Machine learning approaches for prediction of SH2-peptide interactions

Kousik Kundu[1,2], Rileen Sinha[1], Michael Reth[2,3], Michael Huber[4], Rolf Backofen[1,2]

{kousik, 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

[3]Department of Molecular Immunology, Max Planck-Institute of Immunology, Freiburg D-79108

[4]Department of Biochemistry and Molecular Immunology, Institute of Biochemistry and Molecular Biology, University Clinic, RWTH Aachen University, Pauwelsstr. 30, Aachen D-52074

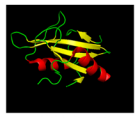
bioss
CENTRE FOR BIOLOGICAL SIGNALLING STUDIES

Summary:

Computational identification of SH2-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 SH2 domains and phosphotyrosine containing peptides

Introduction:

Src homology 2 (SH2) domains are structurally conserved protein domains, found in many intracellular signal-transducing proteins. Phosphorylation of tyrosine residues by tyrosine kinases is an important part of signal transduction. SH2 domains are the largest family of peptide recognition modules (PRMs) that recognize phosphotyrosine containing peptides. Hence, these domains have a vital role in cellular signaling. Around 120 SH2 domains have been identified in 110 human proteins and each SH2 domain binds with a specific subset of peptides. Therefore, peptide motif recognition by specific SH2 domains is important for understanding its biological function. Currently only a few programs have been published for the prediction of the SH2-peptide interactions but most of them are based on position specific weight matrices (PWMs) which ignore modeling the dependencies between the amino acids. Furthermore, these tools either don't model for all human SH2 domains or/and are not publically available. In the current study we are developing a machine learning approach for prediction of SH2-peptide interactions, which shall be made publically available.



Materials:

- The data are taken from microarray experiments and peptide array library

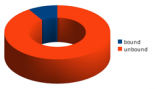
- Microarray Data:

Dataset I:
SH2 domains: 105
Peptides: 20
Possible interactions 105 X 20 = 2100
Positive interactions: 160
Negative interactions: 2100-160 = 1940

Dataset II:
SH2 domains: 85
Peptides: 41
Possible interactions 85 X 41 = 3485
Positive interactions: 314
Negative interactions: 3485 - 314 = 3171

- Peptide Array Library Data:

Dataset III:
SH2 domains: 61
Peptides: 920
Positive interactions: 7455

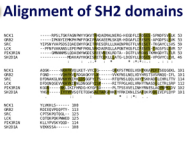


- Total interactions

Positive interactions: 160 + 314 + 7455 = 7929
Negative interactions: 1940 + 3171 = 5111

Feature Selection:

- Alignment of SH2 domains



- Alignment of short peptides

BLNK	SSDDDYDDVDI
CRK	SEETPYSYPTG
CRK	ALSPLYGVPET
CRK	IDPFTYEDPNE
CRKL	FREGFYLFPDG
CRKL	NEDDGYDVPKP
CSK	SSSPVYGSSAK
GRB10	YETDYYRKGGK
GRB2	FDDPSYVNVQN
GRB2	HVNATYVNVKC
NCK1	KEDPIYDEPEG
NCK1	NSVVLYSTPPI
SRC	DPKPHYDNTAG
SRC	TGNHTYQEIAV

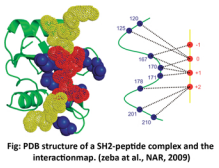


Fig: PDB structure of a SH2-peptide complex and the interactionmap. (zeba at al., NAR, 2009)

Microarray data
&
Peptide array library data

Feature Selection

Important Features

Attribute Evaluator
CfsSubsetEval
ClassifierSubsetEval

Search Method
BestFirst
RankSearch

Classification:

- Main classifiers used:

NaiveBayes
Bayesian Network
Support Vector Machine (SVM)
Meta Learner

- Best Result:

Bayesian Network Classifier with TAN algorithm

- Cross Validation: 10 fold

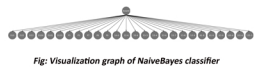


Fig: Visualization graph of NaiveBayes classifier

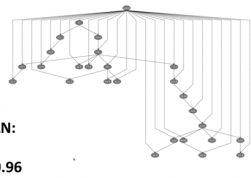


Fig: Visualization graph of BayesNet TAN Algorithm

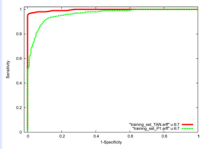


Fig: The ROC curves for the NaiveBayes algorithm (red) and the BayesNet TAN algorithm (dashed green line)

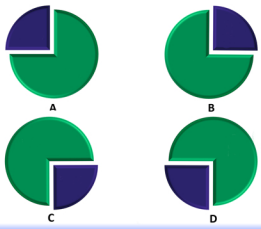
- NaiveBayes:
  Sensitivity: 0.92
  Specificity: 0.91

- BayesNet TAN:
  Sensitivity: 0.96
  Specificity: 0.99

Results:

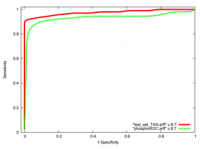
- Making Training and Test sets

The non-redundant dataset is partitioned into 2 parts. 75% data for training set and 25% data for test set.



- Test on In-vivo data

The model has been tested on the fairly reliable experimental dataset of Phospho binding domain, PhosphoELM (Positive) and literature based (Negative)



ROC Area: 0.96
ROC Area: 0.91

Fig: The ROC curves for the test set (25%, red) and the experimental data (dashed green line)

Comparison with other tools

- Our tool has been compared with other available tools which clearly showed that our tool has higher accuracy.

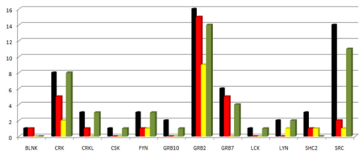


Fig: Compared tools for the prediction of the experimentally validated bound pairs

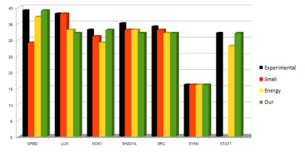


Fig: Compared tools for the prediction of the literature based unbound pairs

Discussion:

Research has shown that the similar domains interact with similar substrates but dissimilar domains don't necessarily have dissimilar specificities. Here, we created a model for predicting the binding partners of almost all human SH2 domains based on amino acids position pairs of SH2 domains and respective phopho-peptides. The Tool is easy to use and significantly fast.

Fig: Dendrogram of the Human Complement of SH2 Domains. (Bernard et al., Mol. Cell, 2006)

ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

Kousik Kundu is funded by the Centre for Biological Signalling Studies (BIOSS), Freiburg, Germany.