Simple sequence-based kernels do not predict protein-protein interactions

Jiantao Yu$^{1,2}$, Maozu Guo$^{1,*}$, Chris J. Needham$^3$, Yangchao Huang$^1$, Lu Cai$^4$ and David R. Westhead$^{2,*}$

$^1$School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
$^2$Institute of Molecular and Cellular Biology, University of Leeds, Leeds, LS2 9JT, UK
$^3$School of Computing, University of Leeds, Leeds, LS2 9JT, UK
$^4$School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou, Inner Mongolia 014010, China

Associate Editor: Prof. Martin Bishop

ABSTRACT

**Motivation:** A number of methods have been reported that predict protein-protein interactions with high accuracy using only simple sequence-based features such as amino-acid 3-mer content. This is surprising, given that many protein interactions have high specificity that depends on detailed atomic recognition between physico-chemically complementary surfaces. Are the reported high accuracies realistic?

**Results:** We find that the reported accuracies of the predictions are significantly over-estimated, and strongly dependent on the structure of the training and testing datasets used. The choice of which protein pairs are deemed as non-interactions in the training data has a variable impact on the accuracy estimates, and the accuracies can be artificially inflated by a bias towards dominant samples in the positive data which result from the presence of hub proteins in the protein interaction network. To address this bias we propose a positive-set-specific method to create a 'balanced' negative set maintaining the degree distribution for each protein, leading to the conclusion that simple sequence-based features contain insufficient information to be useful for predicting protein-protein interactions, but that protein domain-based features have some predictive value.

**Availability:** Our method, named 'BRS-nonint', is available at http://www.bioinformatics.leeds.ac.uk/BRS-nonint/.

**Contact:** D.R.Westhead@leeds.ac.uk, maozuguo@hit.edu.cn

**Supporting Data:** All the data sets used in this study are derived from publicly available data, and are available at http://www.bioinformatics.leeds.ac.uk/BRS-nonint/PPI_RandomBalance.html.

1 INTRODUCTION

Protein-protein interactions are responsible for many critical functions and processes in biology, and are highly relevant to disease states. While experimental measurements are able to discover protein interactions at high-throughput, they can be biased to specific types of interaction or unlikely to discover all interactions, and cannot be applied to all relevant species. As a result, a number of researchers have studied how well machine learning methods are able to predict protein-protein interactions (PPIs) from protein sequence information alone (Ben-Hur and Noble, 2005; Bock and Gough, 2001; Chou and Cai, 2006; Gomez et al., 2003; Guo et al., 2008, 2010; Martin et al., 2005; Nanni and Lumini, 2006; Park, 2009; Pitre et al., 2006, 2008; Roy et al., 2009; Shen et al., 2007; Sprinzak and Margalit, 2001; Yu et al., 2010). They have reported impressive performance using a variety of simple sequence features (such as counts of 3-mers of neighbouring residues) and developed sophisticated pairwise kernels to allow classification using support vector machines (SVMs). The performance measures reported are much better than one would expect, and we have found that they are very sensitive to the content of the data sets used in training and cross-validation. In particular, the presence of ‘hub’ proteins that interact with many other proteins in the positive data set leads to a strong bias that invalidates most performance estimates.

Recently, Park (2009) has benchmarked four methods on yeast and human data, found significant performance differences and highlighted some effects of training and testing data on performance estimates. However, the datasets used in this benchmark still suffer from bias of data set content reported here. In this work we present a method to remove this bias and demonstrate the effect on predicting PPIs from simple sequence features: that simple sequence-based kernels do not predict protein-protein interactions. We now summarise some approaches taken to predict PPI from sequence data, all of which should be reinterpreted following our findings. Ben-Hur and Noble (2005) used a support vector machine method with a pairwise kernel and evaluated a number of sequence features, including the 'spectrum kernel' where the protein sequences are represented by the counts of 3-mers of neighbouring residues (Leslie et al., 2002). This leads to a fixed length (n=8,000) feature vector for each protein regardless of the sequence length. The pairwise kernel allows comparison of pairs of proteins, avoiding any problems that may be introduced by concatenating feature
vectors for each protein, and on a yeast dataset performance of the SVM gives a Receiver-Operator Characteristic (ROC) area under curve (AUC) score of 0.81. Similarly Martin et al. (2005) used symmetric 3-mers (n=4,200) and a pair-wise kernel giving accuracy of 70.3% on human data (and 69.0% for yeast). Accuracy of 83.9% was reported to be achieved by Shen et al. (2007) on human PPI data using 3-mers of residues grouped into 7 categories (n=343) and an S-kernel similar in function to the previous pair-wise kernels. Guo et al. (2008; 2010) represented residues in 7 categories with an auto-covariance method and reported high accuracies using shuffled sequences as a negative set, but poor results on randomly sampled negative data. Roy et al. (2009) proposed using simple amino acid composition features (normalized counts of single or pairs of amino acids) and reported results on a par with those obtained using protein domain information.

