An Efficient Semi-Supervised Learning Approach to Predict SH2 domain Mediated Interactions

Kousik Kundu [1,†] **and Rolf Backofen** [1,2]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany
[2]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany

† Present address: Department of Human Genetics, The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK and Department of Haematology, University of Cambridge, Cambridge, UK

Summary

Src homology 2 (SH2) domain is an important subclass of modular protein domains that plays an indispensable role in several biological processes in eukaryotes. SH2 domains specifically bind to the phosphotyrosine residue of their binding peptides to facilitate various molecular functions. For determining the subtle binding specificities of SH2 domains, it is very important to understand the intriguing mechanisms by which these domains recognize their target peptides in a complex cellular environment. There are several attempts have been made to predict SH2-peptide interactions using high-throughput data. However, these high-throughput data are often affected from a low signal to noise ratio. Furthermore, the prediction methods have several additional shortcomings, such as linearity problem, high computational complexity etc. Thus, computational identification of SH2-peptide interactions using high-throughput data remains challenging. Here, we propose a machine learning approach based on an efficient semi-supervised learning technique for the prediction of 51 SH2 domain mediated interactions in human proteome. In our study, we have successfully employed several strategies to tackle the major problems in computational identification of SH2-peptide interactions.

Key words

Src homology 2 domain, Signal transduction, Protein-protein interaction, Phosphotyrosine peptides, Support vector machine, Semi-supervised learning.

1 Introduction

In 1986, Tony Pawson and co-workers first discovered the Src homology 2 (SH2) domain from the oncogenic v-FPS/FES cytoplasmic tyrosine kinase encoded in the Fujinami sarcoma virus [1]. Since then a number of SH2 domains have been identified in several eukaryotic species [2, 3]. Although SH2 domains are found across the eukaryotes, they are more abundant in metazoans [4, 5]. Currently, 122 SH2 domains from 112 unique human proteins have been reported in the `UniProtKB/Swiss-Prot` database, release 2015-06 [6]. SH2 domains are approximately 100 amino acids in length, and are structurally conserved domains that contain a central β sheet and

two α helices [7]. These domains are known to mediate intracellular signaling pathways by specifically recognizing short linear phosphotyrosine (pY) containing peptides [8]. Although SH2 domains mainly target phosphotyrosine (pY) residue of the binding peptides, their binding specificity is determined by the neighbor residues of the pY, particularly from -2 to $+4$ (pY at 0^{th} position) [8, 9]. For example, a Leu or Pro residue at position $+3$ (xx-pY-xx[L/P]x) is strongly preferred by CRK SH2 domains, where x represents any naturally occurring amino acid. Alternatively, a hydrophobic residue (Φ) at position -2 (Φx-pY-xxxx) is preferred by PTPN11 SH2 domains [10]. Previous studies showed that the mutations in some SH2 domains can cause several human diseases, such as XLP syndrome [11], X-linked α-gammaglobulinemia [12], Noonan syndrome [13] etc.

In recent years, various high-throughput techniques, such as peptide array, microarray etc., have been introduced to define the binding specificities of SH2 domains. The enormous amount of data generated by these techniques are invaluable to build efficient computational methods. However, these data are often affected by false positive and false negative interactions. Most of the popular computational methods, that use high-throughput data as their training sets, are based on the position weight matrices (PWMs), which do not consider the dependencies among the amino acids in the peptide sequences [14, 15]. Here, we present a machine learning algorithm to build non-linear models that can exploit the dependencies between the amino acids in the binding peptides. In addition, these PWM-based models are essentially generative models, as they rely only on the information on positive interaction data and completely ignore the information on negative interaction data; whereas machine learning methods rely on both positive and negative interaction data and produce discriminative models, which have advantages over generative models [16, 17].

One of the major problems of using high-throughput data is, in general, the available information on positive interactions is much higher than the negative interactions, which lead to a severe data imbalance problem. For example, the positive interaction data can be up to 15 times more abundant than negative interaction data for an SH2 domain [18]. In the machine learning literature, it is known that the severely imbalanced class distribution in a training set negatively affects the performance of the predictive model; generally, these models are biased towards the majority class. To mitigate

this problem, we employed an efficient semi-supervised technique where the *self-training* strategy was used to balance the training sets. Therefore, as a consequence, we achieved powerful discriminative models. Finally, we performed a genome-wide prediction of the SH2 domain mediated interactions in human proteome to uncover the biologically relevant interactions. The prediction tool (`SH2PepInt`) has been implemented in a newly developed web server, namely `MoDPepInt` [19].

