

Sequence analysis

Advance Access publication May 28, 2014

MoDPepInt: an interactive web server for prediction of modular domain–peptide interactions

Kousik Kundu^{1,†}, Martin Mann^{1,†}, Fabrizio Costa¹ and Rolf Backofen^{1,2,*}¹Bioinformatics Group, Department of Computer Science, 79110 Freiburg and ²Centre for Biological Signalling Studies (BLOSS), 79104 Freiburg, University of Freiburg, Germany

Associate Editor: John Hancock

ABSTRACT

Summary: MoDPepInt (Modular Domain Peptide Interaction) is a new easy-to-use web server 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, respectively. 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 multidomain 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 datasets achieving competitive performance against various state-of-the-art approaches.

Availability and implementation: The MoDPepInt server is available under the URL <http://modpepint.informatik.uni-freiburg.de/>

Contact: backofen@informatik.uni-freiburg.de

Supplementary information: [Supplementary data](#) are available at [Bioinformatics](#) online.

Received on February 27, 2014; revised on April 25, 2014; accepted on May 15, 2014

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 and cell polarity (Pawson and Nash, 2003). Modular domains that specifically bind with short linear peptides are known as peptide recognition modules. 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 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 models that can accurately predict binding specificity. We have collected into a unified web-based system called MoDPepInt (Modular Domain Peptide Interaction), three different tools: SH2PepInt, SH3PepInt and PDZPepInt for three different modular domains, namely, SH2, SH3 and PDZ (Kundu *et al.*, 2013a,b; Kundu and Backofen, 2014). Currently, we offer single domain models for 51 SH2 human and 69 SH3 human domains, and multidomain models for 226 PDZ domains across human, mouse, fly and worm. 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) a domain-peptide prediction system for SH2, SH3 and PDZ in a single platform and (ii) the largest number of modeled domains (see [Supplementary Table S1](#)).

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).

PDZ domains have the tendency to bind the unstructured C-terminal 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 five C-terminal residues with IUPred scores >0.4 (Akiva *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.

