OXFORD

Review and comparative assessment of sequence-based predictors of protein-binding residues

Jian Zhang and Lukasz Kurgan

Corresponding author. Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA. Tel.: +1-804-827-3986; Fax: +1-804-828-2771; E-mail: lkurgan@vcu.edu

Abstract

Understanding of molecular mechanisms that govern protein–protein interactions and accurate modeling of protein–protein docking rely on accurate identification and prediction of protein-binding partners and protein-binding residues. We review over 40 methods that predict protein–protein interactions from protein sequences including methods that predict interacting protein pairs, protein-binding residues for a pair of interacting sequences and protein-binding residues in a single protein chain. We focus on the latter methods that provide residue-level annotations and that can be broadly applied to all protein sequences. We compare their architectures, inputs and outputs, and we discuss aspects related to their assessment and availability. We also perform first-of-its-kind comprehensive empirical comparison of representative predictors of protein-binding residues using a novel and high-quality benchmark data set. We show that the selected predictors accurately discriminate protein-binding and non-binding residues and that newer methods outperform older designs. However, these methods are unable to accurately separate residues that bind other molecules, such as DNA, RNA and small ligands, from the protein-binding residues. This cross-prediction, defined as the incorrect prediction of nucleic-acid- and small-ligand-binding residues as protein binding, is substantial for all evaluated methods and is not driven by the proximity to the native protein-binding residues. We discuss reasons for this drawback and we offer several recommendations. In particular, we postulate the need for a new generation of more accurate predictors and data sets, inclusion of a comprehensive assessment of the cross-predictions in future studies and higher standards of availability of the published methods.

Key words: Protein–protein binding; prediction of protein-binding residues; protein–protein interactions; protein–nucleic acids interactions

Introduction

Proteins are biomacromolecules that interact with a variety of other molecules including DNA, RNA, small ligands and other proteins [1–4]. Protein–protein interactions drive many cellular processes, such as signal transduction, transport and metabolism, to name but a few. Knowledge of these interactions at a molecular level is important to develop novel therapeutics [5–7], annotate protein functions [8], study molecular mechanisms of diseases [9, 10] and delineate protein–protein interaction networks [11]. Several databases, such as Mentha [12], BioLip [13] and Protein Data Bank (PDB) [14], archive information about protein–protein interactions at molecule (protein) and molecular (residue or atomic) levels. The Mentha resource includes annotations of >86 000 protein–protein interactions at the protein level. BioLip archives 17 000 interactions and includes annotations of protein-binding residues. PDB provides access to 71 000 protein–protein complexes with detailed atomic-level structures. However, these annotations of protein–protein interactions are highly incomplete, especially if we factor in the facts that protein–protein interactions are promiscuous [15] and that we currently know >67 million proteins [16]. Most of these proteins lack functional annotations including the information about the protein–protein interactions. Computational methods that predict protein–protein interactions from the sequences can help to bridge this gap.

Jian Zhang is a Lecturer in School of Computer and Information Technology at the Xinyang Normal University and a visiting scholar at the Virginia Commonwealth University. His research interests are focused on machine learning and bioinformatics.
Lukasz Kurgan is a Qimonda Endowed Professor at the Virginia Commonwealth University in Richmond. His research concerns high-throughput structural and functional characterization of proteins and small RNAs. More details about his research group can be found at http://biomine.cs.vcu.edu/.

Table 1. Categorization of methods that predict protein–protein interactions depending on the inputs (protein sequence versus structure) and outputs (interacting proteins versus residues)

Inputs	Outputs	
	Interacting proteins	Interacting residues
Structure	pSTR-to-PRO: methods that predict whether a given pair of structures interact	pSTR-to-RES: methods that predict protein-binding residues for a given pair of structures
		sSTR-to-RES: methods that predict protein-binding residues for a given single structure
Sequence	pSEQ-to-PRO: methods that predict whether a given pair of sequences interact	pSEQ-to-RES: methods that predict protein-binding residues for a given pair of sequences
		sSEQ-to-RES: methods that predict protein-binding residues for a given single sequence

Numerous computational methods for the prediction of protein–protein interactions have been developed in the recent years [17–22]. These methods can be divided into two groups based to the inputs that they use to perform predictions: structure based versus sequence based [22]. Moreover, the inputs of the structure-based methods could be either experimentally determined structures or structures that are predicted from protein sequences, typically using homology modeling. The use of the putative protein structures lowers the predictive quality of the predicted protein–protein interactions, and the extent of this decrease depends on the quality of the predicted structures [22]. Protein–protein docking and homology-based modeling are the two commonly used approaches that are used to implement the structure-based methods [23]. The former approach samples possible orientations and conformations of protein–protein complexes and then uses empirical scoring functions to select the most energetically favorable structure of the complex [24–27]. The latter uses structure similarity to select proteins with similar structures from a database of known protein–protein complexes and transfers the annotations of interactions from these complexes onto the input protein [28, 29]. However, the use of the structure-based methods is limited by a relatively small set of proteins with experimentally determined structures and by computational cost of generating putative protein structures. These methods may also suffer substantial reduction in the predictive performance if the putative structures they use are not accurate [22]. In contrast, the sequence-based methods for the prediction of protein–protein interactions only need the protein sequence to predict protein–protein interactions. They can be applied to a much larger population of proteins with known sequences and do not require the computationally costly modeling of the structure. The sequence-based methods are subdivided based on granularity of the putative annotations of binding that they produce into two types: protein-level based versus residue-level based. The protein-level-based methods predict whether a given pair of proteins interacts. This can be done using both sequence-based and structure-based methods. The residue-level-based methods predict binding residues in a single protein sequence or in a pair of interacting protein sequences. Table 1 summarizes these different types of the structure- and sequence-based methods for the prediction of interacting protein and residues.

The availability of many predictors of protein–protein interactions prompted publication of six reviews, which cover both structure- and sequence-based methods [17–22]. Table 2 summarizes these reviews. Three reviews describe and discuss various predictors of protein binding, while the other three additionally perform empirical analysis. The first three articles discuss physicochemical characteristics of binding residues and binding interfaces including their evolutionary conservation and topological features [18, 19, 21]. The review by Esmaielbeiki et al. also classifies protein interface prediction methods and summarizes their inputs and predictive models [21]. The other three reviews empirically assess the predictive performance of several predictors, primarily focusing on the structure-based prediction of protein–protein interactions [17, 20, 22]. While these six articles cover a large number of structure-based methods, Table 2 reveals that they review no more than 12 sequence-based methods, which do not include recent methods published after 2013. Our analysis shows that there are 44 sequence-based methods, and 21 of them were published in the past 3 years. Also, these reviews empirically evaluate only a couple of older sequence-based methods.

The discussion of the available reviews indicates a clear need for a comprehensive review and empirical benchmarking of the sequence-based methods. To this end, we cover a comprehensive set of 44 sequence-based predictors of protein-binding residues, including methods that provide predictions at the protein and residue levels. We discuss their inputs, predictive models, outputs and we offer practical and insightful analysis of their availability. We also empirically evaluate a set of seven representative sequence-based predictors of protein-binding residues, which includes five methods that were released in the past 3 years; see Table 2. This assessment was performed on a novel and large benchmark data set that is characterized by a more comprehensive set of native annotations of binding residues than the currently used data sets. The latter stems from the fact that we are the first to transfer annotation of protein binding within clusters of protein–protein complexes that involve the same proteins. We are also the first to offer a detailed analysis of the sources of predictive errors.

Overview of the sequence-based predictors of protein–protein interactions

Sequence-based predictors of protein- and residue-level protein–protein interactions

First, we perform literature search to select relevant methods. We search PubMed database on 31 July 2016 by combining results of two queries: 'protein-binding AND sequence' and 'protein-protein interaction AND sequence' and we found 1585 articles. Next, we select recent and relevant publications based on reading the abstracts. In particular, we select articles that were published in the past decade and that describe predictive methods. Among these selected methods, we consider the

Table 2. Summary and comparison of recent reviews of predictors of protein–protein binding

Review article (year published)	Type of methods covered	Scope of review of SEQ methods		Scope of evaluation of SEQ methods					
		Number of SEQ methods reviewed	Number of recent SEQ methods reviewed (2014–16)	Number of SEQ methods evaluated	Number of recent SEQ methods evaluated (2014–16)	Size of test data set	Test data set is dissimilar to training data sets	Test data set includes full protein sequences	Assess prediction of binding to other ligands
This review	SEQ	44	21	7	5	448	√	√	√
[21] (2016)	SEQ, STR	9	0	N/A	N/A	N/A	N/A	N/A	N/A
[19] (2015)	SEQ, STR	2	0	N/A	N/A	N/A	N/A	N/A	N/A
[20] (2015)	SEQ, STR	4	0	2	0	176	×	×	×
[22] (2015)	SEQ, STR	2	0	1	0	90	×	×	×
[18] (2011)	SEQ, STR	4	0	N/A	N/A	N/A	N/A	N/A	N/A
[17] (2009)	SEQ, STR	12	0	0	0	149	×	×	×

The two main types of methods are structure based (STR) and sequence based (SEQ). N/a means that a given aspect is outside of the scope, √ and × represent that a given feature is and it is not considered by the authors, respectively.

newest version of methods that have multiple versions. We found 44 relevant articles. Supplementary Figure S1 shows that there were 7 methods released between 2006 and 2009, 16 between 2010 and 2013 and 21 since 2014. This increasing trend in the number of methods released in recent years demonstrates strong interest in this predictive task.