2 Materials

2.1 Dataset

In our study, all the high-throughput interaction data were obtained mainly from three sources; one high-density peptide array [20] and two protein microarray experiments [21, 22]. Additionally, we extracted interaction data from a manually curated high-quality `PhosphoELM` database [23] for evaluating our models. To unveil novel interactions by genome-wide prediction, we extracted all the tyrosine containing proteins from `UniProtKB/Swiss-Prot` database [6].

1. **Peptide array data:** The observed binding interaction data in the peptide array experiment was deposited in the `NetPhorest` database [20]. From `NetPhorest` database, a total 14678 positive interactions that involved 61 SH2 domains and 920 peptides were retrieved. After removing all the redundancy, we finally obtained 7544 positive interactions (Dataset I).

2. **Microarray data:** We incorporated the interaction data from two protein microarray experiments [21, 22]. From the microarray experiment in [21], we retrieved 2100 interactions with 160 positive and 1940 (2100 − 160) negative interactions, which involved 115 SH2 domains, and 20 peptides from ErbB2 and ErbB3 receptor proteins (Dataset II). Note that in Dataset II, we did not consider the interactions related to the ErbB1 receptor protein (see **Note 1**).

   From other microarray experiment in [22], we considered 3485 interactions with 314 positive and 3171 (3485 − 314) negative interactions, involving 85 SH2 domains, and 41 singly phosphorylated peptides from EGFR, FGFR, and IGIFR receptor proteins (Dataset III).

3. **Manually curated data:** For evaluating our models, we retrieved the binding information of SH2 domains from a manually curated database, called `PhosphoELM` database. We have extracted a total 878 binding interactions that involved 63 SH2 domains and 359 peptides (Dataset IV).

4. **Genome-wide prediction data:** All the human protein sequences were derived from `UniProtKB/Swiss-Prot` database [6]. A total 20225 proteins, which contain 298 637 tyrosine residues, were considered. Finally, a total 298 637 tyrosine containing peptides were generated as our test set for genome-wide prediction.

2.2 Data compilation

We have combined all the high-throughput data derived from peptide array and microarray experiments together, but surprisingly, we found there were several disagreements on the binding information between different experiments. Furthermore, these high-throughput data are often affected by a high rate of false positive and false negative interactions. The refinement of these noisy data from our training set is explained below.

1. In the two microarray experiments, i.e., Dataset II and Dataset III, there were 10 proteins that contained two SH2 domains (N and C terminal) each. Since these datasets do not report the assignment of which peptide specifically binds with which of the two SH2 domains of a protein, we discarded all the interactions related to these double-SH2 domain containing proteins.

2. We combined these two microarray data, and collected a total 474 (160 + 314) positive and 5111 (1940 + 3171) negative interactions. The apparent equilibrium disassociation constant (K_D value) or affinity constant was applied to determine the positive and negative interaction classes. We used the same K_D cutoff as mentioned in [21, 22], and thus the SH2-peptide interactions with K_D values less than 2 μM were considered as binding (positive data) interactions, while remaining interactions were considered as non-binding (negative data) interactions. Nevertheless, various inconsistencies in the SH2-ErbB1 (ErbB1 or EGFR protein was common in both experiments) interactions were observed between Dataset II and Dataset III (see **Note 1**).

3. We could only consider 7544 positive data from the peptide array experiment [20], since there was evidence only for positive interactions (see **Note 2**). Surprisingly, we also observed 149 interactions for which there is a conflict between the peptide array experiment [20] and the microarray array experiments [21, 22], i.e., these interactions are positive in Dataset I, but negative in Dataset II and Dataset III. To reduce the noisy and conflicting information from our training sets, we discarded these 149 interactions. Therefore, as a consequence, the positive data in Dataset I was reduced to 7395 (7544 − 149), and the negative data in Dataset II and Dataset III was reduced to 4962 (5111 − 149).

4. Among the 474 positive interactions collected from microarray experiments, 247 interactions were already present in Dataset I. After removing the redundancy, we comprised 227 (474 − 247) positive interactions from Dataset II and Dataset III. These non-redundant positive interactions were kept for the validation.