There exist a number of data sets of protein pairs which have been determined to interact by high-throughput methods (Costanzo et al., 2010; Gavin et al., 2006; Ito et al., 2001; Stelzl et al., 2005; Yu et al., 2008) and can be used as positive data to train and validate prediction methods. In particular, many experiments have been conducted on yeast, and a PPI network consisting of 156,673 non-redundant interactions exists in BioGRID (Stark et al., 2006), and a high-confidence dataset of 6,568 non-redundant interactions has been published (Batada et al., 2007). For human, the Human Protein References Database (HPRD) contains 38,945 non-redundant high quality interactions (Peri et al., 2004), and the BioGRID database contains 30,851 non-redundant interactions (Stark et al., 2006). The choice of a set of negative examples (non-interacting protein pairs) to use in training and validating prediction methods is particularly important since few techniques can conduct a large-scale measurement of non-interacting pairs (Smialowski et al., 2010; Doerr, 2010). The latest negative dataset, 

2 METHODS

2.1 Datasets

The datasets of interactions were processed to filter out redundant samples, self-interactions, and proteins whose sequence contained elements other than the 20 standard amino acids. For yeast, 154,828 non-redundant positive interactions (47,335 physical interactions and remainder genetic) were extracted from the BioGRID database (Stark et al., 2006) (3.0.64.tab1). For human, 36,134 non-redundant positive physical interactions were extracted from the HPRD (Peri et al., 2004) (Release9_041310), which was used in (Shen et al., 2007). In addition 27,307 human interactions extracted from the BioGRID database (3.0.64.tab1) were used to filter ‘potential interactions’ in creating negative candidates (see below) (16,993 PPIs overlap between these two sets). Sequence data used for yeast was from the Saccharomyces Genome Database (http://download.yeastgenome.org/) (file orf_trans.fasta), and for human was from UniProt (The UniProt Con-

sorium, 2010) (file uniprot_sprot.fasta). For evaluating Pfam features, version 24.0 of the Pfam Database (Finn et al., 2010) (pfam_A.hmm) was used. HMMER3.0 (http://hmmer.org) was used to scan the protein sequences. Smaller datasets were constructed to test the performance of Pfam features. For yeast, we used the high confidence physical interaction dataset from (Batada et al., 2007) (HC-BIOGRID-2.0.31.tab) and filtered out all interactions that involved a protein with no Pfam domain hits, resulting in a dataset containing 5,621 interactions. For human, we took the intersection of the two human datasets described above, and again filtered out all interactions that involved a protein with no Pfam hits, resulting in a dataset containing 15,804 physical interactions. We consider a ‘Pfam hit’ when the match has an E-value (full sequence) of less than 0.01, similar to the methods of Gomez et al. (2003).

2.2 Choosing a subset

Owing to the computational demand of the large data sets, we chose a subset of the interactions. Subsets were formed by choosing at random 300 proteins for yeast or 1500 for human, and then selecting all interactions involving any of those proteins. This creates an interaction data set that maintains the vertex degree distribution of the large data set. Typically subsets contained ~15,000 interactions between ~4,000 proteins for yeast and ~10,000 interactions between ~5,300 proteins for human. We created ten subsets at random and report average performances of 10-fold cross-validations (10x10cv) in each of these subsets.

