BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

MoDPepInt: An interactive webserver for prediction of modular domain-peptide interactions

Kousik Kundu [1,†], Martin Mann [1,†], Fabrizio Costa [1] and Rolf Backofen [1,2,3,4 *]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany, [2]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany, [3]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany, [4]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: MoDPepInt (Modular Domain Peptide Interaction) is a new, easy-to-use webserver for the prediction of binding partners for modular protein domains. Currently we offer models for SH2, SH3 and PDZ domains via the tools SH2PepInt, SH3PepInt and PDZPepInt. More specifically our server offers predictions for 51 SH2 human domains and 69 SH3 human domains via single domain models, and predictions for 226 PDZ domains across several species, via 43 multi-domain models. All models are based on support vector machines with different kernel functions ranging from polynomial, to Gaussian, to advanced graph kernels. In this way we model non-linear interactions between amino acid residues. Results were validated on manually curated data sets achieving competitive performance against various state-of-the-art approaches.

Availability: The MoDPepInt server is available under the URL: `http://modpepint.informatik.uni-freiburg.de/`

Contact: backofen@informatik.uni-freiburg.de

1 INTRODUCTION

Protein-protein interactions are often mediated by modular protein domains in eukaryotes and play an essential role in diverse biological processes such as signal transduction, cellular growth, cell polarity etc. (Pawson and Nash, 2003). Modular domains that specifically bind with short linear peptides are known as peptide recognition modules (PRMs). Each domain family recognizes peptides with specific characteristics. For example, phosphotyrosine (pY) containing peptides, proline-rich peptides and C-terminus peptides are recognized by SH2, SH3 and PDZ domains, respectively. However, individual domains from the same family show different binding specificity. Accurate models that can help understand the mechanisms responsible for the highly selective binding affinity are therefore of interest. Recently, several high-throughput techniques, such as protein microarray, phage display and SPOT synthesis, have been developed which can detect the binding specificity of various modular domains. However efficient bioinformatics tools are needed in order to extract meaningful knowledge from the enormous amount of data produced.

To this end, we used state-of-the-art machine learning approaches to build support vector machine (SVM) models that can accurately predict binding specificity. We have collected into a unified web-based system called MoDPepInt, three different tools: SH2PepInt, SH3PepInt and PDZPepInt for three different modular domains, namely SH2, SH3 and PDZ (Kundu *et al.*, 2013b,a; Kundu and Backofen, 2014). Currently we offer single domain models for 51 SH2 human and 69 SH3 human domains, and multi-domain models for 226 PDZ domains across the species. To assess the quality of our models we have used manually curated interaction data achieving competitive performance against various state-of-the-art approaches.

In summary, MoDPepInt unique features include (i) the largest number of modeled domains, and (ii) a comprehensive SH2, SH3 and PDZ domain-peptide prediction system in a single platform.

2 APPLICATION AND FUNCTIONALITY

2.1 Input

All tools have a unified input format. Query sequences (up to a maximum number of 500) can be supplied either in a FASTA format or using UniProt database accession numbers. PDZPepInt offers predictions also for domains that are newly developed and/or not comprised in the original 226 PDZ domains: the unknown query domain should be supplied in FASTA format. Multiple query domain sequences can also be provided.

2.2 Filters

Several filters are available to increase predictive accuracy. SH2 domains generally recognize phosphotyrosine (pY) residues of binding proteins. For this reason in SH2PepInt we offer a *phosphotyrosine* filter that only considers those peptides whose tyrosine phosphorylation has already been experimentally verified and reported in PhosphoSitePlus database (Hornbeck *et al.*, 2012).

As SH3 domains mainly bind with proline rich peptides, in SH3PepInt, we offer a *proline rich* filter that uses 31 regular expressions to select proline rich peptides (Carducci *et al.*, 2012).

*to whom correspondence should be addressed.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