There are three types of sequence-based predictors of protein–protein interactions, which are defined according to their inputs (single versus pair of protein sequences) and outputs (sequence versus residue level). The pSEQ-to-PRO methods predict whether a given pair of protein sequences interacts. The pSEQ-to-RES approaches predict protein-binding residues for a pair of input protein sequences. Finally, the sSEQ-to-RES methods predicts binding residues in a single input protein sequence. Table 3 reveals that 23 of the 44 methods belong to the pSEQ-to-PRO group, 5 are in the pSEQ-to-RES group and 16 in the sSEQ-to-RES category. Many methods were published in the past 3 years, primarily from the pSEQ-to-PRO and sSEQ-to-RES types. Among the 44 methods, 28 (or 64%) were released to the research community as freely available webservers or source code. Table 3 provides the corresponding URLs (Uniform Resource Locators) to facilitate finding these predictors. The availability of the source code means that users will need to download the program, install it and run it on their own computer. Most of the recently published method are provided this way. While this might be an attractive option for bioinformaticians, especially in when these programs need to be incorporated into other computational platforms, these tasks could be prohibitively difficult for biologists. The webservers cater to less computer-savvy users. The users only need a web browser that is connected to the Internet to perform prediction. They simply arrive at the given URL, enter their sequence(s) and click start. The predictions are performed on the server side and the results are delivered back to the users via the web browser and/or email. Unfortunately, 11 of the 28 available methods are no longer maintained or take >30 min to predict a single protein. On the positive note, the number of the publicly available prediction tools that were developed in the past 3 years is twice the number of the tools that were created in the previous 7 years.

sSEQ-to-RES: methods that use single sequence to predict protein-binding residues

The three types of sequence-based predictors of protein–protein interactions use different inputs and generate different outputs. They also require different types of data sets to build predictive models and use different test protocols and measures to perform empirical assessment. Consequently, each of the three types of methods would require a uniquely structured review. The methods in the sSEQ-to-RES group offer more detailed residue-level annotations compared with the sequence-level annotations generated by the pSEQ-to-PRO methods. Moreover, they can be used for any of the millions of proteins with known sequences, compared with the pSEQ-to-RES methods that are limited to proteins that have known binding protein partners (they take interacting protein pairs as the inputs). Therefore, given their more detailed predictions and broad applicability, we focus our review and comparative assessment on the sSEQ-to-RES methods. The other two categories of methods will be the subject of future studies.

Nowadays, the sSEQ-to-RES predictors include methods that focus on the protein-binding residues and also methods that predict residues that interact with a variety of other ligands. Examples include methods that predict RNA- and DNA-binding

Table 3. Summary of the sequence-based predictors of the protein–protein interactions

Type	Method	Ref.	Year	Predictor	URL
pSEQ-to-PRO	Shen et al.	[30]	2007	N/A	N/A
	Predict_PPI	[31]	2008	Web server	http://www.scucic.cn/Predict_PPI/index.htm.
	Yu et al.	[32]	2010	N/A	N/A
	Meta_PPI	[33]	2010	Source code	http://home.ustc.edu.cn/~jfxia/Meta_PPI.html
	PRED_PPI	[34]	2010	**Web server**	http://cic.scu.edu.cn/bioinformatics/predict_ppi/default.html
	BRS-nonint	[35]	2010	**Web server**	http://www.bioinformatics.leeds.ac.uk/BRS-nonint/
	Zhang et al.	[36]	2011	Source code	http://www.csbio.sjtu.edu.cn/bioinf/CS/
	SPPS	[37]	2011	**Web server**	http://mdl.shsmu.edu.cn/SPPS/
	PPIPP	[38]	2011	Web server	http://tardis.nibio.go.jp/netasa/ppipp/
	Yousef et al.	[39]	2013	N/A	N/A
	PPIevo	[40]	2013	Web server	http://lbb.ut.ac.ir/Download/LBBsoft/PPIevo/
	You et al.	[41]	2013	N/A	N/A
	MCDPPI	[42]	2014	Source code	http://csse.szu.edu.cn/staff/youzh/MCDPPI.zip
	You et al.	[43]	2015	Source code	https://sites.google.com/site/zhuhongyou/data-sharing/
	VLASPD	[44]	2015	**Source code**	http://www.comp.polyu.edu.hk/~cslhu/resources/vlaspd/
	Profppikernel	[45]	2015	**Source code**	https://rostlab.org/owiki/index.php/Profppikernel
	You et al.	[46]	2015	N/A	N/A
	Jia et al.	[47]	2015	**Web server**	http://www.jci-bioinfo.cn/PPI
	Huang et al.	[48]	2015	N/A	N/A
	Gao et al.	[49]	2016	N/A	N/A
	Sze-To et al.	[50]	2016	N/A	N/A
	Huang et al.	[51]	2016	N/A	N/A
	An et al.	[52]	2016	N/A	N/A
pSEQ-to-RES	PIPE	[53]	2006	Web server	http://pipe.cgmlab.org/
	Shi et al.	[54]	2010	N/A	N/A
	Chang et al.	[55]	2010	N/A	N/A
	PIPE-Sites	[56]	2011	Web server	http://pipe-sites.cgmlab.org/
	PETs	[57]	2015	**Source code**	https://github.com/BinXia/PETs
sSEQ-to-RES	ISIS	[58]	2007	N/A	N/A
	SPPIDER	[59]	2007	**Web server**	http://sppider.cchmc.org/
	Du et al.	[60]	2009	N/A	N/A
	Chen et al.	[61]	2009	Source code	http://ittc.ku.edu/~xwchen/bindingsite/prediction
	PSIVER	[62]	2010	**Web server**	http://tardis.nibio.go.jp/PSIVER/
	Chen et al.	[63]	2010	Source code	http://mail.ustc.edu.cn/~bigeagle/BMCBioinfo2010/index.htm
	HomPPI	[64]	2011	**Web server**	http://homppi.cs.iastate.edu/
	Wang et al.	[65]	2014	N/A	N/A
	LORIS	[66]	2014	**Source code**	https://sites.google.com/site/sukantamondal/software
	SPRINGS	[67]	2014	**Source code**	https://sites.google.com/site/predppis/
	CRF-PPI	[68]	2015	**Source code**	http://csbio.njust.edu.cn/bioinf/CRF-PPI
	Geng et al.	[69]	2015	N/A	N/A
	iPPBS-Opt	[70]	2016	**Web server**	http://www.jci-bioinfo.cn/iPPBS-Opt
	PPIS	[71]	2016	**Source code**	http://csbio.njust.edu.cn/bioinf/PPIS
	SPRINT	[72]	2016	**Source code**	http://sparks-lab.org/yueyang/server/SPRINT/
	SSWRF	[73]	2016	**Source code**	http://csbio.njust.edu.cn/bioinf/SSWRF/

We group these predictors into three types: pSEQ-to-PRO, pSEQ-to-RES and sSEQ-to-RES. The 'web server' and 'source code' indicate that a given method is available as the online web server and standalone source code, respectively. The bold font indicates that the corresponding predictor is available and provides prediction for a single protein in < 30 min. 'N/A' means that neither web server nor source code is available.

residues [74–79] and a variety of other, small ligands [80]. The latter group of methods includes predictors of nucleotide-binding residues [81, 82], metal-binding residues [83], residues that interact with vitamins [84, 85], calcium [86], as well as methods that predict binding to multiple types of small ligands [87]. Picking a suitable sSEQ-to-RES predictor of protein-binding residues could be a daunting task given that currently already 16 of them were published. We provide practical information concerning the architecture of these methods, their outputs and their predictive performance to facilitate an informed selection. Table 4 summarizes architectures and outputs of these predictors and discusses how they were assessed in the past studies.

There are two main types of architectures of these predictive models. One is based on the sequence alignment and the other uses predictive models, which are generated by machine learning algorithms. The alignment-based methods rely on the assumption that proteins with similar sequences share similar binding partners and binding residues [64]. They require a data set of proteins with known annotations of protein-binding residues. They perform predictions by transferring annotations of binding residues from proteins in that data set that are sufficiently similar to the input protein; for example, having sequence similarity >30% or the log(Evalue)< −50. The machine learning-based methods predict propensity for protein binding for each residue in the input sequence using a predictive model,

Table 4. Summary of the single sequence-based predictors of protein-binding residues

Method	Year	Architecture					Evaluation			Outputs and performance measurement	
		Window	Sequence only	Solvent accessibility	Evolutionary conservation	Predictive model	k-fold cross-validation on training data set	Leave-one-out cross-validation on training data set	Test on test data set (similarity to the training data set)	Binary values	Propensity scores
ISIS	2007	9	×	×	√	NN	×	×	√ (N/A)	ACC	×
SPPIDER	2007	11	×	×	√	KNN	10	×	√ (50%)	SN, SP, ACC, MCC	AUC
Du et al.	2009	11	√	√	√	SVM	5	×	×	SN, SP, ACC, MCC, F1	AUC
Chen et al.	2009	21	×	×	√	RF	×	×	√ (30%)	SN, SP, ACC, MCC	AUC
PSIVER	2010	9	×	√	√	NB	×	√	√ (25%)	SN, SP, ACC, MCC, F1	AUC
Chen et al.	2010	19	√	×	√	SVM	5	×	√ (30%)	SN, SP, ACC, MCC, PRE, F1	×
HomPPI	2011	×	×	×	√	Alignment	×	×	√ (30%)	SN, SP, ACC, MCC	×
Wang et al.	2014	11	×	×	√	SVM	5	×	√ (25%)	SN, PRE, ACC	×
LORIS	2014	9	√	√	√	RLF	×	√	√ (25%)	SN, SP, PRE, ACC, MCC, F1	×
SPRINGS	2014	9	√	√	√	NN	×	√	√ (25%)	SN, SP, PRE, ACC, MCC, F1	×
CRF-PPI	2015	9	√	√	√	RF	×	√	√ (25%)	SN, SP, PRE, ACC, MCC, F1	AUC
Geng et al.	2015	9	×	√	√	NB	×	√	√ (25%)	SN, SP, PRE, ACC, MCC, F1	×
iPPBS-Opt	2016	15	√	×	×	KNN	10	×	×	SN, SP, ACC, MCC	AUC
PPIS	2016	9	√	√	√	RF	×	√	√ (25%)	SN, SP, PRE, ACC, MCC, F1	×
SPRINT	2016	9	√	√	√	SVM	10	×	√ (30%)	SN, SP, ACC, MCC	AUC
SSWRF	2016	9	√	√	√	SVM, RF	×	√	√ (25%)	SN, SP, PRE, ACC, MCC, F1	AUC

We summarize key aspects including their architecture (input features and classifiers used to perform predictions), evaluation and performance measurements that were used in past studies, and their outputs. The first four sub-columns under the architecture list various classes of features. √ means that a given aspect (feature class) is relevant or considered, while × indicates that it is not considered. The 'predictive model' column lists machine learning algorithms that are used to build predictive models including neural networks (NN), K-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), naïve Bayes (NB), regularized logistic function (RLF) and radial basis function (RBF). One method is based on the sequence alignment. We show the number of folds k in the 'k-fold cross-validation on the training data set' column. For the 'binary values' column, SN, SP, PRE, ACC, MCC and F1 stand for sensitivity or recall, specificity, precision, accuracy, Mathew's correlation coefficient and F1-measure, respectively. For the 'propensity scores' column, AUC is the area under ROC curve. The definition of these measurements is provided in the 'Measures of predictive performance' section. Methods that have listed values in the 'binary values' column output binary predictions of binding residues (protein binding versus other residues). Methods that have listed values in the 'propensity scores' column output propensities for the protein binding (a numeric score that quantifies likelihood that a given residue binds proteins).

instead of relying on the sequence similarity. The predictive models are generated by machine learning algorithms with the aim to differentiate between protein-binding and the remaining residues in a training data set of annotated protein sequences. These methods provide accurate predictions for proteins that are not limited by high levels of similarity with the proteins from the training data set. In particular, the machine learning-based methods produce accurate results for proteins that share low (<30%) similarity with proteins from the training data set, and thus they complement predictions that can be obtained using the alignment-based approaches. Among the 16 sSEQ-to-RES predictors listed in Table 4, there is one alignment-based method (HomPPI [64]) and 15 machine learning-based methods.

