Identification of SH2-peptide interactions using support vector machine

Kousik Kundu[1,2], Fabrizio Costa[1], Michael Huber[3] , Michael Reth[2,4], 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 Biochemistry and Molecular Immunology, Institute of Biochemistry and Molecular Biology, University Clinic, RWTH Aachen University, Pauwelsstr. 30, Aachen D-52074
[4]Department of Molecular Immunology, Max Planck-Institute of Immunology, Freiburg D-79108

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 by mediating the correct formation of protein complexes. 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 for 51 human SH2 domains, which shall be made publically available.

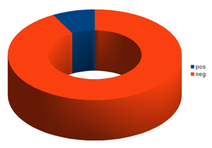
Materials:

- **The data are taken from microarray experiments and peptide array library**
  - **Microarray Data:**

Dataset I:

SH2 domains: 51
Peptides: 20
Possible interactions 51 X 20 = 1020

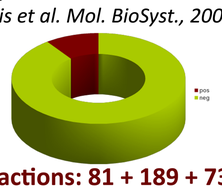
Positive interactions: 81
Negative interactions: 1020 - 81 = 939
Jones et al. Nature, 2006



Dataset II:

SH2 domains: 46
Peptides: 41
Possible interactions 46 X 41 = 1886

Positive interactions: 189
Negative interactions: 1886 - 189 = 1697
Alexis et al. Mol. BioSyst., 2008

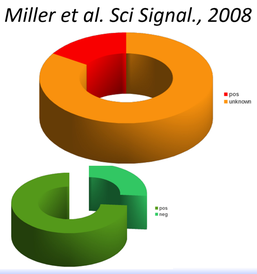


  - **Peptide Array Library Data:**

Dataset III:

SH2 domains: 51
Peptides: 906

Positive interactions: 7395
Unknown interactions: 38811
Miller et al. Sci Signal., 2008

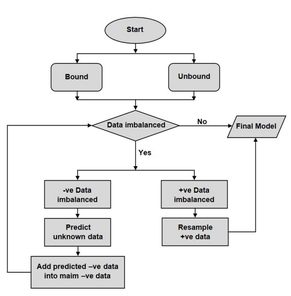


- **Total interactions**

Positive interactions: 81 + 189 + 7395 = 7665
Negative interactions: 939 + 1697 = 2636

Data imbalanced problem and features:

- **Preparation of dataset**



- **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      DPKFHYDNTAG
SRC      TGNHTYQEIAV
```

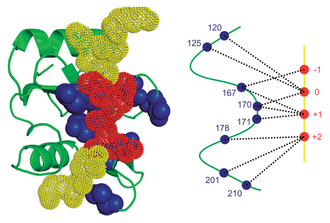


Fig: PDB structure of a SH2-peptide complex and the interactionmap. (zeba at al., NAR, 2009)

- **Important amino acids in peptides:**
  Research has shown the neighbor residues (-2 to +4) of the pY, are the most important for domain-peptide binding specificities. Therefore.

- **Total features: 20 x 7 = 140**

Classification:

- **Main classifier used:**

  Support Vector Machine (SVM)

  SVM is a machine learning approach based on statistical learning theory.

- **SVM learning implementation softwares:**
  - **WEKA**
  - **SVM light**

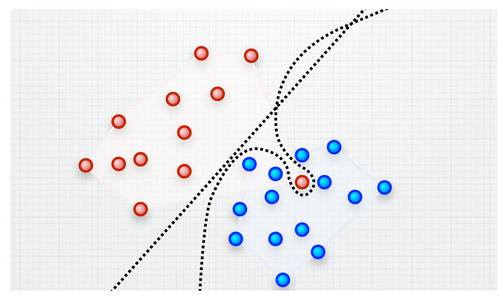


Fig: Example of linear and polynomial kernels

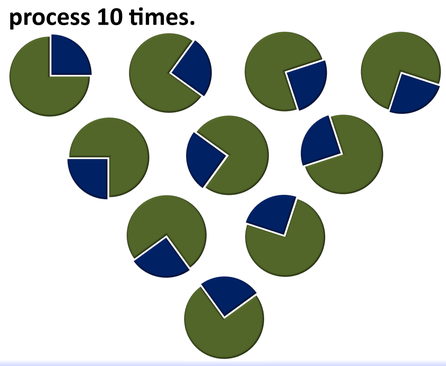
We used different degrees with different C values to get best results for each SH2 model.

Best performances:
Kernel : Linear, Polynomial
Degree : 1, 2
C value: 1, 10

Results:

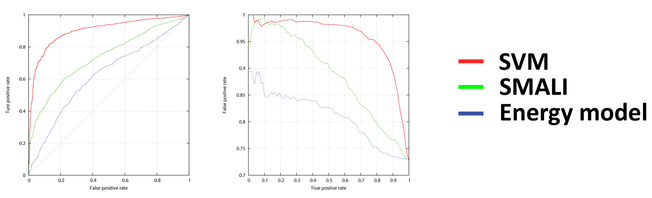
- **Making Training and Test sets**

The non-redundant dataset is partitioned into 2 parts. 75% data for training set and 25% data for test set and we repeated the process 10 times.



- **Result and comparison with other tool on the test-set.**

The models have been tested on the respective test-sets. There were 51 models for each SH2 domains and then we compared the performance with the other tools.



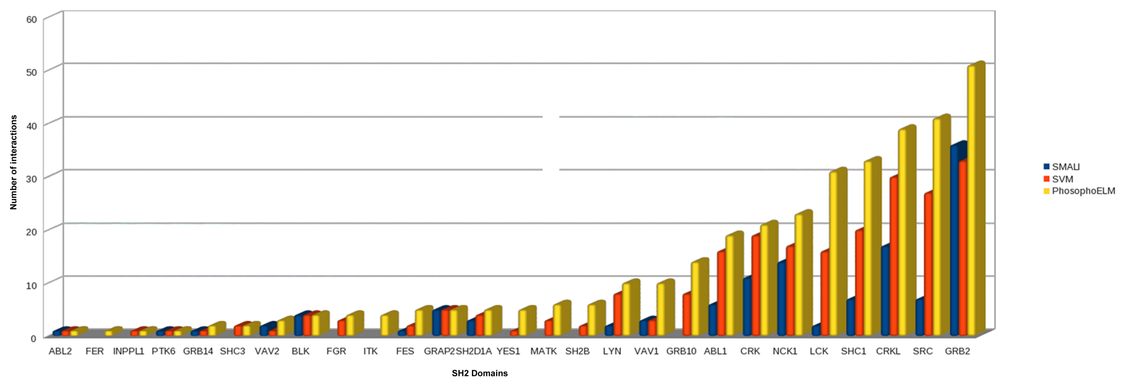
ROC:
SVM- 0.90
SMALI- 0.73
Energy model- 0.63

PR:
SVM- 0.96
SMALI- 0.89
Energy model- 0.81

Figs: The ROC curves and PR for the test set (25%).

Comparison with other tools on in-vivo data

- **Our tool has been compared with SMALI on reliable PhosphoELM data, which clearly showed that our tool has higher accuracy.**



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 51 human SH2 domains based on amino acids position on the respective phospho-peptides. The tool is easy to use and significantly fast.

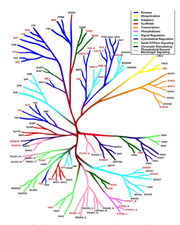


Fig: Dendrogram of the Human Complement of SH2 Domains. (Bernard et al., Mol. Cell, 2006)

BIOSS is the Cluster of Excellence of Albert-Ludwigs-University Freiburg. Funded by the German Research Foundation (DFG)