5. We only considered those SH2 domains that have at least 40 positive interaction data, otherwise no complex model can be reliably fit. We used positive interaction data from Dataset I and the negative interaction data from Dataset II and Dataset III, and finally, we composed our training sets with 6742 positive interactions and 2523 negative interactions for 51 human SH2 domains.

3 Methods

Here, we present a machine learning method to produce non-linear models, which can exploit the inter-dependencies between the amino acids in the binding peptides. Additionally, we present a semi-supervised technique that can efficiently refine the high-quality negative interactions from a noisy dataset.

3.1 Feature encoding

1. Although SH2 domains specifically bind to the phosphotyrosine (pY) residue of their binding peptides, the neighbor residues of the pY are known to be highly predictive for domain-peptide interactions [14, 15].

2. We restricted the peptide sequences to 7 amino acids in length, namely we extracted the amino acids in position from -2 to $+4$ where the pY holds the 0^{th} position.

3. In the feature encoding, we did not include the central residue (pY), since it was always same in the peptides from both classes (positive and negative) and thus not holding any discriminative information. Therefore, a peptide sequence was mapped into a binary vector x living in a $120 \times 6 = 120$ dimensional space, i.e., for each position, we reserved 20 dimensions for each amino acid, and encoded the amino acid type with a 1 in the corresponding dimension and 0 elsewhere.

4. A data set for each domain D_j was compiled, which was encoded as a set of pairs $(x_1, c_1), .., (x_n, c_n)$ where x_i is the binary feature vector for peptide P_i with the class label $c_i \in \{-1, 1\}$. If the domain D_j interacts with the peptide P_i, then the correspondence class label is 1, otherwise in case of non-interaction, the class label is -1.

3.2 Data modeling

Previous research showed that the contextual dependencies between the amino acids in the cognate peptide sequences are highly important to describe the binding specificities of SH2 domains (see **Note 3**) [24]. Any methods that ignore these kind of dependencies often produce sub-optimal models (see **Note 4**). Therefore, in order to build the predictive models, which allow the dependencies between the amino acids in the binding peptides, a polynomial kernel support vector machine (SVM) has been employed. We have used the SVM implementation in C language provided in SVM^{light} [25].

1. A polynomial kernel is a kernel function that computes the similarity between training samples (vectors) in the polynomial feature space to learn a non-linear model. The polynomial kernel function for two vectors: X and X' with degree d is defined in [26] as:

$$K(X, X') = (1 + \langle X, X' \rangle)^d, \tag{1}$$

where "1" is a constant, which is required to consider the effects of all degrees that are less than d. A feature space with two inputs: $X1$ and $X2$, and $d = 2$ (see **Note 5**) is, therefore, defined as:

$$\begin{aligned} K(X, X') &= (1 + \langle X, X' \rangle)^2 \\ &= (1 + X_1 X_1' + X_2 X_2')^2 \\ &= 1 + 2X_1 X_1' + 2X_2 X_2' + (X_1 X_1')^2 + (X_2 X_2')^2 + 2X_1 X_1' X_2 X_2'. \end{aligned} \tag{2}$$

2. One of the main hyper-parameters in SVM is the cost parameter or C, which is used to trade off generalization of data fitting. Basically, it provides some flexibility in an enlarged feature space for data separation.

3. All the model parameters, i.e., $d \in \{1, 2, 3\}$ and $C \in \{0.01, 0.1, 1, 10\}$, have been optimized on training sets under a *stratified* cross-validation setting (see **Note 6**).

4. The optimization of these hyper-parameters is important to counter balance the overfitting phenomena (see **Note 7**). More specifically, for each model, the best parameter combination was chosen on a held out data set (validation set). However, the model performance was evaluated on a separate test set, which was never seen in the validation or training phase.

3.3 Semi-supervised negative data

Data imbalance is one of the major problems in high-throughput experiments where availability of the data from one class is much higher than the other class (see **Note 8**). To deal with this problem, we employed a semi-supervised learning (SSL) approach (see **Note 9**).

1. We resorted to the *self-training* strategy (see **Note 10**), although there are several strategies available to tackle the SSL problem.

2. For each domain, the initial high-throughput data was extracted from high density peptide array and microarray experiments (see dataset compilation) to train the base classifiers.