The machine learning-based methods perform predictions in the following two steps. First, each residue in the input protein chain is encoded with a feature vector. Second, the vector is input into the predictive model that generates predictions. In the first step, the vector of numeric features quantifies structural and physicochemical characteristics of the predicted residue and its neighbors in the sequence. These neighbors form a window that is centered on the predicted residue. Use of the window is motivated by the fact that the knowledge of the characteristics of the neighboring residues provides useful clues for the prediction of the residue in the center of the window [74]. The length of the window varies widely between 9 and 21 residues among different methods, with 9 residues being the most commonly used value, especially for the recent predictors (Table 4). The features are computed from two types of inputs: directly from the protein sequence and from putative structural information that is predicted from the protein sequence. The former type of features includes physicochemical properties and evolutionary conservation of amino acids as well as amino acid composition. The latter features are derived from the putative relative solvent accessibility that is obtained with other predictive tools, such as SANN [88] and PSIPRED [89]. The relative solvent accessibility is defined as a predicted solvent accessible surface of a given amino acid in the input sequence divided by the maximal possible solvent accessible surface area of that amino acid. This information is useful because the protein-binding residues are likely to be located on the solvent accessible protein surface. While a few of the early methods use solely the features computed directly from the sequence [58, 59, 61], most of the methods published in the past 3 years combine both types of features (Table 4). The most popular by far feature type is the evolutionary conservation, which is typically computed from the position-specific scoring matrix generated by the PSI-BLAST algorithm [90]. In the second step that performs prediction of protein-binding residues, the features are input into a predictive model (classifier) that computes predictions in the form of binary values (protein-binding versus other residues) and/or propensities for binding (a numeric score that quantifies likelihood that a given residue binds proteins). Half of the 16 methods generate both propensities and binary values, while the other eight generate only the binary values. For the former eight methods, which can be identified based on the 'Propensity scores' column in Table 4, these propensities are typically converted into binary values by using a threshold. More specifically, residues with the putative propensities below the threshold are predicted not to bind proteins, while residues with the propensities above the threshold are predicted to bind proteins. The most popular machine learning algorithm that is used to generate these predictive models is support vector machine [91]; it was used in 5 of the 16 predictors (Table 4). The second most popular algorithm is random forest [92].

The sSEQ-to-RES predictors are assessed using a variety of test types and measures of predictive performance, typically using test sets of proteins that were not used to build these models. These tests aim to estimate predictive performance that end users should expect to observe on his/her proteins of interest, which is why evaluation is done on proteins that are not used to build the predictive models. These tests include cross-validation on the training data sets and tests on 'independent' (different from the training data set) test data sets. Most of the methods were evaluated using both test types (Table 4). In the k-fold cross-validation, the training data set is divided into k equally sized parts (folds). Each time, k − 1 folds are used to train the predictive model and the remaining fold is used as the test set. This is repeated k times so that each fold is used once as the test set. The leave-one-out cross-validation is an extreme case of the k-fold cross-validation where k is equal to the number of all proteins in the data set. Another important aspect of the assessment of these single-sequence-based methods for the prediction of protein-binding residues is the fact that proteins in the independent test sets share low sequence similarity with the training proteins, typically <25 or 30% (Table 4). This is because proteins with higher levels of similarity can be accurately predicted by the alignment-based methods.

There are two groups of measures of predictive performance of the sSEQ-to-RES predictors that address evaluation of the two types of outputs: the propensity scores and binary values. The measures that target the binary predictions include sensitivity, specificity, precision, accuracy, Matthews correlation coefficient (MCC) and F1-measure (Table 4). Sensitivity and specificity measure the fraction of protein-binding residues or non-protein-binding residues that are correctly identified as such, respectively. Accuracy quantifies the fraction correctly predicted protein-binding and non-protein-binding residues. Precision is defined as the ratio of correctly predicted protein-binding residues among all predicted protein-binding residues. The MCC and F1 measures take into account both correctly predicted protein-binding residues and correctly predicted non-protein-binding residues. These two measures are regarded as balanced, which means that they can provide an accurate measurement of predictive performance for imbalanced data sets. The data sets in this area are typically imbalanced, with a significant majority of the residues being non-protein-binding and only a relatively small number of protein-binding residues. The AUC, which quantifies the area under ROC (receiver operating characteristic) curve, is used to evaluate the putative propensities. The ROC curve represents a tradeoff between sensitivity and false-positive rate = 1 − specificity. Higher value corresponds to more accurate predictions. Three of the four methods that were published in 2016 generate propensity scores and were evaluated using AUC.

Overall, we show that most of the sSEQ-to-RES predictors were developed in the past 3 years and that their predictive models were generated with machine learning algorithms that use a variety of feature types as inputs. The empirical assessment of these methods relies on the independent test sets that share low sequence similarity with proteins used to generate the predictive models and a mixture of several measures of predictive performance. The diversity of these measures and strict standards on the similarity are hallmarks of a mature field of research. Similar standards are in place in other areas of prediction of one-dimensional descriptors of protein functions and structure, such as secondary structure, solvent accessibility, residue contacts and others [93].

Comparative empirical assessment of single-sequence methods that predict protein-binding residues

Benchmark data sets

The source data for our benchmark data sets were collected from the BioLip database [13] in October 2015. These data contain 5913 DNA-binding chains, 20 731 RNA-binding chains, 163 589 protein-binding chains and 112 797 ligand-binding chains. A given residue is defined as binding if the distance between an atom of this residue and an atom from a given ligand is <0.5Å plus the sum of the Van der Waal's radii of the two atoms [13]. Our goal is to create a large, high-quality and non-redundant data set that uniformly samples the annotated protein sequences. First, to ensure the high quality we remove protein fragments. Next, we map BioLip sequences into UniProt records with identical sequences to allow future users of this data set to map these proteins to other databases and to collect additional functional and structural annotations. This also allows us to improve quality of annotations of binding by mapping binding residues across different protein–protein complexes where one of the protein is shared; this way we transfer annotations of binding residues from all of these complexes onto the UniProt sequence. We ensure that the resulting data set is non-redundant by using Blastclust [90] to cluster protein sequences with a threshold of 25% similarity. For each cluster of proteins that share >25% similarity, we select a protein that was most recently released in UniProt. The resulting data set includes 1291 protein sequences.

Next, we ensure that proteins in our data set share low similarity with the proteins in the data sets used to develop the sSEQ-to-RES predictors that are included in the comparative assessment. This facilitates fair comparison that adheres to the standards in this field. First, we collect the training data sets of the seven predictors that we assess: SPPIDER [59], PSIVER [62], LORIS [66], SPRINGS [67], CRF-PPI [68], SPRINT [72] and SSWRF [73]; the selection of these predictors is explained in the 'Selection of single-sequence methods that predict protein-binding residues for the comparative assessment' section. SPPIDER has used training set S435 with 435 protein chains. SPRINT has used a large training set with 1199 proteins. Finally, PSIVER, LORIS, SPRINGS, CRF-PPI and SSWRF adopted the same training set Dset186. We chose to limit similarity of the proteins in our data set to the proteins in all of these training data sets to 25% given that this threshold is the most often used in the prior studies (Table 4). We use Blastclust to cluster our 1291 proteins together with the proteins from the three training data sets at 25% similarity. We remove proteins from our set that are in clusters that include any of the proteins from the training data sets. The resulting 1120 protein sequences share <25% similarity with each other and with the training proteins used by the considered seven predictors. Because some of the seven predictors are computationally expensive, we randomly pick 40% of the 1120 proteins as the final benchmark data set. The selected set of 448 proteins constitutes our benchmark test data set, which we name Dataset448. This data set is substantially larger that the data sets used in prior reviews of the predictors of protein–protein binding [20–22], which use data sets with between 90 and 176 proteins (Table 2).

Besides testing the overall predictive performance of the considered methods on our benchmark data set, we also investigate whether these predictors can accurately identify protein-binding residues among residues that bind other types of ligands (other-ligand-binding residues). Dataset448 contains 15 810 protein-binding residues (13.6% of all residues in the data set), 557 DNA-binding residues (0.5%), 696 RNA-binding residues (0.6%), 7175 residues that interact with small ligands (6.2%) and 93 857 non-binding residues (80.6%) that do not bind any of these ligands. We also name the residues that do not bind proteins, which include the non-binding residues and the residues that bind DNA, RNA or small ligands, as 'non-protein-binding residues'. To quantify and compare the ability of these predictors to identify protein-binding residues among all ligand-binding residues, we define two subsets of the Dataset448 data set. The PBPdataset336 is a data set of 336 protein-binding proteins, which excludes proteins from Dataset448 that bind only ligands that are not proteins. The nPBPdataset112 is a data set that includes 112 proteins from Dataset448 that bind only the ligands that are not proteins.

Moreover, we also develop a test data set that mimics the approach to develop the test data sets used in prior works in this area. This data set is limited to the proteins that bind proteins (excludes the 112 proteins from Dataset448 that bind other ligands), and where annotations of the protein-binding residues are collected from a single protein–protein complex. To do the latter we randomly pick one complex from the set of complexes with the same protein that we use to transfer annotations of binding residues. This data set is named PBCdataset336 and includes 336 protein-binding proteins that are annotated based on a single protein–protein complex. The PBCdataset336 data set includes 28% fewer protein-binding residues when compared with the PBPdataset336 data set. In other words, transfer of protein-binding annotations from multiple complexes with the same protein increases the number of protein-binding residues by 28%.

