Bioinformatics, 32(6), 2016, 940–942
doi: 10.1093/bioinformatics/btv666
Advance Access Publication Date: 14 November 2015
Applications Note

OXFORD

Structural bioinformatics

MIEC-SVM: automated pipeline for protein peptide/ligand interaction prediction

Nan Li, Richard I. Ainsworth, Meixin Wu, Bo Ding, Wei Wang*

Department of Chemistry and Biochemistry, UC, San Diego, La Jolla, CA 92093-0359 USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on July 9, 2015; revised on October 13, 2015; accepted on November 7, 2015

Abstract

Motivation: MIEC-SVM is a structure-based method for predicting protein recognition specificity. Here, we present an automated MIEC-SVM pipeline providing an integrated and user-friendly workflow for construction and application of the MIEC-SVM models. This pipeline can handle standard amino acids and those with post-translational modifications (PTMs) or small molecules. Moreover, multi-threading and support to Sun Grid Engine (SGE) are implemented to significantly boost the computational efficiency.

Availability and implementation: The program is available at http://wanglab.ucsd.edu/MIEC-SVM.

Contact: wei-wang@ucsd.edu

Supplementary information: Supplementary data available at *Bioinformatics* online.

1 Introduction

Accurate prediction of the binding specificity of proteins, particularly those whose interacting partners are conformationally flexible, remains a great challenge. Many computational methods have been developed to meet this challenge (Chen *et al.*, 2008; Hui *et al.*, 2013) and one of them is the MIEC-SVM method. This method has been successfully applied to predict the binding specificity of diverse systems including SH3 domains (Hou *et al.*, 2008, 2009, 2012), PDZ domains (Li *et al.*, 2011), chromodomains (He *et al.*, 2015) and HIV protease (Ding *et al.*, 2013a,b; Hou *et al.*, 2009). Here, we present an automated MIEC-SVM pipeline that allows users to predict interacting partners using the existing MIEC-SVM models as well as to construct new models for the protein families of their own interest.

2 MIEC-SVM pipeline

The MIEC-SVM method aims to characterize the energetic patterns of proteins binding to their partners. Based on computational modeling of the complex structures, molecular interaction energy components (MIECs) such as van der Waals and electrostatic interaction energies between protein and peptide residues at the interaction interface are computed using biophysical models. Using the available binding specificity data, a classification model using a support vector machine (SVM) is then trained to predict binding and non-binding events. The MIEC-SVM method integrates energetic characteristics (MIEC) and a machine learning method (SVM) to reduce the error and noise in pure binding free energy calculations or bioinformatics methods based on sequence alone. Our previous studies have demonstrated that the MIEC-SVM method shows superior performance in predicting specificities of modular domains binding to peptides, scoring docking poses and identifying drug resistant mutants (Ding *et al.*, 2013a,b; Hou *et al.*, 2012; Li *et al.*, 2011).

To make the method more accessible to the scientific community, we construct a pipeline to facilitate the use of existing MIEC-SVM models and the construction of new ones (See Supplementary Material for detailed instructions). The pipeline is divided into three consecutive modules (Fig. 1A): complex structure building, MIEC generation and SVM training/prediction. The modular design of the program allows easy customization of the standard protocol for a particular application. The pipeline is implemented in Perl and can be installed in Linux/Unix systems with Perl support. Computationally demanding steps are implemented with multi-threading for multi-core CPUs. Moreover, SGE (Sun Grid Engine) is integrated into the energy decomposition part of

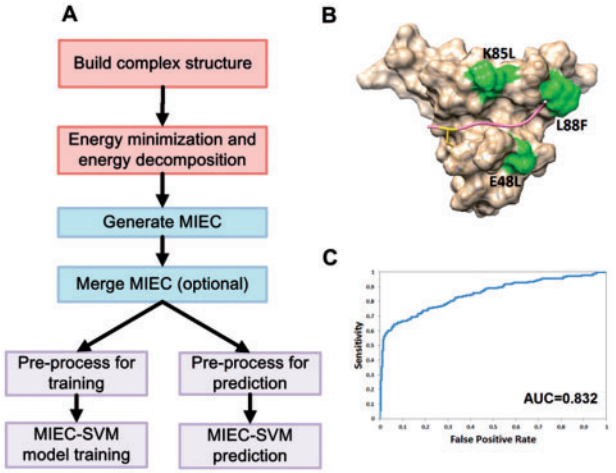

Fig. 1. Flowchart of the MIEC-SVM pipeline. (A) The pipeline is divided into three modules: model building (step 1-2), MIEC construction (step 3-4) and MIEC-SVM training/prediction (step 5-6); (B) the binding conformation of SUV92 mutant E48L/K85L/L88F, mutated sites are colored in green, the peptide colored in pink and the methylated lysine colored in yellow; (C) prediction performance of chromo domain MIEC-SVM on the binding specificity of the SUV92 mutant. The AUC of ROC curve is 0.832

the pipeline to support users with access to high performance computing clusters.