3. As an unlabeled data set, we considered the SH2-peptide pairs that did not show any positive signals in the peptide array experiment (see **Note 11**). Note that randomly generated negative peptides were not considered (see **Note 12**).

4. For each domain, a polynomial SVM was used to predict confidence negative instances from the unlabeled data set, and iteratively added them to the main training set until the data set was balanced (see **Note 13**). Here, the confidence of negative data was scored as the distance from the hyperplane.

5. Finally, the model selection process was performed to select the best model complexity for each specific SH2 domain. Fifty one models were built for 51 domains.

6. The flowchart for iterative balancing technique for positive and negative data is depicted in Figure 1.

3.4 Predictive performance

1. For measuring the predictive performance, we computed 5 measures, i.e., sensitivity, specificity, precision, area under the receiver operating characteristics curve (AUC ROC), and area under the precision recall curve (AUR PR) (see **Note 14**).

2. Two different strategies were taken to evaluate the predictive performance of our models: (i) a *stratified* 5-fold cross-validation and (ii) we randomly split the data where we considered 75% as training set and 25% as test set; we repeat the process 10 times to create 10 train/test data sets.

3. We compared our methods with PWM-based `SMALI` approach [15] and the energy model [27], and in 5-fold cross-validation setting, we achieved an average AUC ROC of 0.83 and an average AUC PR of 0.93, which outperformed other two methods; `SMALI` and the energy model achieved an average AUC ROC of 0.71 and 0.62, respectively, and an average AUC PR of 0.87 and 0.81, respectively [18].

4. In order to achieve same specificity as `SMALI` (0.95 on average), we identified our threshold accordingly, and as a consequence, our models achieved an average sensitivity of 0.45, outperforming `SMALI` and the energy model, which achieved an average sensitivity of 0.26 and 0.17, respectively [18].

5. For the validation purpose, we evaluated our models on a manually curated and highly reliable data set, namely `PhosphoELM` (Dataset IV). Our models achieved a true positive rate (TPR) of 0.64, which is much better than the TPR of 0.33 achieved by `SMALI` [18]. Note that the comparison with energy model could not be possible, as the class determination threshold was not specified in [27].

6. Our method efficiently predicts the binding partners of most of the SH2 domains, however, it might get challenged for some SH2 domains whose training data are very small or have *within-class imbalance* problem (see **Note 15**).

3.5 Genome-wide prediction

It is always interesting to see the interactions that are novel and biologically relevant. In order to uncover such interactions, we performed a genome-wide prediction. Subsequently, a term-centric enrichment analysis was performed to unveil novel functionalities of the predicted interactions.

1. All the generated peptides were restricted to 7 amino acids in length, i.e., -2 to $+4$ amino acids with Tyr at 0^{th} position.

2. We used our models to predict the binding partners of all 51 human SH2 domains.

3. All the predicted interactions were filtered based on some criteria to achieve more confidence interactions. We have used two filters: (i) phosphotyrosine (pY) (see **Note 16**) and (ii) co-cellular localization (see **Note 17**).

4. For each SH2 domain, we considered top 50 predictions and performed a *term-centric* enrichment analysis using `DAVID` tool [28] to unveil the novel and biologically relevant interactions (see **Note 18**). By doing this, several biologically meaningful interactions were observed [18].

5. All the top predictions and their *term-centric* analysis for all 51 human SH2 domains are available under the URL: `http://www.bioinf.uni-freiburg.de/Software/SH2PepInt/Genome-wide-predictions.tar.gz`

3.6 MoDPepInt web-server

We implemented our prediction tool (SH2PepInt) for SH2-peptide interactions into a newly developed web server, called MoDPepInt (Modular Domain Peptide Interaction) [18, 19]. Currently, the MoDPepInt web server offers three different tools: (i) SH2PepInt, (ii) SH3PepInt, and (iii) PDZPepInt for predicting the binding interactions of three different modular domains, SH2, SH3, and PDZ, respectively [18, 17, 29].

1. The web server has two different modes: (i) basic mode and (ii) expert mode. We designed a meta-web server for the basic mode, where only the input is required. The input is submitted simultaneously to all tools, and a summary table is produced. The expert mode is more flexible, where user can choose the SH2 domains of interest and use desired filters to obtain high confident interactions.