Table 5 summarizes the data sets used in this review. These data sets are used to evaluate and compare existing methods and will become a useful resource to validate and compare future methods. The Dataset448 data set is provided the Supplement and includes the protein identifiers, sequences and annotations of protein-, RNA-, DNA- and small-ligand-binding residues. The PBPdataset336 and nPBPdataset112 data sets can be derived from this data set based on the included annotations of ligand-binding residues.

Selection of single-sequence methods that predict protein-binding residues for the comparative assessment

We empirically compare computationally efficient methods that are available as either webservers or source code/downloadable software. This ensures that these methods are accessible to the end users. The criteria to select predictors for inclusion in the empirical assessment are as follows: (1) a working webserver or source code was available as of August 2016 when the predictions were collected; (2) ability to complete prediction of an average length protein sequence with 200 residues within 30 min; and (3) generation of both binary score and numeric propensity for protein binding. The latter is necessary to compute the commonly used measures for the evaluation of predictive quality. Of the original list of 16 methods we exclude ISIS [58] and methods by Du et al. [60], Wang et al. [65] and Geng et al. [69], which lack availability of the webserver or source code. The HomPPI method [64] required prohibitively long runtime. We could not include the two older predictors by Chen et al. [61, 63] because their webservers were no longer maintained at the time of our experiment. Moreover, two methods

Table 5. Summary of the benchmark data sets that are used in this comparative review

Data sets			Dataset448	PBPdataset336	nPBPdataset112	PBCdataset336
Number of proteins			448	336	112	336
Number of protein-binding residues[a]			15 810	15 810	0	11,982
Fraction of protein-binding residues			13.6%	18.6%	0.0%	14.3%
Breakdown of non-protein -binding residues[b] by ligand types	Other-ligand- binding residues[c]	Number of DNA-binding residues	557	320	237	N/A
		Fraction of DNA-binding residues	0.5%	0.4%	0.8%	N/A
		Number of RNA-binding residues	696	444	252	N/A
		Fraction of RNA-binding residues	0.6%	0.5%	0.8%	N/A
		Number of ligand-binding residues	7,175	5,215	1,960	N/A
		Fraction of ligand-binding residues	6.2%	6.1%	6.2%	N/A
	Non-binding residues[d]	Number of non-binding residues	93,857	64,673	29,184	71,713
		Fraction of non-binding residues	80.6%	76.1%	92.5%	85.7%
Total number of residues			116,500	84,941	31,559	83,695

[a]Protein-binding residues bind to proteins.
[b]Non-protein-binding residues do not bind to proteins and they include residues that bind to other molecules and that do not bind to proteins and the other molecules.
[c]Other-ligand-binding residues bind to DNA, RNA or small ligands and they do not bind to proteins.
[d]Non-binding residues do not bind to proteins and the other molecules.

that do not generate propensities, iPPBS-Opt [70] and PPIS [71], were also excluded.

We include seven methods that satisfy the three criteria: SPPIDER [59], PSIVER [62], LORIS [66], SPRINGS [67], CRF-PPI [68], SPRINT [72] and SSWRF [73]. These methods rely on a variety of architectures defined by the use of different input features and different types of predictive models that were computed using different training data sets. Their input features include a number of combinations of features derived directly from the protein sequences and indirectly from the putative relative solvent accessibility. The predictive models they use were generated by several machine learning algorithms, such as the k nearest neighbors [59], naïve Bayes [62], logistic regression [66], neural network [67], random forest [68, 73] and support vector machine [72, 73]. In the nutshell, they cover a broad range of currently available predictors and that their predictions are likely to differ from each other.

Measures of predictive performance

The outputs generated by the sSEQ-to-RES predictors include propensities and binary values. The authors of the 16 predictors use six measures of predictive performance to assess the binary predictions (Table 4). We use the same criteria to evaluate predictions of the seven methods on our benchmark data sets:

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (4)$$

$$F1 - measure = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision} \quad (5)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \quad (6)$$

where TP, TN, FP and FN indicate the number of true positives (correctly predicted protein-binding residues), true negatives (correctly predicted non-protein-binding residues), false positives (non-protein-binding residues incorrectly predicted as protein binding) and false negatives (protein-binding residues incorrectly predicted non-protein-binding residues), respectively. The binary predictions are generated from the propensities using a threshold as follows: residues with putative propensities > threshold are labeled as protein binding and the remaining residues as non-protein-binding residues. To allow for a side-by-side comparison between different predictors, we set the threshold value such that the number of predicted protein-binding residues equals to the number of native protein-binding residues. This way the number of predicted protein-binding residues is correct and more importantly equal between different methods. This ensures that the values of the six criteria can be directly compared between the seven predictors.

We introduce two new measures that provide further insights about the non-protein-binding residues that are predicted as protein binding. The non-protein-binding residues include the other-ligand-binding residues that bind other types of ligands (RNA, DNA and small ligands) as well as the non-binding residues that do not interact with proteins, DNA, RNA and small ligands. We measure the rate of cross-prediction, which is defined as the fraction of the other-ligands-binding residues that are incorrectly predicted as protein binding, and the rate of over-prediction, which quantifies the fraction of the non-binding residues incorrectly predicted as protein binding. Correspondingly, we introduce OPR (over-prediction rate) and CPR (cross-prediction rate):

$$OPR = \frac{FP_{non-binding}}{N_{non-binding}} \quad (7)$$

Table 6. Predictive performance on the Dataset448 data set

Predictor	Year released	Predicted binary values (protein- versus non-protein-binding residues)							Predicted propensities		
		Sensitivity	Specificity	Precision	Accuracy	F1-measure	MCC	CPR	AUC	AULC$_{ratio}$	AUCC
SPPIDER	2007	0.20	0.87	0.19	0.78	0.19	0.06	0.33	0.52	1.69	0.60
PSIVER	2010	0.19	0.87	0.19	0.78	0.19	0.06	0.25	0.57	1.58	0.54
SPRINT	2016	0.19	0.87	0.19	0.78	0.19	0.06	0.38	0.58	1.55	0.66
SPRINGS	2014	0.23	0.88	0.23	0.79	0.23	0.11	0.24	0.62	2.19	0.50
LORIS	2014	0.27	0.89	0.27	0.80	0.27	0.15	0.19	0.65	2.75	0.44
CRF-PPI	2015	0.27	0.89	0.27	0.80	0.27	0.16	0.20	0.67	2.72	0.45
SSWRF	2016	0.32	0.89	0.31	0.82	0.31	0.21	0.20	0.69	3.49	0.39
Random	N/A	0.13	0.86	0.13	0.76	0.13	0.00	0.13	0.50	0.96	0.50

Methods are sorted by their AUC values. CPR is the cross-predicted rate (ratio of other-ligand-binding residues predicted as protein binding). The last row corresponds to a method that predicts binding residues at random. In other words, we assign each residue with a random value of propensity for protein binding. The binary predictions are based on the threshold for which the number of predicted and native protein-binding residues is equal.

$$CPR = \frac{FP_{DNA} + FP_{RNA} + FP_{small\ ligand}}{N_{DNA} + N_{RNA} + N_{small\ ligand}} \qquad (8)$$

where $FP_{nonbinding}$, FP_{DNA}, FP_{RNA} and $FP_{small\ ligand}$ represent the numbers of different types of false positives including the non-binding, DNA-, RNA- and small-ligand-binding residues that are predicted as protein binding; $N_{non-binding}$, N_{DNA}, N_{RNA} and $N_{small\ ligand}$ stand for the number of non-binding, DNA-, RNA- and small-ligand-binding residues. Higher values of OPR and CPR measures mean that the amount of the over-prediction and cross-prediction is higher, and this leads to more incorrect predictions of protein-binding residues. Similar assessment of the cross-prediction was recently performed in the context of the prediction of DNA- and RNA-binding residues [30].

We evaluate the putative propensities with the AUC measure, which was also used by the authors of the sSEQ-to-RES predictors (Table 4). Moreover, we expand this evaluation motivated by the fact that the benchmark data sets are imbalanced. The latter means that the number of protein-binding residues is substantially smaller, by about 7–1 margin, than that of the number of the non-protein-binding residues (Table 5). Given the imbalanced nature of the data sets, even modest values of the false-positive rates (non-protein-binding residues predicted as protein binding) correspond to severe over-prediction of the number of binding residues. Therefore, we introduce a new measure for the evaluation of the putative propensities that focuses on the low range of false-positive rates of the corresponding ROC curve. The AULC (Area Under the Low false positive rate ROC Curve) quantifies the AUC where the number of predicted protein-binding residues is equal or smaller than the number of native protein-binding residues. This means that this score quantifies AUC for the predictions where the number of putative protein-binding residues is not over-predicted. Instead of using the raw values of AULC, which are relatively small and would be difficult to interpret, we compute ratio of AULC for a given predictor to the AULC of a method that predicts binding residues at random (AULCratio). AULCratio = 1 means the prediction from a given sSEQ-to-RES method is equivalent to a random result. AULCratio > 1 indicates a better than random predictor. Such ratio was recently used in a study that evaluates methods that predict disordered flexible linkers using a similarly unbalanced data set [94].

We also propose two new measures of the putative propensities that are motivated by the OPR and CPR criteria. They are analogous to AUC but instead of measuring the AUC defined by the true-positive rates against the false-positive rates, they quantify the area under the curve defined by the OPRs/CPRs against the true-positive rates. The corresponding two measures are named AUOC and AUCC and they quantify the area under the OPR and CPR curves, respectively. Importantly, higher values of AUOC and AUCC correspond to the predictors that more heavily over- and cross-predict protein-binding residues. The values of AUOC and AUCC range between 0 (optimal predictor) and 0.5 (equivalent to a method that predicts binding residues at random). Thus, methods characterized by stronger predictive performance should have low values of these two measures.

Assessment of the predictive performance on Dataset448

We empirically evaluate the single-sequence methods that predict protein-binding residues on the novel Dataset448 data set. This data set includes complete protein sequences (test data sets used to assess predictors in the past rely on fragments of protein chains collected from PDB) with more complete annotations of binding residues (based on mapping of annotations between compatible protein–protein complexes) that cover multiple types of ligands: proteins, DNA, RNA and small ligands. We also include results from a 'random' predictor as a point of reference to assess the existing predictors. The random predictor assigns a random value propensity for each residue. The binary predictions are assigned by selecting a cutoff that ensures that the number of putative binding residues predicted by the random method is equal to the number of native binding residues. This is consistent with the other predictors and ensures that the random results provide the correct number of binding residues.