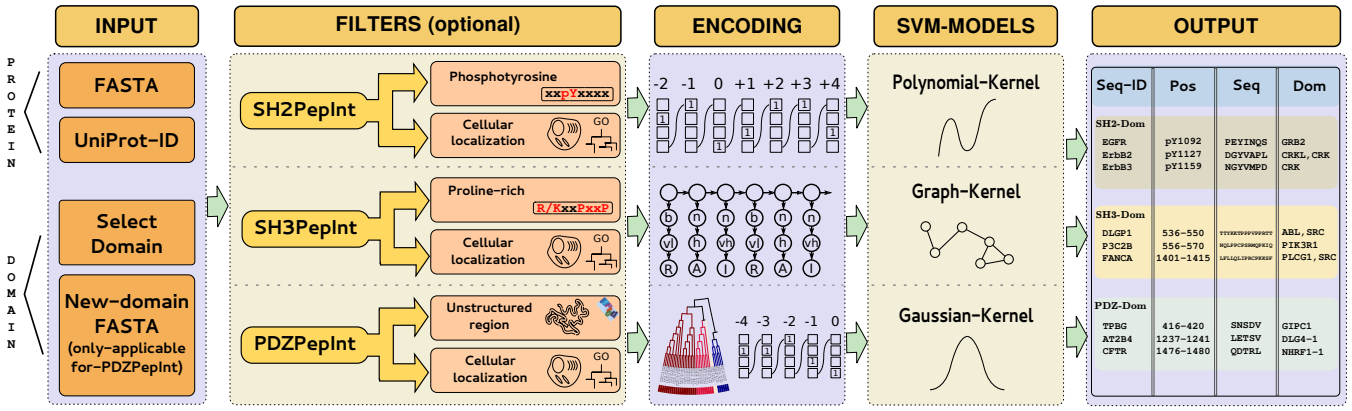

Fig. 1: Schematic representation of the MoDPepInt pipeline.

PDZ domains have the tendency to bind the unstructured regions of binding proteins, hence in PDZPepInt we offer a filter to select for *intrinsically unstructured/disordered regions* based on the IUPred algorithm (Dosztanyi *et al.*, 2005), which selects peptides with IUPred scores above 0.4 (Akiva *et al.*, 2012).

Finally, a *cellular localization* filter is available for all tools. This filter considers only those interactions where both the protein containing the peptide and the protein containing the modular domain have the same cellular localization according to the Gene Ontology Database (Ashburner *et al.*, 2000).

2.3 Processing and output

An internal queuing system (which currently uses 40 computation nodes) balances the submitted jobs in parallel. MoDPepInt is implemented in C++, perl and shell scripting, with runtimes typically ranging in the order of few minutes.

The output for all three tools is formatted as a downloadable table. We report for each domain-ligand protein interaction pair: (i) the sequence ID, (ii) the ligand binding position, (iii) the ligand binding sequence and (iv) the ligand binding domains.

See Figure 1 for the schematic representation of the MoDPepInt pipeline.

3 DISCUSSION

MoDPepInt collects three protein-protein interaction predictive models that can be efficiently tuned using data derived from various high-throughput experimental techniques. The resulting models exhibit significant performance improvement in comparison with other existing tools. The main sources of performance improvement are due to: (i) non-linear modeling, (ii) balanced discriminative training and (iii) datasets pooling.

SH2PepInt uses polynomial kernels and it is trained on additional high confidence negatives obtained via semi-supervised techniques.

SH3PepInt uses graph kernels on a complex representation of both the peptide sequence and of the aligned domains. The adoption of a graph-type representation allows the inclusion of the physico-chemical properties of amino acids which increase the generalization capacity of the models. Furthermore, the method does not need any prior alignment of the peptides. This is a big advantage since poly-proline rich peptides are hard to align.

PDZPepInt uses Gaussian kernels and it is trained on interaction data from additional highly related domains. Using pooling from closely related domains allows to leverage the limited information available for some domains and helps extrapolating to unseen, but alignable, novel domains.

Once trained, all models can be used to efficiently scan entire proteomes to identify novel interactions with typical runtimes of few minutes.

ACKNOWLEDGEMENT

Funding: This work was funded by the Bundesministerium für Bildung und Forschung (e-bio; FKZ 0316174A to RB) and the Excellent Initiative of the German Federal and State Governments (EXC 294 to RB).

Conflict of Interest: None declared.

REFERENCES

Akiva, E., Friedlander, G., Itzhaki, Z., and Margalit, H. (2012). A dynamic view of domain-motif interactions. *PLoS Comput Biol*, **8**(1), e1002341.

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*, **25**(1), 25–9.

Carducci, M., Perfetto, L., Briganti, L., Paoluzi, S., Costa, S., Zerweck, J., Schutkowski, M., Castagnoli, L., and Cesareni, G. (2012). The protein interaction network mediated by human SH3 domains. *Biotechnol Adv*, **30**(1), 4–15.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**(16), 3433–4.

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*, **40**(Database issue), D261–70.

Kundu, K. and Backofen, R. (2014). Cluster based prediction of PDZ-peptide interactions. *BMC Genomics*, **15**(Supp 1), S5.

Kundu, K., Costa, F., and Backofen, R. (2013a). A graph kernel approach for alignment-free domain-peptide interaction prediction with an application to human SH3 domains. *Bioinformatics*, **29**(13), i335–i343.

Kundu, K., Costa, F., Huber, M., Reth, M., and Backofen, R. (2013b). Semi-Supervised Prediction of SH2-Peptide Interactions from Imbalanced High-Throughput Data. *PLoS One*, **8**(5), e62732.

Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**(5618), 445–52.