In order to test stability of performance measures on subsets of different sizes, and also to examine the effect of excluding genetic interactions, we sampled four subsets of 5,000-20,000 interactions from the yeast physical interactions in the BioGRID database (Stark et al., 2006) (3.0.64.tab2). 10 cross-validations were performed on each subset respectively.

2.3 Negative data set construction

From the positive set of N interactions, the complement graph can be formed as the set of all possible pairs of proteins in the positive set for which a positive interaction is not present. Since we created subsets of the whole PPI network, we formed the negative candidate graph from the subset’s complement graph. Interactions that exist in the BioGRID database (Stark et al., 2006) were removed from the negative candidate graphs to reduce the potential for real interactions appearing erroneously in the negative sets. Self-interactions were also excluded. For simple random sampling, N edges were chosen from the negative candidate graph, ensuring that each protein that appeared in the positive subset also appeared at least once in the negative dataset. Each protein was taken in turn and a connected edge randomly sampled from the negative candidate graph and added to the negative set until all N edges were chosen. For balanced random sampling, the number of times each protein appears in the negative set is equal to the number of times it appears in the positive set. Taking proteins in the positive data set in turn, beginning with the protein of largest vertex degree, we randomly sampled edges from the negative candidate graph connected to the current protein until the vertex degree in the negative set was equal to that in the positive set.

The balanced random sampling method is available in the BRS-nonint software. Input to this software is the positive interaction dataset, as well as the pairs which should be avoided in choosing negative data. All of interactions (one pair per line) should be input with a format of ‘proteinA proteinB interacting’.

2.4 SVMs and kernel functions

The methods of Ben-Hur and Noble (2005) and Shen et al. (2007) were evaluated, and the SVMs were implemented by modifying the code of libsvm-2.91 (http://www.csie.ntu.edu.tw/~cjlin/libsvm). From (Ben-Hur and Noble, 2005), the TPKS800 kernel represents each protein as a length
8,000 feature vector \(v_i\) of normalised counts (normalised to sum to one) of each possible triple of amino acids (the 3-mer spectrum kernel). The kernel takes cosine form, \(k(X, Y) = \langle v_X, v_Y \rangle / \sqrt{\langle v_X, v_X \rangle \langle v_Y, v_Y \rangle}\), for proteins \(X\) and \(Y\) with feature vectors \(v_X\) and \(v_Y\), respectively, and the pairwise kernel for proteins pairs (A-B and C-D) is calculated as a tensor product, \(K_{\text{S-kernel}}((A, B), (C, D)) = k(A, C) / k(B, D) + k(A, D) / k(B, C)\). From (Shen et al., 2007), the S-kernel represents each protein as a length 343 feature vector \(v_i\) of normalised counts (normalised by the largest) of each possible conjoint triad – 3-mers of residues grouped into 7 categories. Squared Euclidean distances between proteins are calculated, \(d(X, Y) = \|v_X - v_Y\|^2\) and the pairwise S-kernel is \(K_{\text{S}}(A, B, C, D) = \exp(-\gamma \|d(A, C) + d(B, D) + d(A, D) + d(B, C)\|)\), SVM parameters \(C = 128\) and \(\gamma = 0.25\) were used as in (Shen et al., 2007). For the Pfam kernel, a feature vector of counts of Pfam hits (see above) was created, and the TPPK pairwise kernel was used on these feature vectors (length 1,932 and 3,400 for yeast and human respectively, as this many distinct Pfam features were detected). Similar results were obtained using other representations of the Pfam features (such as scores, E-values or binary presence/absence of domains).

2.5 Performance evaluation
Performances were reported by the AUC values (area under the ROC curve, a perfect classifier has AUC=1, random performance gives AUC=0.5) with 10-fold cross-validation. Variation in performance between ten subsets or ten cross-validations was quantified as ‘average±range’. Similar to ROC in (Ben-Hur and Noble, 2005), we also report ROC(0.5%) (the area under the ROC curve for false positive rate ≤0.5%), which measures just the true positives detected before 0.5% of false positives.

3 RESULTS
The structures of the various data sets are illustrated in Figure 1, which clearly shows the hub proteins in the positive set, reflecting the well-known power-law distribution of vertex degrees, and how this is replicated in the negative set generated by the method of balanced random sampling but not simple random sampling.