The ROC curves for considered seven sSEQ-to-RES predictors and the random predictor on the Dataset448 data set are provided in Supplementary Figure S2A. Four of the seven predictors produce AUCs > 0.6, which corresponds to modest levels of predictive performance. All seven methods outperform the random predictor that secures AUC = 0.5. The SSWRF method secures the highest AUC = 0.69, which suggests that this is a fairly accurate predictor. Because the threshold to compute the binary predictions is set to ensure that the number of protein-binding residues predicted by each method equals the number of the native protein-binding residues, results summarized in Table 6 can be used to directly compare different predictors. The SSWRF predictor that has the highest AUC also obtains the highest sensitivity = 0.32. This means that about one of three predicted
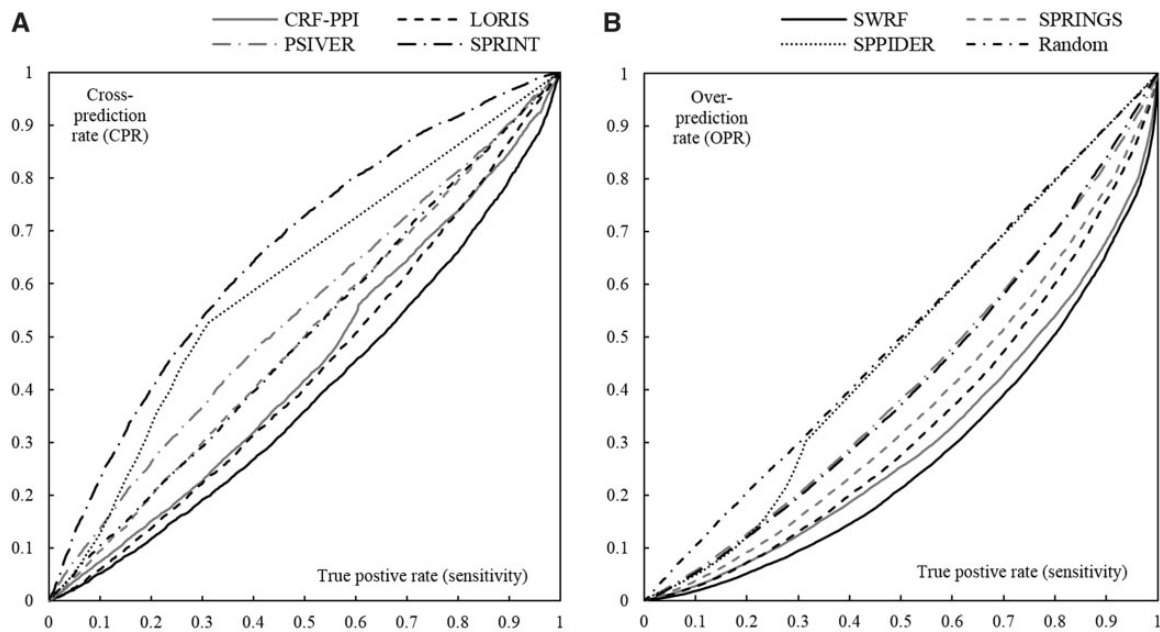
Figure 1. The CPR and OPR curves as a function of sensitivity (fraction of correctly predicted protein-binding residues) based on predictions on the Dataset448 data set. CPR is the fraction of native other-ligands-binding residues incorrectly predicted as protein binding, while OPR is the fraction of native non-binding residues incorrectly predicted as protein binding.
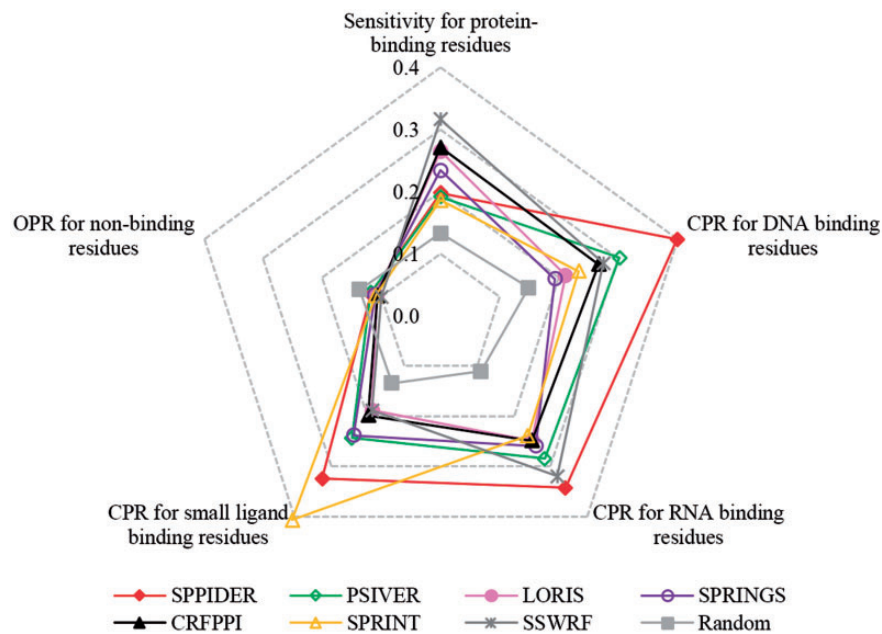

Figure 2. CPRs for the native DNA-, RNA- and small-ligand-binding residues and the corresponding sensitivity for the protein-binding residues and OPR for the non-binding residues based on predictions on the Dataset448 data set.

protein-binding residues generated by this method is correct. This should be considered as an accurate result given that fraction of correctly predicted putative protein-binding residues (sensitivity) is three times higher than the fraction of the non-protein-binding residues incorrectly predicted as binding, i.e. sensitivity = 3*false positive rate = 3*(1 − specificity). The accuracy of SSWRF = 0.82 and MCC = 0.21; the latter reveals a modest level of correlation between the predicted and native binding residues. Overall, three methods secure sensitivity that at least doubles their false positive rate (SSWRF, LORIS and CRF-PPI) and

these methods also obtain the highest specificity, precision, accuracy, F1-measure, MCC and AUC values. The predictive performance for the other four methods is rather modest, with MCC < 0.12 and AUC < 0.63. To compare, the random predictor secures MCC = 0, AUC = 0.5 and accuracy = 0.76. We also calculate the AULCratio, which quantifies how much better is the AUC value of a given predictor for the predictions with low false-positive rate (left side of the ROC curve) from the AUC of a method that makes random predictions. This measure reveals that SSWRF is 3.5 times better that random, and that three other
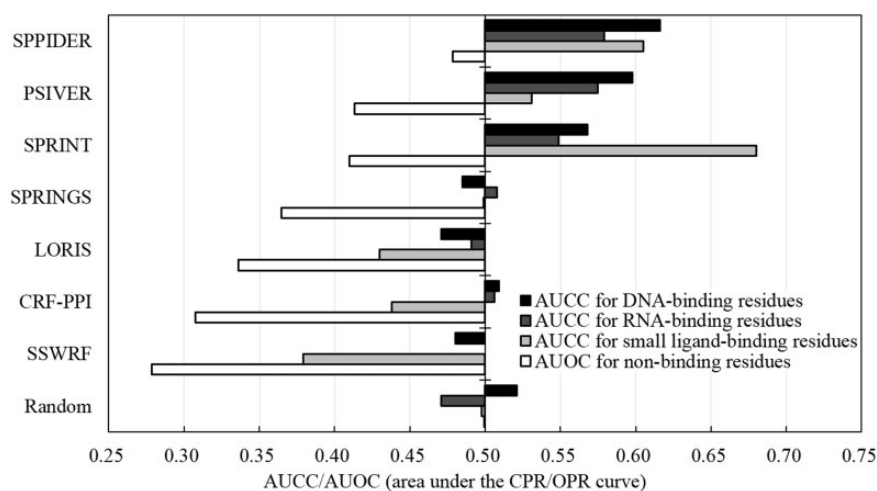
Figure 3. AUCC and AUOC values (*x*-axis) for the native DNA-, RNA-, small-ligand- and non-binding residues based on predictions on the Dataset448 data set. AUOC is the area under the OPR for the native non-binding residues, while AUCC is the area under the CPR for the native DNA-, RNA- and small-ligand-binding residues. A predictor that generates predictions at random is shown at the bottom of the figure and it secures AUOC and AUCC at about 0.5. Values of AUOC < 0.5 (>0.5) and AUCC < 0.5 (>0.5) indicate that a given predictor is better (worse) than random.

methods (CRF-PPI, LORIS and SPRINGS) are at least two times better. Moreover, even the three other less accurate methods are at least 55% better than random. The three best performing methods, which include SSWRF, CRF-PPI and LORIS, are also among the newest, which demonstrates that progress has been made in the recent years.

Assessment of the cross-prediction between other-ligand-binding and protein-binding residues on Dataset448

Besides the evaluation of the overall predictive quality, we are the first to assess the extent of the cross-prediction, defined as incorrect prediction of residues that bind other ligands (DNA, RNA and small ligands) as protein binding. The relatively low sensitivity coupled with low precision and F1-measure (Table 6) suggest high levels of cross-predictions for all considered methods. We quantify that using CPR (defined as the ratio of native other-ligand-binding residues predicted as protein binding) and AUCC; see Table 6. We observe that CPR is higher than sensitivity for SPPIDER, PSIVER, SPRINGS and SPRINT, while the random predictor secures CPR that is equal to its sensitivity. In other words, these four methods predict a higher fraction of the native other-ligand-binding residues as protein binding when compared with the fraction of native protein-binding residues that they predict as protein binding. This means that in fact these four methods predict ligand-binding residues rather than protein-binding residues. The CPR values for SSWRF, CRF-PPI and LORIS are lower than the corresponding sensitivities, which reveals that these methods predict proportionally more protein-binding residues among the native protein-binding residues than among the native other-ligand-binding residues. However, the CPR values of these methods are still relatively high, at about 0.2. They predict 20% of the native other-ligand-binding residues as protein binding compared with between 27% and 32% of the native protein-binding residues predicted as protein binding.