2. For SH2-peptide interactions, two filters have been used in order to increase the prediction accuracy. The filters are: (i) phosphotyrosine and (ii) cellular localization (see **Note 16** and **17**).

3. The MoDPepInt server is available under the URL: http://modpepint.informatik.uni-freiburg.de/SH2PepInt/Input.jsp

4 Notes

1. Eleven peptides from ErbB1 proteins were used in both microarray experiments [21, 22]. We retrieved the interaction data, which involved those 11 peptides and 85 SH2 domains (also common in both microarray experiments). Interestingly, we observed there were severe inconsistencies in the interaction data produced by these two microarray experiments, as in similar settings, one microarray experiment [21] showed positive signals ($K_D < 2\ \mu M$) for 32 interactions, whereas other microarray experiment [22] showed positive signals for 120 interactions (see Figure 2). All the 32 positive interactions observed in [21] were also observed in [22]. The possible reason for the lack of interactions in [21], could be caused due to low concentration of proteins on the surface of the slides, which can happen when the protein printing tip is slightly mis-aligned. In this case, there is just not enough pro-

tein present on the surface of the slide that can cross the background threshold, even though there is, however, a tight interaction.

2. From peptide array experiment, we retrieved the interaction information for 61 domains and 920 peptides. Thus, the possible domain-peptide interactions should be $920 \times 61 = 56120$. Among them, a total 7544 interactions showed the binding signals in the experiment, and remaining 48576 ($56120 - 7544$) interactions did not show any signals. However, one should not assume all these 48576 interactions as non-binding type, since the binding signal could not be observed may be due to some experimental stringencies. For example, an interaction might be considered as non-binding, if the binding threshold value is less than the detection limit, even though there is, however, a weak interaction. Poor domain folding and/or peptide synthesis problem could also be the reason for these kind of error (false negative) in the data. Thus, refinement of these false negative interactions is vital to build a good predictive model.

3. In 2010, Liu *et al.* published an interesting study where they explained the importance of the contextual dependencies between the amino acids in the binding peptides [24]. More precisely, the binding peptides are composed of permissive and non-permissive amino acids where permissive amino acids allow the interaction and non-permissive amino acids inhibit the interactions [24, 9]. It is known that CRK SH2 domain interacts peptides where Leu or Pro amino acid is present at +3 (pY at 0^{th} position), but surprisingly, presence of other amino acids in other positions can also influence the binding specificity. For example, basic residues (Arg and His) are disfavored at position +1 and +2, whereas Pro is prohibited at position +1. More interestingly, the acceptance of Ala at +1 completely depends on the amino acid at +3. Ala is accepted only if there is a Pro at +3, where as it is rejected, in case, there is a Leu at +3 [9].

4. All the approaches that based on position weight matrices (PWMs) [14, 15] and linear machine learning models [30] do not consider the positional dependencies between the amino acids in the binding peptides, and therefore can not accurately determine the binding specificities of SH2 domains and eventually, produce sub-optimal results.

5. As $d = 1$ only provides linear model, we used the $d > 1$ for building non-linear models. However, the degree of the polynomial kernel is optimized via cross-validation, and therefore a simpler linear model ($d = 1$) can still be chosen for some SH2 domains when it offers better performance.

6. In a cross-validation setting, a stratification procedure is used to maintain approximately the same proportion of the two types of class labels, i.e., positive and negative, in each fold. Cross-validation with stratification procedure is known as *stratified* cross-validation.

7. Overfitting is a common problem in machine learning methods. It normally occurs when the machine learning algorithms capture the noise of the data. If the model fits too well to the data, it causes overfitting, and eventually, produce sub-optimal predictive model. Unfortunately, this important aspect is often ignored in the bioinformatics prediction methods. To overcome this overfitting problem, we used an appropriate technique, called *regularization*. The regularized predictor is more robust to noise, and guarantees better prediction quality on unseen data. Although there are several ways to counter balance the overfitting issue, we adopted an efficient strategy where we minimized the model complexity by tuning the degree of the polynomial (d) and the cost parameter (C).

8. It is known that machine learning algorithms work poorly on highly imbalanced data, and negatively affect the performance of adaptive predictors [31]. These algorithms are generally biased towards the majority class, and hence often produce poor discriminative models.