3 Example: MIEC-SVM application on SUV92 chromodomain mutant

SUV92 is a H3K9 specific methyl-transferase and specifically recognizes H3K9me3 with its chromodomains. Our previous study (He *et al.*, 2015) showed that the binding specificity of the SUV92 chromodomain to 457 selected peptides can be altered by only a few mutations. The MIEC-SVM model trained from 13 human chromodomains performs well on predicting the binding specificity of the SUV92 mutants. Here, we use one of the SUV92 chromodomain mutants (E48L/K85L/L88F, Fig. 1B) to demonstrate the application of the MIEC-SVM pipeline.

The first module is 'complex building'. The complex structure of the SUV92 mutant and peptide is generated through virtual mutagenesis. For each mutant-peptide interaction, 8 complex structures were generated from the corresponding templates, which were taken from the molecular dynamics simulation trajectory of the SUV92 wild type. The peptide sequence and the mutated sites are input to the pipeline using two plain text files. For residues that are non-standard amino acids, molecular information of these residues can be input to the database 'MolDB'. For the SUV92 mutant, 'MolDB' includes the information of several modified amino acids, such as mono-, di-, tri-methylated lysine, di-methylated arginine, phosphorylated serine and phosphorylated threonine.

Depending on whether the peptide contains post-translation modifications (PTMs), two different protocols are used to build the complex structure. For standard amino acids, the program Scwrl (Krivov *et al.*, 2009) generates the side chain conformations of the mutated residues. For residues with PTMs, the residue is first mutated to the un-modified version using Scwrl. Then, PTMs are added to the residue and the complex conformation is optimized by the sander program in the AMBER package(Case *et al.*, 2010). The complex modeling step is followed by the MM/GBSA energy decomposition, with the igb = 2 option, to generate the energy decomposition profile. For each residue pair between the SUV92 mutant and the peptide, four energy components are used to characterize the interaction, including van de Waals (VDW), electrostatics (ELE), generalized Born (GB) and surface area (SA).

The second module is 'MIEC construction'. We construct the MIEC profile for each mutant-peptide complex as such: (P1-VDW, P1-ELE, P1-GB, P1-SA, P2-VDW . . . , P*n*-GB, P*n*-SA, Q1-VDW,Q1-ELE . . . Q*m*-GB, Q*m*-SA), where P*i* is the *i*th residue pair between receptor and ligand, and Q*j* is the *j*th residue pair between the adjacent ligand residues. In the chromodomain application, we include all residue pairs whose minimum distances are less than 10 Å. Users can also combine different energy components (Hou *et al.*, 2009) or contributions from multiple residue pairs (Ding *et al.*, 2013a,b) to construct MIECs.

In the third module of 'MIEC-SVM', the MIEC profiles from the eight templates are averaged. Each component of the MIEC profile of the mutant is linearly scaled to the range of −1 and 1. The SUV92 mutant MIEC profile is input to the MIEC-SVM model trained on the 13 human chromodomains interacting with 457 peptides. By comparing to the peptide microarray data between the mutant and the 457 peptides (He *et al.*, 2015), we demonstrate that the prediction performance of the chromodomain MIEC-SVM model on the SUV92 mutant is 0.832 measured by AUC of ROC curve (Fig. 1C). Note that the previously trained MIEC-SVM models for predicting chromodomain binding specificity(He *et al.*, 2015) and HIV protease drug resistance(Ding *et al.*, 2013a,b) are also available to users. Moreover, this module also includes 'SVM training' for users to build their own model of interest with feature selection using LASSO logistic regression.

4 Conclusion

The MIEC-SVM method is a powerful tool to predict the binding specificity of proteins. The pipeline presented here automates the model training and prediction, which will greatly facilitate the broader application of this method.

Funding

The work was partially supported by NIH (R01GM085188 to W.W.).

Conflict of Interest: none declared.

References

Chen,J.R. *et al.* (2008) Predicting PDZ domain-peptide interactions from primary sequences. *Nat. Biotechnol.*, **26**, 1041–1045.

Case,D.A. *et al.* (2010) AMBER 11. University of California, San Francisco.

Ding,B. *et al.* (2013a) Characterizing binding of small molecules. II. Evaluating the potency of small molecules to combat resistance based on docking structures. *J. Chem. Inf. Model*, **53**, 1213–1222.

Ding,B. *et al.* (2013b) Characterization of small molecule binding. I. Accurate identification of strong inhibitors in virtual screening. *J. Chem. Inf. Model*, **53**, 114–122.

He,W. *et al.* (2015) Deciphering and engineering chromodomain-methyllysine peptide recognition. *In submission*.

Hou,T. *et al.* (2012) Characterization of domain-peptide interaction interface: prediction of SH3 domain-mediated protein-protein interaction network in yeast by generic structure-based models. *J. Proteome Res.*, **11**, 2982–2995.

Hou,T. *et al*. (2009) Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. *Mol. Cell Proteomics*, **8**, 639–649.

Hou,T. *et al.* (2008) Characterization of domain-peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. *J. Mol. Biol.*, **376**, 1201–1214.

Hou,T. *et al*. (2009) Predicting drug resistance of the HIV-1 protease using molecular interaction energy components. *Proteins*, **74**, 837–846.

Hui,S. *et al*. (2013) Predicting PDZ domain mediated protein interactions from structure. *BMC Bioinformatics*, **14**, 27.

Krivov,G.G. *et al*. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.

Li,N. *et al.* (2011) Characterization of PDZ domain-peptide interaction interface based on energetic patterns. *Proteins*, **79**, 3208–3220.