The AUCC values, which assess CPRs across different true-positive rates (fractions of correctly predicted protein-binding residues), tell the same story. The CPR curves shown in Figure 1A show that CPR values are relatively high across the entire spectrum of the true-positive rates and all predictors. Curves of four methods (SPPIDER, PSIVER, SPRINGS and SPRINT) are located above a diagonal that corresponds the results from the random predictor. Correspondingly, their AUCC values are >0.5 (Table 6), which suggests that these methods perform worse than the random predictions. This agrees with our observation that their CPRs are higher than sensitivities. While AUCC values are <0.5 for the other three predictors (SSWRF, CRF-PPI and LORIS), these values that range between 0.39 and 0.45 are relatively poor given that AUCC of the random predictor equals 0.5. The OPR values that quantify fraction of native non-binding residues incorrectly predicted as protein binding are lower than CPRs and the corresponding curves are located well below the diagonal line (Figure 1B). This means that the seven predictors generate proportionally more correctly predicted protein-binding residues than the native non-binding residues incorrectly predicted as protein binding. When taken together, the CPR and OPR curves (Figure 1) convey that the modern sSEQ-to-RES predictors predict ligand-binding residues rather than protein-binding residues. In other words, they accurately discriminate between protein-binding and non-binding residues (OPR curves), but they also confuse protein-binding residues with the residues that bind DNA, RNA and small ligands (CPR curves).

Motivated by these results, we further analyze the cross-predictions for specific types of the other ligands: DNA-, RNA- and small-ligand-binding residues. Figure 2 compares the CPR values for these ligands with the corresponding sensitivity for the native protein-binding residues and OPR for the native non-binding residues. The figure also includes results from the random predictor. A well-performing predictor should have higher sensitivity relative to the values of CPRs and OPR, while the random method has comparable values of CPR, OPR and sensitivity. In general, while the seven methods have high sensitivity and low OPR, their CPR values are high and comparable with the sensitivity. The CPR values for SPPIDER, PSIVER and SPRINGS are equally high for the native DNA-, RNA- and small-ligand-binding residues. The SPRINT method significantly over-predicts protein-binding among the native small-ligand-binding residues and also produces high CPR values for the native DNA- and RNA-binding residues. SSWRF, CRF-PPI and LORIS confuse
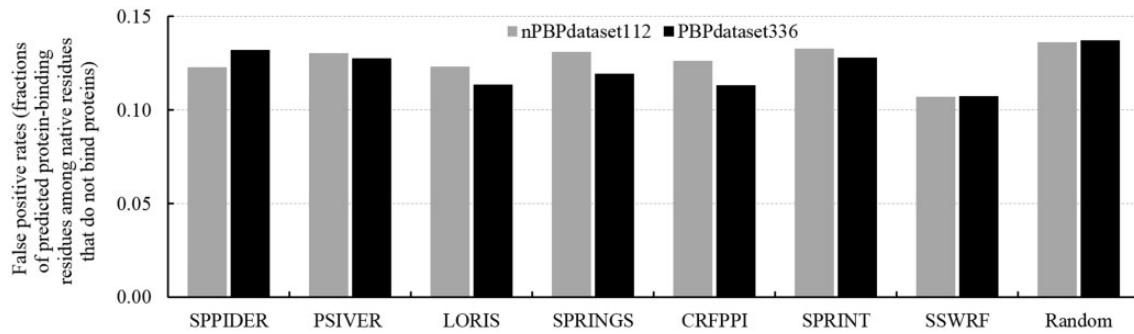
Figure 4. Comparison of fractions of incorrectly predicted protein-binding residues among native residues that do not bind proteins in the nPBPdataset112 and PBPdataset336 data sets. These predictions are based on the threshold for which the number of predicted and native protein-binding residues is equal based on the Dataset448 data set that combines nPBPdataset112 and PBPdataset336. The right-most set of results is for a method that predicts binding residues at random.
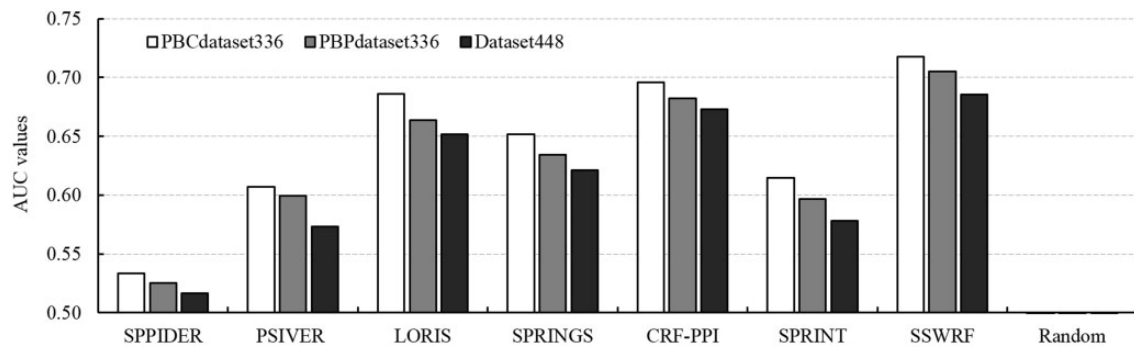

Figure 5. Comparison of the overall predictive performance measured with AUC for the considered seven predictors on the Dataset448, nPBPdataset112 and PBPdataset336 data sets. The right-most set of results is for a method that predicts binding residues at random. Its AUC values are at 0.5 and thus the corresponding bars are not visible.

protein-binding residues with DNA- and RNA-binding residues (high CPR values for the nucleic-acid-binding residues) but they secure reasonable low CPR for the native small-ligand-binding residues. In other words, these three methods can distinguish protein-binding from small-ligand-binding residues, but not from the nucleic-acid-binding residues.

We also analyze the AUCC and AUOC values that quantify the area under the OPR curve for the native non-binding residues and CPR curves for the native DNA-, RNA- and small-ligand-binding residues, respectively (Figure 3). The corresponding CPR and OPR curves are given in the Supplementary Figure S3. The AUCC/AUOC values > 0.5 indicate that a given predictor is worse than random, while AUCC/AUOC < 0.5 means that it is better than random. The white bars in Figure 3 that correspond to the AUOC values show that all seven methods are better than random when predicting native non-binding residues. The light gray bars reveal that SSWRF, CRF-PPI and LORIS produce accurate predictions for the native small-ligand-binding residues. However, these three methods perform poorly (they are equivalent to a random predictor) for the native DNA- and RNA-binding residues. Moreover, SPPIDER, PSIVER, SPRINGS and SPRINT substantially over-predict protein-binding residues among the native DNA-, RNA- and small-ligand-binding residues. Overall, these results agree with the analysis based on the CPR and OPR values from Figure 2.

Overall, our analysis demonstrates that SPPIDER, PSIVER, SPRINGS and SPRINT predict residues that bind proteins, RNA, DNA and small ligands instead of just the protein-binding residues. Namely, these methods predict protein-binding residues at the same or higher rate among the native RNA-, DNA- and

small-ligand-binding residues as among the native protein-binding residues. SSWRF, CRF-PPI and LORIS predict residues that bind proteins, RNA and DNA. In other words, while these three methods relatively accurately separate protein-binding residues from the non-binding and small-ligand-binding residues, they confuse protein-binding and nucleic-acid-binding residues.

Assessment of the predictive performance on proteins that do not interact with proteins from the nPBPdataset112 data set

We empirically observe that the modern sSEQ-to-RES predictors overpredict protein-binding residues. There could be two potential ways for that overprediction. First, these false-positive predictions (incorrectly predicted protein-binding residues among the residues that do not bind proteins) could be in proximity of protein-binding residues and thus they could be predicted as protein binding because these methods use a window in the sequence to make predictions. Second, they overpredict protein-binding residues irrespective of the proximity to the native protein-binding residues. We investigate that by evaluating false-positive rates on the nPBPdataset112 data set that includes proteins that do not have protein-binding residues. We compare these rates with the false-positive rates on the PBPdataset336 data set that includes solely the protein-binding proteins. Figure 4 illustrates that the false-positive rates in the nPBPdataset112 are comparable with the rates on the PBPdataset336 data set across the seven predictors and the random predictor. They range between 0.11 and 0.13 on both data

sets. Given that the predictions were computed such that the number of predicted protein-binding residues equals to the number of native binding residues and because the fraction of native protein-binding residues equals 0.14 (which is why the random method has false-positive rates on both data sets at 0.14), these false-positive rates are rather high. This suggests that the corresponding overprediction of protein-binding residues is not driven by the proximity to native binding residues. Instead, this could be explained by our empirical observation in Figure 2 that shows that these methods do not discriminate between protein- and other-ligand-binding residues. In other words, they substantially cross-predict the residues that bind ligands other than proteins as protein binding. This results in high false-positive rates for proteins that do not have protein-binding residues but that have residues that bind other ligands, which is the case of the proteins in the nPBPdataset112 data set.

Comparison with results from previous studies

Our empirical residues in Table 6 are different from the results that were published in the articles that introduce these predictors. In these articles, SPPIDER, CRF-PPI, SPRINT and SSWRF were reported to obtain AUC values of 0.62, 0.71, 0.71 and 0.71 using their respective test data sets, whereas they secure lower AUC values of 0.52, 0.67, 0.58 and 0.69 on our Dataset448 (Table 6), respectively. The other three methods do not report AUC and it is virtually impossible to compare measures based on the binary predictions given that they depend on the selection of the threshold value. There are three potential reasons for these differences that stem from the use of different test data sets: (1) we use complete protein sequences based on UniProt records instead of potential fragments of protein chains based on PDB records that were used in past studies; (2) following the work in[74] we improve the coverage of the annotations of protein-binding residues by transferring annotations from identical proteins across multiple complexes while the other studies use a single complex; (3) we include proteins that bind other ligands in our test data set to investigate the cross-predictions instead of just the protein-binding proteins like it was done in previous studies.

To verify whether the differences in AUC values are a result of these improvements to the test data set, we create a different version of our test data set that mimics the test data sets from the prior works. The PBCdataset336 data set (Table 5 provides details on this data set) was derived from Dataset448 by (i) removing 112 proteins that do not bind to proteins; (ii) selecting at random a single chain among multiple protein–protein complexes with the same protein and using just this chain to annotate protein-binding residues. We compare the AUC values for the seven considered predictors and the random method on the Dataset448, PBPdataset336 (an intermediate data set that includes only the protein-binding proteins and the complete set of protein-binding annotations) and PBCdataset336 data sets in Figure 5. Complete assessment of predictive performance of these methods on the three data sets is given in Supplementary Table S1 (for the PBPdataset336 and PBCdataset336 data sets) and Table 6 (for the Dataset448 data set). The corresponding ROC curves are provided in Supplementary Figure S2.