3.1 Prediction performance and dataset bias
SVM performance results are reported in Table 1. The TPPK kernel (Ben-Hur and Noble, 2005) predicted PPIs with an AUC of 0.95 on the yeast dataset, when trained and tested with a negative set constructed by simple random sampling. We also reproduced the reported results of 0.81 AUC on the set of the 10,517 yeast interactions used in (Ben-Hur and Noble, 2005). However, when a balanced negative dataset was used, the performance fell to 0.5 AUC, indicating that performance was close to random (AUC=0.46 on the same dataset of Ben-Hur and Noble using balanced sampling). This highlights the bias inherent in the structure of the datasets constructed using a random sampling approach that does not account for the number of times each protein appears in the positive and negative datasets. This is also confirmed by similar results for predicting PPIs on human datasets (0.83 AUC vs 0.55 AUC on balanced dataset), again indicating the bias, and revealing the poor performance of the classifier based on these simple sequence features.

Using the S-kernel\(_{343}\) we obtained performance of 0.67 AUC on human data with a negative set generated by simple random sampling and our own implementation of the methods of Shen et al. (2007). Using a negative set constructed by balanced sampling this was reduced to 0.55 AUC. On yeast data the AUC score dropped from 0.72 to 0.53 when the balanced negative set was used.

The data in Table 2 show the effect of using different subset sizes, and restricting the yeast data set to physical interactions only (NB. the human data in Table 1 already consist entirely of physical interactions). Similar to the above results, the AUC values drop from around 0.9 (simple random sampling) to 0.6 (balanced sampling), and these values are stable for different subset sizes. Comparing these values to the Table 1 results for yeast, shows that omitting yeast genetic interactions leads to AUC values about 0.05 lower for simple random sampling and 0.1 higher for balanced sampling. The decrease in AUC from simple random sampling when genetic interactions are omitted is probably caused by a decrease in data set bias: genetic interactions lead to hubs of higher...
vertex degree in the positive data set. The increase in AUC for balanced random sampling could perhaps indicate that physical interactions are slightly more predictable from simple sequence features than genetic interactions, but this is not borne out in all data sets (see for instance the smaller data set of physical interactions used in Table 3, vide infra).

Table 1. Performance evaluation of classifier

<table>
<thead>
<tr>
<th>Kernel</th>
<th>YEAST subsets</th>
<th>HUMAN subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>~15,000</td>
<td>~10,000</td>
</tr>
<tr>
<td></td>
<td>~4,000</td>
<td>~5,300</td>
</tr>
<tr>
<td>random sampling</td>
<td>0.95±0.01</td>
<td>0.50±0.03</td>
</tr>
<tr>
<td>random balanced</td>
<td>0.72±0.02</td>
<td>0.53±0.01</td>
</tr>
<tr>
<td>balanced sampling</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>balanced balanced</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Figures reported are the AUC values (area under the ROC curve) performed with 10-fold cross validations on ten subsets of the yeast and human datasets as detailed in Methods. Values given are average±range.

Table 2. Performance evaluation of classifiers

<table>
<thead>
<tr>
<th>Kernel</th>
<th>YEAST subsets of different sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>~5,000</td>
</tr>
<tr>
<td></td>
<td>~15,000</td>
</tr>
<tr>
<td>number of interactions</td>
<td>random sampling</td>
</tr>
<tr>
<td>random sampling</td>
<td>0.92±0.02</td>
</tr>
<tr>
<td>balanced sampling</td>
<td>0.08</td>
</tr>
<tr>
<td>random sampling</td>
<td>0.91±0.01</td>
</tr>
<tr>
<td>balanced sampling</td>
<td>0.07</td>
</tr>
<tr>
<td>random sampling</td>
<td>0.88±0.02</td>
</tr>
<tr>
<td>balanced sampling</td>
<td>0.08</td>
</tr>
<tr>
<td>random sampling</td>
<td>0.89±0.01</td>
</tr>
<tr>
<td>balanced sampling</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Figures reported are the AUC values (area under the ROC curve) performed with 10-fold cross validations on four subsets of yeast dataset which is composed of 47,335 physical interactions as detailed in Methods. Values given are average±range.

