deviation. The domains are sorted by increasing average performance for the Graph Kernel method.

Performance comparison with multi-domain model:

AUC PR for the Multi-Domain Gaussian Graph Kernel (MD-G-GK), the Single Domain Gaussian Graph Kernel (SD-G-GK), the Single Domain Linear Graph kernel (GK) and the MUSI tool for six related SH3 domains. The error bars represent respective standard deviation.



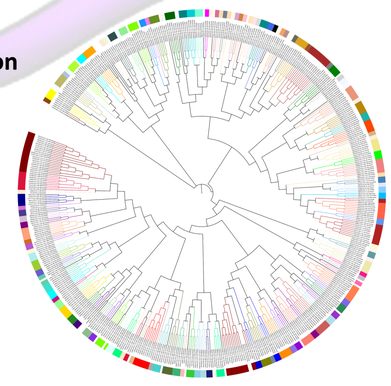
PDZPepInt

Approach:

- PDZ domain clustering based on their binding specificity
- Models high-order correlations between amino acid position
- Negative data filtering using semi-supervised technique
- Genome-wide analysis

Data: Microarray and peptide array data

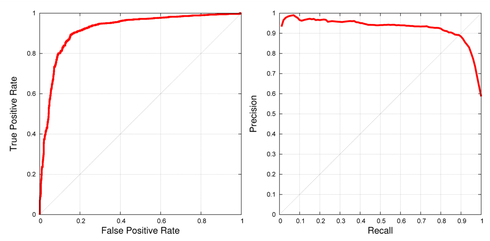
Phylogenetic tree of all available PDZ domains from human, mouse, fly and worm. The MCL clustering output was mapped onto the phylogenetic tree. A total number of 138 PDZ families are presented by 138 colors. iTOL was used for the visualization.



Results:

Performance:

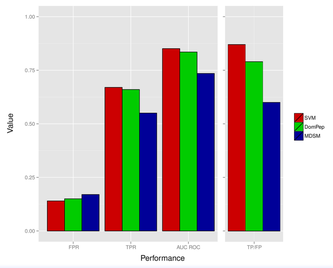
A 5-fold cross-validation performance
(a) The AUC-ROC and (b) the AUC-PR obtained by sequence-based feature encoding method.



AUC ROC - 0.92
AUC PR - 0.94

Performance evaluation on independent test-set:

Performance comparison of three different tools. Red, green and blue bars indicate the predicted performances by our tool (SVM), DomPep and MDSM, respectively. The figure clearly shows that our tool (SVM) achieved better performance.



References

1. Kundu K, Costa F, Backofen R: A graph kernel approach for alignment-free domain-peptide interaction prediction with an application to human SH3 domains. Bioinformatics 2013, 29(13):i335-i343
2. Kundu K, Costa F, Huber M, Reth M, Backofen R: Semi-Supervised Prediction of SH2-Peptide Interactions from Imbalanced High-Throughput Data. PLoS One 2013, 8(5):e62732
3. Kundu K, Backofen R: Cluster Based Prediction of PDZ-Peptide Interactions, BMC Genomics 2014, 15 (Suppl 1):S5