We observe a consistent, across the seven methods, trend in the AUC values as we increase similarity between our test data sets and the test data sets from the other works. To compare, as expected the results for the random predictor do not change between the data sets. The AUCs of the seven predictors on Dataset448, which includes full sequences, comprehensive

annotations and a complete set of proteins are the lowest. The AUC on the PBPdataset336 data set, which includes only protein-binding proteins, goes up, and it again increases on the PBCdataset336, which is the most similar to the older test data sets. The relative increase of the AUC between PBCdataset336 and Dataset448 defined as $(AUC_{PBCdataset336} - AUC_{Dataset448})/AUC_{Dataset448}$ ranges between 3.3% and 6.3%. The AUCs on the PBCdataset336 data set that imitates the test data sets from the articles that introduce these predictors are similar to the previously reported AUCs, i.e. we obtain 0.70 versus 0.71 reported in [67] for CRF-PPI; we measure 0.72 versus 0.71 reported in [73] for SSWRF. Our AUC for SPRINT that equals 0.61 is lower than the 0.71 reported in [72]. The likely reason is that SPRINT was designed to predict protein–peptide interactions, which are a subset of the protein–protein interactions that we evaluate. Also, the test data set used to evaluate SPRINT shared higher similarity to their training data set at up to 30% compared with our data sets that share up to 25% similarity (Table 4). This is in contrast to the test data set used to assess CRF-PPI and SSWRF, which rely on the same similarity of 25%. Finally, we measure AUC = 0.53 for SPPIDER, which is lower than 0.62 reported by the authors of this method [59]. However, 0.62 is also a low value and the authors of SPPIDER used the test data set that shares much higher sequence similarity with their training proteins at up to 50% (Table 4) compared with our data set that shares up to 25% similarity with the proteins from their training data set. This may explain why our estimate of predictive performance is lower.

Overall, this experiment suggests that our benchmark test data set provides reliable estimates of predictive performance. We observe that the predictive quality of the considered methods that we measured is comparable with that assessed by the authors when compatible data sets are used. Importantly, we also note that the predictive quality drops down when we consider full protein chains and a more complete set of transferred annotations of protein-binding residues. We hypothesize that the reason for this is that the current predictors were built on training data sets that make the same assumptions as the older test data sets by using fragments of protein chains and incomplete annotations of binding.

Summary and conclusions

Accurate identification of protein-binding residues is essential to improve our understanding of molecular mechanisms that govern protein–protein interactions and to improve protein–protein docking studies. Recent years have witnessed the development of a large number of computational methods that predict protein–protein interactions. Previous reviews of these methods mainly focused on the structure-based methods, while paying little attention to the many sequence-based methods. The influx of the sequence-based methods in past 3 years motivates this first-of-its-kind study in which we comprehensively review and empirically evaluate sequence-based methods for the prediction of protein–protein interactions.

We categorize the sequence-based methods into three groups according into their inputs and outputs: the 'pSEQ-to-PRO' methods that predict whether a given pair of sequences interacts, the 'pSEQ-to-RES' techniques that predict protein-binding residues for a pair of input protein sequences and the 'sSEQ-to-RES' methods that predict protein-binding residues in a single input protein chain. We focus our review and empirical evaluation on the 'sSEQ-to-RES' predictors because they provide more detailed residue-level annotations and can be applied to

all protein sequence, without the need to know the pairs of protein partners. We review the architectures of these methods, discuss their inputs and outputs, summarize how they were assessed and comment on their availability.

We also perform a comprehensive empirical comparison of representative seven sSEQ-to-RES methods that are computationally efficient and available to the end users as either webservers or source code. We have developed a high-quality and large benchmark data set that is characterized by the more complete annotation of protein-binding residues and which includes annotations of residues that bind to other ligands. We share this data set with the community to facilitate future comparative studies (see Supplement). Our empirical analysis demonstrates that the selected predictors perform well in discriminating protein-binding residues from non-binding residues. Their overall AUC values range from 0.52 to 0.69 and they all outperform the random predictor. We found that more recent methods have higher predictive performance than the older method, with the newest SSWRF that obtains the highest AUC. Given that we set the number of predicted protein-binding residues equal to the number of native ones protein-binding residues, SSWRF yields sensitivity $= 32\%$ and specificity $= 89\%$. This means that it correctly identifies 32% native protein-binding residues and 89% of native non-protein-binding residues. These results show that progress has been made in this field in the recent years. We hypothesize that this progress is owing to the use of more informative features to encode input residues in the recently designed predictors.

However, we found that these predictors incorrectly cross-predict many residues that bind other ligands as protein-binding residues. We investigate this cross-prediction bias for each predictor and across different types of ligands. For instance, we uncover that when the number of predicted and native protein-binding residues is equal, the best predictor SSWRF cross-predicts 28% DNA-binding residues, 32% RNA-binding and 19% ligand-binding residues as protein binding. When compared with the sensitivity of this predictor, which equals 32%, this reveals that SSWRF predicts as many binding residues among the native protein-binding residues as among the native nucleic-acid-binding residues. Overall, we conclude that four methods SPPIDER, PSIVER, SPRINGS and SPRINT predict residues that bind proteins, RNA, DNA and small ligands instead of just the protein-binding residues; their CPRs for these types of ligands are comparable or higher than their sensitivity. The other three methods, SSWRF, CRF-PPI and LORIS, predict residues that bind proteins, RNA and DNA; their CPRs for nucleic acids are similar to their sensitivity.

Furthermore, we also investigate the source of these cross-predictions. Our empirical analysis shows similar rates of cross-predictions among protein-binding proteins and proteins that do not have protein-binding residues. Thus, we conclude that cross-predictions are not driven by the proximity to the native protein-binding residues, which could be influential owing to the use of the sliding windows by the sSEQ-to-RES predictors. Instead, our results suggest that these methods confuse the protein-binding residues with residues that bind the other ligands. We hypothesize that this is because these predictors do not use a sufficiently rich set of inputs and because they use biased training data sets. Their inputs focus on the sequence conservation and solvent accessibility as means to separate protein-binding from non-protein-binding residues (Table 4). While protein-binding residues are more solvent exposed and conserved than non-binding residues [95], the same is true for other ligands, such as nucleic acids [96]. Thus, these two factors

would predict both protein-binding and nucleic-acid-binding residues. Their training data sets are solely focused on the protein-binding proteins that include a relatively large number of protein-binding residues and relatively few residues that bind other ligands. This way, the predictive models derived from these data sets cannot be properly optimized to discriminate protein-binding from other-ligands-binding residues.

Our new benchmark data set presents a bigger challenge than the previously used test data sets. The empirically evaluated predictive performance of selected methods is lower on this data set compared with the results reported by the authors. The differences likely stem from the fact that the training data sets used to build these methods use fragments of protein sequences and incomplete annotations of protein-binding residues when compared with our data set. We demonstrate that our results are in agreement with the reported predictive performance when our data set is scaled back to the format of the older test data sets.

Our study prompts five recommendations. First, a new generation of more accurate sSEQ-to-RES predictors is needed. These predictors should not only separate the protein-binding residues from the non-binding residues but, most importantly, also from residues that bind the other ligands. The authors of such studies are urged to compute CPR, OPR, AUCC and AUOC values to quantify the extent of the ability of their method to satisfy this objective. Second, the currently used annotations of protein-binding residues should be extended by transferring annotations across the same proteins in multiple protein–protein complexes. This will improve completeness of data that are used to both build and validate the predictors. Third, the authors of the sequence-based predictors of protein–protein interactions should be required to make their methods publicly available, preferably as both webservers and standalone applications, and to maintain this availability over an extended period of time. Of the 44 methods that we review, 16 are unavailable and another 11 are no longer maintained, which totals to >60% of the published methods that are not accessible to the end users. Fourth, standard benchmark data sets should be periodically compiled and made available. This will facilitate evaluation and comparative analysis of the predictive performance of the existing and new methods. We start this initiative with the inclusion of our benchmark data set in the Supplement to this article. Fifth, the current methods predict protein-binding residues, but these residues are not grouped into specific sites of interaction on the protein surface (binding sites). An ability to group the predicted binding residues into binding sites would be particularly relevant for proteins that interact with multiple protein partners in multiple sites. Such clustering of putative binding residues was performed in the context of prediction of several small ligand types including nucleotides, metal ions and heme group [82, 87]. The authors have used putative structure predicted from the protein sequence to spatially cluster the predicted binding residues into the corresponding binding sites.

Key Points

- The article reviews >40 sequence-based predictors of protein–protein interactions, with focus on 16 methods that predict protein-binding residues from a single sequence.
- Empirical results demonstrate that current predictors accurately discriminate protein binding from non-

binding residues, but they also incorrectly cross-predict a large number of DNA-, RNA- and small-ligand-binding residues as protein binding.

• The cross-predictions are driven by the inability of the predictors to separate protein-binding and other-ligand-binding residues rather than a proximity to the native protein-binding residues.

• New data sets in this field should include more complete annotations of protein-binding residues and a larger number of nucleic acids and small-ligand-binding residues and should be mapped into the full protein sequences.

• A new generation of accurate predictors that use the improved data sets and that use novel predictive inputs and architectures to reduce the cross-predictions are needed.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

References

1. Ding XM, Pan XY, Xu C, *et al*. Computational prediction of DNA-protein interactions: a review. *Curr Comput Aided Drug Des* 2010;**6**:197–206.

2. Chen K, Kurgan L. Investigation of atomic level patterns in protein–small ligand interactions. *PLoS One* 2009;**4**:e4473.

3. Sudha G, Nussinov R, Srinivasan N. An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. *Prog Biophys Mol Biol* 2014;**116**:141–50.

4. Fornes O, Garcia-Garcia J, Bonet J, *et al*. On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. *Adv Protein Chem Struct Biol* 2014;**94**:77–120.

5. Sperandio O. Editorial: toward the design of drugs on protein-protein interactions. *Curr Pharm Des* 2012;**18**:4585.

6. Petta I, Lievens S, Libert C, *et al*. Modulation of protein-protein interactions for the development of novel therapeutics. *Mol Ther* 2016;**24**:707–18.

7. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 2007;**450**:1001–9.

8. Orii N, Ganapathiraju MK. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. *PLoS One* 2012;**7**:e49029.

9. Kuzmanov U, Emili A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Med* 2013;**5**:37.

10. Nibbe RK, Chowdhury SA, Koyuturk M, *et al*. Protein-protein interaction networks and subnetworks in the biology of disease. *Wiley Interdiscip Rev Syst Biol Med* 2011;**3**:357–67.

11. De Las Rivas J, Fontanillo C. Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief Funct Genomics* 2012;**11**:489–96.