9. In semi-supervised learning, a small amount of labeled data and a large amount of unlabeled data are trained. Note that for using the small amount of labeled data, a strong model assumptions need to be made. It is very important step as if the model assumptions do not match the nature of the problem, then it would be critical for the predictive performance. There are several techniques, such as *expectation maximization (EM)*, *co-training*, *self-training*, and *graph-based* methods, have been developed to handle the SSL problem. Each technique is used based on the requirements of the problems.

10. The *self-training* strategy relies only on the good discriminative properties of the base classifier. This is a simple wrapper method, which iteratively uses the initial labeled data to train the classifier, which then assigns a label to the remaining previously unlabeled data. In our application, this is the most suited strategy that can efficiently tackle the semi-supervised problem. Note that this approach is only applicable when at least a few confidence positive and negative data available to train the base classifier.

11. In theory, if an SH2 domain does not show a positive signal for a peptide in a peptide array experiment, the SH2 domain is considered to be a non-binder to that particular peptide, and the SH2-peptide pair is believed as a negative interaction pair. However, it is known that high-throughput data are highly affected by false negative interactions. Therefore, in order to filter high-confidence negative, we applied *self-training* strategy.

12. In common practice, random peptides are used to generate artificial negative instances. However, previous research showed that the randomly generated instances significantly decrease the prediction quality of a model [32]. Hence, instead of taking random peptides, we used experimental data.

13. For some domains, negative data was already much higher than the positive data in the base classifier. In these cases, we used a *rebalancing* technique where we over-sampled the positive class to balance the base classifier. Note that we did not under-sample of the positive class in order not to throw away the valuable information.

14. One major problem in machine learning is that the mainstream algorithms are not designed to efficiently deal with the skewed class distribution; these algorithms are more accurate only on the majority class. For example, if a data set is imbalanced, containing a few positive and many negative data, a rational choice based on maximizing the predicted accuracy (in an equal cost scenario) would most certainly be biased towards the majority class, and therefore the predictive model will almost always predict a negative response. Hence, in a binary classification, a single standard statistical measurement

(e.g., accuracy, AUC ROC) will not be appropriate and can mislead the predictive performance. In this case, the model will achieve high specificity, precision, and AUC ROC but very poor sensitivity and AUC PR. Thus, it is always important to show multiple statistical measures to describe the performance of a predictive model.

15. The *within-class imbalance* and the *small-disjuncts* problem typically occur when the class concept is composed by many sub-clusters/sub-concepts, and each of the sub-cluster represented by a very small number of examples. If a small sub-cluster in training set is over-represented in the test set, our models sometimes might fail to identify those interactions. However, we could tackled this problem for the negative data, as we could select many peptides from different array experiments for which no definitive interaction information was available. Unfortunately, the information for the positive interactions was very limited for some SH2 domains, and therefore the total number of positive interactions were very few in the training set for those SH2 domains. Since this is a standard problem in machine learning, some oversampling techniques (e.g., SMOTE) have been proposed in the literature to tackle this problem, however, they have several drawbacks, e.g., requiring an explicit instance representation [18, 33].

16. This filter was used to get all the tyrosine containing peptides whose phosphorylation evidence was experimentally verified. At the time of analysis, the phosphorylation evidence of a total 30 228 peptides from 10 688 proteins was available in the `PhosphoSitePlus` database [34]. Note that we ignored some phosphopeptide containing proteins that were not present in the `UniProtKB/Swiss-Prot` database. Finally, a total number of 27 481 phosphopeptide from 9621 human proteins were used. Since SH2 domains are known to interact with phosphopeptides, this filter will provide more probable interactions.

17. It is highly unlikely to see an SH2-peptide interaction where the SH2 domain and the peptide containing protein reside different compartments of the cell. To filter out all these kind of unlikely interactions, we implemented a co-cellular localization filter. In this setting, we only considered an SH2-peptide interaction, if the SH2 domain containing protein and the peptide containing protein share at least one `GO`-term

that is annotated in the `Gene Ontology (GO)` database [35].

18. The `DAVID` [28] tool allows to perform a *term-centric* enrichment analysis on more than 40 different annotation categories, and reports enriched annotation terms associated with the predicted proteins. p-value is used to determine the enrichment; the smaller p-values indicate higher enrichment.