3.2 Predictions based on other sequence features

The above results suggest that the prediction of protein interactions from simple sequence features is very difficult. An alternative is to employ evolutionary information, in the form of recognisable conserved domains within the sequence, because it is known that some domains have propensities to interact with other domains, and their presence in the sequence could therefore be predictive of interactions to some extent. We used Pfam (Finn et al., 2010) to identify domains and used a vector of counts of these domains in each sequence with the TPPK kernel to predict PPIs. We found that the Pfam domains are useful features for predicting PPIs, demonstrated by the result of 0.73 AUC on yeast and 0.75 AUC on human using a negative set generated by balanced sampling, and therefore contain stronger information about protein-protein interactions than we are able to extract from a spectrum kernel representation using current state of the art machine learning methods (see Table 3 for comparison with the TPPK and S-kernel spectrum kernel on the same PPI network datasets). The Pfam features encode information that is useful in predicting PPIs, rather than encoding for particular proteins, demonstrated by a similar result of 0.74 AUC on yeast and 0.77 AUC on human using a negative set generated by simple random sampling: here there is minimal bias from the positive and negative data set composition since the features allow the generalisation of the important relationships between Pfam features present in protein sequences, rather than learning to identify that specific proteins tend to appear more in one set.

Table 3. Performance evaluation of classifiers

<table>
<thead>
<tr>
<th>Yeast HC PPI network</th>
<th>Human HC PPI network</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,621 interactions</td>
<td>15,804 interactions</td>
</tr>
<tr>
<td>2,245 proteins</td>
<td>6,198 proteins</td>
</tr>
</tbody>
</table>

Figures reported are the AUC values (area under the ROC curve) performed with 10-fold cross validations for the Pfam kernel and the spectrum kernel. Values given are average±range.

4 DISCUSSION

The high accuracy reported in the literature for predicting protein-protein interactions from simple sequence features appears to be an artefact of the data sets used to train and validate methods. In simple terms, if data sets are used where some proteins (hubs) appear many more times in the positive set than the negative set, then a machine learning method will learn this, predict positive interac-
tions preferentially for these proteins, and seem to be highly accurate. The sequence feature vectors used are of high dimension and allow machine learning methods to identify such specific proteins. The accuracies are unrealistic because the intended application is finding interactors for a specific protein from the entire proteome, of which even hub proteins only interact with a small fraction. It seems in fact that simple sequence features are little better than random in predicting protein interactions, when the task is to distinguish positive and negative interactions that occur in equal number for each protein in balanced positive and negative data sets.

It is clear that the choice of negative data is critical in training and performance evaluation for machine learning methods; in particular there must be no systematic differences between positive and negative datasets used in training and evaluation that are not present in the datasets of intended application. While it is clear that the Negatome (Smialowski et al., 2010) data set will be useful in method development, it is also probable that this set will have its own biases, for instance towards well studied proteins in the literature, and proteins that appear in the structural database. It will therefore have to be used with care in prediction method development.

While simple sequence features are not strongly predictive of interactions, the presence of conserved domains does contain some predictive information. But even then, accuracy is only moderate. Given that the a priori probability of interaction of a protein and a partner chosen randomly from the entire proteome is quite low, any of these methods will be susceptible to a high rate of false positive predictions. Protein interactions depend exquisitely on the three-dimensional atomic structures of the proteins concerned, and their detailed spatial and temporal pattern of expression in the cell. Our ability to predict these interactions from information only in the sequence is, at present, too poor to be useful.

ACKNOWLEDGEMENTS

We are grateful for the collaborations between the Bioinformatics Group, University of Leeds, United Kingdom and the Natural Computing Laboratory, Harbin Institute of Technology, China. We also thank three anonymous referees who helped to improve this work.

Funding: The work was supported by the Natural Science Foundation of China [grant numbers 60932008, 60761001]; and the Natural Science Foundation of Heilongjiang Province in China [grant number ZJG0705], and the University of Leeds.

Conflict of Interest: none declared.

REFERENCES