12. Calderone A, Castagnoli L, Cesareni G. Mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods* 2013;**10**:690–1.

13. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013;**41**:D1096–103.

14. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.

15. Patil A, Kinoshita K, Nakamura H. Hub promiscuity in protein-protein interaction networks. *Int J Mol Sci* 2010;**11**:1930–43.

16. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.

17. Ezkurdia I, Bartoli L, Fariselli P, *et al*. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform* 2009;**10**:233–46.

18. Fernández-Recio J. Prediction of protein binding sites and hot spots. *Wiley Interdiscip Rev Comput Mol Sci* 2011;**1**:680–98.

19. Aumentado-Armstrong TT, Istrate B, Murgita RA. Algorithmic approaches to protein-protein interaction site prediction. *Algorithms Mol Biol* 2015;**10**:7.

20. Xue LC, Dobbs D, Bonvin AM, *et al*. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett* 2015;**589**:3516–26.

21. Esmaielbeiki R, Krawczyk K, Knapp B, *et al*. Progress and challenges in predicting protein interfaces. *Brief Bioinform* 2016;**17**:117–31.

22. Maheshwari S, Brylinski M. Predicting protein interface residues using easily accessible on-line resources. *Brief Bioinform* 2015;**16**:1025–34.

23. Vreven T, Hwang H, Pierce BG, *et al*. Evaluating template-based and template-free protein-protein complex structure prediction. *Brief Bioinform* 2014;**15**:169–76.

24. Huang SY. Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov Today* 2014;**19**:1081–96.

25. Ritchie DW. Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci* 2008;**9**:1–15.

26. Vreven T, Moal IH, Vangone A, *et al*. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J Mol Biol* 2015;**427**:3031–41.

27. Rodrigues JP, Bonvin AM. Integrative computational modeling of protein interactions. *FEBS J* 2014;**281**:1988–2003.

28. Kundrotas PJ, Vakser IA. Accuracy of protein-protein binding sites in high-throughput template-based modeling. *PLoS Comput Biol* 2010;**6**:e1000727.

29. Mukherjee S, Zhang Y. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* 2011;**19**:955–66.

30. Shen J, Zhang J, Luo X, *et al*. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA* 2007;**104**:4337–41.

31. Guo Y, Yu L, Wen Z, *et al*. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* 2008;**36**:3025–30.

32. Yu CY, Chou LC, Chang DT. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* 2010;**11**:167.

33. Xia JF, Zhao XM, Huang DS. Predicting protein-protein interactions from protein sequences using meta predictor. *Amino Acids* 2010;**39**:1595–9.

34. Guo Y, Li M, Pu X, *et al*. PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. *BMC Res Notes* 2010;**3**:145.

35. Yu J, Guo M, Needham CJ, *et al*. Simple sequence-based kernels do not predict protein-protein interactions. *Bioinformatics* 2010;**26**:2610–4.

36. Zhang YN, Pan XY, Huang Y, *et al*. Adaptive compressive learning for prediction of protein-protein interactions from primary sequence. *J Theor Biol* 2011;**283**:44–52.

37. Liu X, Liu B, Huang Z, *et al*. SPPS: a sequence-based method for predicting probability of protein-protein interaction partners. *PLoS One* 2012;**7**:e30938.

38. Ahmad S, Mizuguchi K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS One* 2011;**6**:e29104.

39. Yousef A, Moghadam Charkari N. A novel method based on new adaptive LVQ neural network for predicting protein-protein interactions from protein sequences. *J Theor Biol* 2013;**336**:231–9.

40. Zahiri J, Yaghoubi O, Mohammad-Noori M, *et al*. PPIevo: protein-protein interaction prediction from PSSM based evolutionary information. *Genomics* 2013;**102**:237–42.

41. You ZH, Lei YK, Zhu L, *et al*. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics* 2013;**14 (Suppl 8)**:S10.

42. You Z-H, Zhu L, Zheng C-H, *et al*. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics* 2014;**15**:S9.

43. You Z-H, Li J, Gao X, *et al*. Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. *Biomed Res Int* 2015;**2015**:867516.

44. Hu L, Chan KC. Discovering variable-length patterns in protein sequences for protein-protein interaction prediction. *IEEE Trans Nanobiosci* 2015;**14**:409–16.

45. Hamp T, Rost B. Evolutionary profiles improve protein-protein interaction prediction from sequence. *Bioinformatics* 2015;**31**:1945–50.

46. You Z-H, Chan KC, Hu P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLoS One* 2015;**10**:e0125811.

47. Jia J, Liu Z, Chen X, *et al*. Prediction of protein-protein interactions using chaos game representation and wavelet transform via the random forest algorithm. *Genetics and Molecular Research* 2015;**14**:11791–805.

48. Huang Y-A, You Z-H, Gao X, *et al*. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. *Biomed Res Int* 2015;**2015**:902198.

49. Gao Z-G, Wang L, Xia S-X, *et al*. Ens-PPI: a novel ensemble classifier for predicting the interactions of proteins using auto covariance transformation from PSSM. *Biomed Res Int* 2016;**2016**:456524.

50. Sze-To A, Fung S, Lee E-SA, *et al*. Prediction of protein–protein interaction via co-occurring aligned pattern clusters. *Methods* 2016;**110**:26–34.

51. Huang YA, You ZH, Chen X, *et al*. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinformatics* 2016;**17**:184.

52. An J-Y, Meng F-R, You Z-H, *et al*. Using the relevance vector machine model combined with local phase quantization to predict protein-protein interactions from protein sequences. *Biomed Res Int* 2016;**2016**:4783801.

53. Pitre S, Dehne F, Chan A, *et al*. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics* 2006;**7**:365.

54. Shi MG, Xia JF, Li XL, *et al*. Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. *Amino Acids* 2010;**38**:891–9.

55. Chang DT, Syu YT, Lin PC. Predicting the protein-protein interactions using primary structures with predicted protein surface. *BMC Bioinformatics* 2010;**11 (Suppl 1)**:S3.

56. Amos-Binks A, Patulea C, Pitre S, *et al*. Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences. *BMC Bioinformatics* 2011;**12**:225.

57. Xia B, Zhang H, Li Q, *et al*. PETs: a stable and accurate predictor of protein-protein interacting sites based on extremely-randomized trees. *IEEE Tran Nanobiosci* 2015;**14**:882–93.

58. Ofran Y, Rost B. ISIS: interaction sites identified from sequence. *Bioinformatics* 2007;**23**:e13–16.

59. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. *Proteins* 2007;**66**:630–45.

60. Du X, Cheng J, Song J. Improved prediction of protein binding sites from sequences using genetic algorithm. *Protein J* 2009;**28**:273–80.

61. Chen XW, Jeong JC. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* 2009;**25**:585–91.

62. Murakami Y, Mizuguchi K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* 2010;**26**:1841–8.

63. Chen P, Li J. Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC Bioinformatics* 2010;**11**:402.

64. Xue LC, Dobbs D, Honavar V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics* 2011;**12**:244.

65. Wang DD, Wang R, Yan H. Fast prediction of protein–protein interaction sites based on extreme learning machines. *Neurocomputing* 2014;**128**:258–66.

66. Dhole K, Singh G, Pai PP, *et al*. Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. *J Theor Biol* 2014;**348**:47–54.

67. Singh G, Dhole K, Pai PP, *et al*. SPRINGS: prediction of protein-protein interaction sites using artificial neural networks. *PeerJ PrePrints* 2014:e266v2.

68. Wei Z-S, Yang J-Y, Shen H-B, *et al*. A cascade random forests algorithm for predicting protein-protein interaction sites. *IEEE Trans Nanobiosci* 2015;**14**:746–60.

69. Geng H, Lu T, Lin X, *et al*. Prediction of protein-protein interaction sites based on Naive Bayes classifier. *Biochem Res Int* 2015;**2015**:978193.

70. Jia J, Liu Z, Xiao X, *et al*. iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules* 2016;**21**:95.

71. Liu G-H, Shen H-B, Yu D-J. Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures. *J Membr Biol* 2016;**249**:141–53.

72. Taherzadeh G, Yang Y, Zhang T, *et al*. Sequence-based prediction of protein–peptide binding sites using support vector machine. *J Comput Chem* 2016;**37**:1223–9.

73. Wei Z-S, Han K, Yang J-Y, *et al*. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing* 2016;**193**:201–12.

74. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. *Brief Bioinform* 2016;**17**:88–105.

75. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* 2015;**43**:e121.

76. Nagarajan R, Ahmad S, Gromiha MM. Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res* 2013;**41**:7606–14.

77. Puton T, Kozlowski L, Tuszynska I, *et al*. Computational methods for prediction of protein-RNA interactions. *J Struct Biol* 2012;**179**:261–8.

78. Walia RR, Caragea C, Lewis BA, *et al*. Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics* 2012;**13**

79. Zhang T, Zhang H, Chen K, *et al*. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr Protein Pept Sci* 2010;**11**:609–28.

80. Roche DB, Brackenridge DA, McGuffin LJ. Proteins and their interacting partners: an introduction to protein-ligand binding site prediction methods. *Int J Mol Sci* 2015;**16**:29829–42.

81. Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* 2012;**28**:331–41.

82. Yu DJ, Hu J, Huang Y, *et al*. TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble. *J Comput Chem* 2013;**34**:974–85.

83. Passerini A, Lippi M, Frasconi P. Predicting metal-binding sites from protein sequence. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**:203–13.

84. Yu DJ, Hu J, Yan H, *et al*. Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble. *BMC Bioinformatics* 2014;**15**:297.

85. Panwar B, Gupta S, Raghava GPS. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC Bioinformatics* 2013;**14**:44.

86. Horst JA, Samudrala R. A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognit Lett* 2010;**31**:2103–12.

87. Yu DJ, Hu J, Yang J, *et al*. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**:994–1008.

88. Joo K, Lee SJ, Lee J. Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins* 2012;**80**:1791–7.

89. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;**16**:404–5.

90. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

91. Burges CJ. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 1998;**2**:121–67.

92. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.

93. Kurgan L, Disfani FM. Structural protein descriptors in 1-dimension and their sequence-based predictions. *Curr Protein Pept Sci* 2011;**12**:470–89.

94. Meng F, Kurgan L. DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* 2016;**32**:i341–50.

95. Caffrey DR, Somaroo S, Hughes JD, *et al*. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 2004;**13**:190–202.

96. Luscombe NM, Thornton JM. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* 2002;**320**:991–1009.