Acknowledgement

This chapter is based on our previous publication [18]. This work was funded by Bundesministerium für Bildung und Forschung (e-bio; FKZ 0316174A to RB), and the Centre for Biological Signalling Studies (BIOSS), University of Freiburg.

References

[1] Sadowski I, Stone J. C, and Pawson T, (1986), A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps. Mol Cell Biol, vol. 6, no. 12, pp. 4396–408.

[2] Mayer B. J, Hamaguchi M, and Hanafusa H, (1988), A novel viral oncogene with structural similarity to phospholipase C. Nature, vol. 332, no. 6161, pp. 272–5.

[3] Anderson D, Koch C. A, Grey L, Ellis C, Moran M. F, and Pawson T, (1990), Binding of SH2 domains of phospholipase C gamma 1, GAP, and Src to activated growth factor receptors. Science, vol. 250, no. 4983, pp. 979–82.

[4] Lim W. A and Pawson T, (2010), Phosphotyrosine signaling: evolving a new cellular communication system. Cell, vol. 142, no. 5, pp. 661–7.

[5] Liu B. A, Shah E, Jablonowski K, Stergachis A, Engelmann B, and Nash P. D, (2011), The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes. Sci Signal, vol. 4, no. 202, p. ra83.

[6] Magrane M and Consortium U, (2011), UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford), vol. 2011, p. bar009.

[7] Waksman G, Kominos D, Robertson S. C, Pant N, Baltimore D, Birge R. B, Cowburn D, Hanafusa H, Mayer B. J, Overduin M, Resh M. D, Rios C. B, Silverman L, and Kuriyan J, (1992), Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. Nature, vol. 358, no. 6388, pp. 646–53.

[8] Pawson T, (2004), Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. Cell, vol. 116, no. 2, pp. 191–203.

[9] Liu B. A, Engelmann B. W, and Nash P. D, (2012), The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction. FEBS Lett.

[10] Imhof D, Wavreille A.-S, May A, Zacharias M, Tridandapani S, and Pei D, (2006), Sequence specificity of SHP-1 and SHP-2 Src homology 2 domains. Critical roles of residues beyond the pY+3 position. Journal of Biological Chemistry, vol. 281, no. 29, pp. 20271–82.

[11] Sayos J, Wu C, Morra M, Wang N, Zhang X, Allen D, van Schaik S, Notarangelo L, Geha R, Roncarolo M. G, Oettgen H, De Vries J. E, Aversa G, and Terhorst C, (1998), The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. Nature, vol. 395, no. 6701, pp. 462–9.

[12] Tzeng S. R, Pai M. T, Lung F. D, Wu C. W, Roller P. P, Lei B, Wei C. J, Tu S. C, Chen S. H, Soong W. J, and Cheng J. W, (2000), Stability and peptide binding specificity of Btk SH2 domain: molecular basis for X-linked agammaglobulinemia. Protein Sci, vol. 9, no. 12, pp. 2377–85.

[13] Tartaglia M, Mehler E. L, Goldberg R, Zampino G, Brunner H. G, Kremer H, van der Burgt I, Crosby A. H, Ion A, Jeffery S, Kalidas K, Patton M. A, Kucherlapati R. S, and Gelb B. D, (2001), Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat Genet, vol. 29, no. 4, pp. 465–8.

[14] Obenauer J. C, Cantley L. C, and Yaffe M. B, (2003), Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res, vol. 31, no. 13, pp. 3635–41.

[15] Li L, Wu C, Huang H, Zhang K, Gan J, and Li S. S.-C, (2008), Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. Nucleic Acids Res, vol. 36, no. 10, pp. 3263–73.

[16] Ng A. Y and Jordan M. I, On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. in NIPS, pp. 841–848, (2001).

[17] Kundu K, Costa F, and Backofen R, (2013), A graph kernel approach for alignment-free domain-peptide interaction prediction with an application to human SH3 domains. Bioinformatics, vol. 29, no. 13, pp. i335–i343.

[18] Kundu K, Costa F, Huber M, Reth M, and Backofen R, (2013), Semi-Supervised Prediction of SH2-Peptide Interactions from Imbalanced High-Throughput Data. PLoS One, vol. 8, no. 5, p. e62732.

[19] Kundu K, Mann M, Costa F, and Backofen R, (2014), MoDPepInt: an interactive web server for prediction of modular domain-peptide interactions. Bioinformatics, vol. 30, no. 18, pp. 2668–2669.

[20] Miller M. L, Jensen L. J, Diella F, Jorgensen C, Tinti M, Li L, Hsiung M, Parker S. A, Bordeaux J, Sicheritz-Ponten T, Olhovsky M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk B. E, Yaffe M. B, Brunak S, and Linding R, (2008), Linear motif atlas for phosphorylation-dependent signaling. Sci Signal, vol. 1, no. 35, p. ra2.

[21] Jones R. B, Gordus A, Krall J. A, and MacBeath G, (2006), A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature, vol. 439, no. 7073, pp. 168–74.

[22] Kaushansky A, Gordus A, Chang B, Rush J, and MacBeath G, (2008), A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R. Mol Biosyst, vol. 4, no. 6, pp. 643–53.

[23] Diella F, Gould C. M, Chica C, Via A, and Gibson T. J, (2008), Phospho.ELM: a database of phosphorylation sites–update 2008. Nucleic Acids Res, vol. 36, no. Database issue, pp. D240–4.

[24] Liu B. A, Jablonowski K, Shah E. E, Engelmann B. W, Jones R. B, and Nash P. D, (2010), SH2 domains recognize contextual peptide sequence information to determine selectivity. Mol Cell Proteomics, vol. 9, no. 11, pp. 2391–404.

[25] Joachims T, (1999), Making large-scale SVM learning practical, in Advanced in Kernel Methods-Support Vector Learning (ikopf, B., Burges, C., Smola, A., eds). MIT Press, Cambridge, MA.

[26] Hastie T, Tibshirani R, and Friedman J, (2008), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, second ed.

[27] Wunderlich Z and Mirny L. A, (2009), Using genome-wide measurements for computational prediction of SH2-peptide interactions. Nucleic Acids Res, vol. 37, no. 14, pp. 4629–41.

[28] Huang D. W, Sherman B. T, and Lempicki R. A, (2009), Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, vol. 37, no. 1, pp. 1–13.

[29] Kundu K and Backofen R, (2014), Cluster based prediction of PDZ-peptide interactions. BMC Genomics, vol. 15, no. Suppl 1, p. S5.

[30] Li L, Zhao B, Du J, Zhang K, Ling C. X, and Li S. S.-C, (2011), DomPep–a general method for predicting modular domain-mediated protein-protein interactions. PLoS One, vol. 6, no. 10, p. e25528.

[31] He H and Garcia E. A, (2009), Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp. 1263–1284.

[32] Ben-Hur A and Noble W. S, (2006), Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics, vol. 7 Suppl 1, p. S2.

[33] Chawla N, Bowyer K, Hall L, and Kegelmeyer W. P, (2002), Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, vol. 16, pp. 321–357.

[34] Hornbeck P. V, Kornhauser J. M, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, and Sullivan M, (2012), PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res, vol. 40, no. Database issue, pp. D261–70.

[35] Ashburner M, Ball C. A, Blake J. A, Botstein D, Butler H, Cherry J. M, Davis A. P, Dolinski K, Dwight S. S, Eppig J. T, Harris M. A, Hill D. P, Issel-Tarver L, Kasarskis A, Lewis S, Matese J. C, Richardson J. E, Ringwald M, Rubin G. M, and Sherlock G, (2000), Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, vol. 25, no. 1, pp. 25–9.

Figure legends

Figure 1: For the imbalanced data sets, we encountered two types of problems: (i) for most of the domains, the positive data was much higher than the negative data and (ii) for some domains, the different scenario was occurred when the negative data was higher than the positive ones. To solve the first problem, we used a *self-training* strategy to predict confidence negative interaction data. The process was iteratively done until a balanced data set was reached (left branch). To solve the second problem, we applied a *rebalancing* technique and over-sampled the positive class (right branch). This figure is adapted from [18].

Figure 2: Comparison of the outcome of two different microarray experiments. Eleven peptide sequences from ErbB1 protein and 85 SH2 domains were considered. The green bars indicate the number of SH2-peptide interactions observed in [21], and the red bars indicates the number of SH2-peptide interactions observed in [22]. This figure clearly shows that almost 4 times more interactions were observed in Dataset III (120 interactions observed) in comparison to Dataset II (32 interactions observed).