Bioinformatics, 31(10), 2015, 1521–1525
doi: 10.1093/bioinformatics/btu857
Advance Access Publication Date: 12 January 2015
Discovery Note

OXFORD

Sequence analysis

More challenges for machine-learning protein interactions

Tobias Hamp and Burkhard Rost*

Department of Informatics, Bioinformatics and Computational Biology I12, Technische Universität München, 85748 Garching/Munich, Germany

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 9, 2014; revised on December 6, 2014; accepted on December 23, 2014

Abstract

Motivation: Machine learning may be the most popular computational tool in molecular biology. Providing sustained performance estimates is challenging. The standard cross-validation protocols usually fail in biology. Park and Marcotte found that even refined protocols fail for protein–protein interactions (PPIs).

Results: Here, we sketch additional problems for the prediction of PPIs from sequence alone. First, it not only matters whether proteins A or B of a target interaction A–B are similar to proteins of training interactions (positives), but also whether A or B are similar to proteins of non-interactions (negatives). Second, training on multiple interaction partners per protein did not improve performance for new proteins (not used to train). In contrary, a strictly non-redundant training that ignored good data slightly improved the prediction of difficult cases. Third, which prediction method appears to be best crucially depends on the sequence similarity between the test and the training set, how many true interactions should be found and the expected ratio of negatives to positives. The correct assessment of performance is the most complicated task in the development of prediction methods. Our analyses suggest that PPIs square the challenge for this task.

Availability and implementation: Datasets used in our analyses are available at https://rostlab.org/owiki/index.php/PPI_challenges

Contact: rost@in.tum.de

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

1 Introduction

1.1 Prediction is the acid test for understanding

Machine learning appears to be easily applicable through tools such as Weka (Hall *et al.*, 2009). As opposed to standard statistical analyses, its ability to predict allows distinguishing 'true understanding' from 'rationalization given the fact'. However, if we choose more free parameters than supported by the evidence, we easily over-train or over-optimize, i.e. predictions drop to the descriptive level of statistics, instead of being the acid test for understanding. Cross-validation (CV) and testing on new data are supposed to prevent such a failure.

1.2 Refined CV

Standard CV proceeds as follows. Given a set of observations S to machine learn: split S into two sets S_{train} (to optimize parameters) and S_{test} (to assess performance). Even better: split S into m different sets and use one for testing and the other $m-1$ for training, essentially creating m different solutions to the problem. Many refined versions of the above protocol have been implemented (Kohavi, 1995). However, after several decades of applying machine learning to computational biology, we observe that the above procedures do not suffice to generate sustained performance estimates. Several refinement steps are needed. First, we need to reduce the redundancy

between S_{train} and S_{test}. What exactly constitutes redundancy is not trivial to define, but if we can infer that test protein A has feature X only because it is sequence similar to a training protein A' with feature X (*homology-based inference*), redundancy is too high (Rost, 1999; Rost and Sander, 1993). In this case, even simple statistics can mislead (Rost, 2002). Second, we need to split *S* into three, not two types of sets: training, cross-training and testing (also termed training, testing, hold-out). The cross-training set will be used to decide which method to choose in the end: if you developed two methods m_1 and m_2 and could only publish one, you cannot use the test set for both, optimizing the choice 'm_1 versus m_2' and estimating performance. Data used for method optimization no longer provides independent performance estimates. For this, you need the cross-training complication.

1.3 Refined CV fails for machine-learning PPIs

Park & Marcotte have demonstrated that the above refined CV still fails for the prediction of protein–protein interactions (PPIs) (Park and Marcotte, 2012). For each test interaction A–B between proteins A and B they distinguished three cases: *C1* if both A and B, but not the interaction A–B, were used for training, *C2* if this was the case for either A or B and *C3* if neither A nor B were used for training. By a great margin, all methods performed best for C1 and worst for C3, i.e. for the prediction of proteins without any experimental annotation. Arguably, such new proteins are most interesting to predict.

Here, we present findings that expand on this theme that machine-learning PPIs requires much more careful data set preparations than other applications. Ultimately, PPIs appear to square the noise and the complexity of reducing it, similar to the effect overrepresented protein families had on the prediction of protein function (Rost, 2002).

2 Methods

2.1 Data

2.1.1 Human PPIs

The Hippie database (Schaefer *et al.*, 2012) collects human PPIs with experimental annotations. A reliability score grades each interaction according to the trust in the annotation. We followed the Hippie procedure and reduced version 1.2 (Aug 2011) to the top 10% highest scoring interactions to obtain a high-quality subset (*HumanHQ*). It contained 7237 PPIs from 3915 unique proteins. We applied the same procedure to Hippie version 1.6 (Nov 2013) and used the difference between both sets to test (*HumanHQ_new*; 7201 new PPIs in 3877 proteins; 1561 of these 3877 proteins were new).

2.1.2 Yeast PPIs

Reliable, manually curated yeast PPIs were in the core data set of the Database of Interacting Proteins [*YeastHQ*; Apr 2014; (Salwinski *et al.*, 2004)]. It contains 4796 PPIs among 6434 proteins. Due to slow growth, we could not compile a large enough 'new' test data set.

2.1.3 Redundancy reduction

We reduced redundancy of both, HumanHQ and YeastHQ, by excluding sequence-similar interactions as follows. We considered proteins X to be sequence similar to X′ if their HSSP value (HVAL) was greater than 20 (Mika and Rost, 2003; Rost, 1999; Sander and Schneider, 1991). This corresponds to ∼40% pairwise sequence identity for 250 aligned residues. When we included an interaction A–B in the non-redundant set, we excluded all interactions A′–C and B′–D (A′ similar to A; B′ similar to B). Put differently: A and B were sequence dissimilar to any other protein in the data set. We refer to both sets as *HumanHQ_nr* (842 PPIs) and *YeastHQ_nr* (746 PPIs).

2.1.4 CV and testing (C1, C2, C3)

Here, we introduce 'generalized' Park-Marcotte classes C1–C3. Proteins of a test interaction only need to be similar (HVAL(X,X′) > 20), not identical, to proteins of the training set to be in class C1 or C2. We randomly split each non-redundant set of PPIs (HumanHQ, YeastHQ) into 10 partitions, using nine to train and one to test. All test cases belonged to class C3, because neither A nor B of a test interaction A–B had HVAL > 20 to any protein in the training set. The non-redundant C2 test set first contained each PPI A–B of the full PPI set (HumanHQ or YeastHQ) if exactly one protein (A or B) had HVAL > 20 to any protein in the training set. We redundancy reduced this C2 test set internally in the same way as the full HQ sets before. The C1 test set was created in analogy to C2, except that both proteins A and B needed to have HVAL > 20 to proteins in the training set. We repeated this 10 times so that each of the 10 partitions was the test set exactly once. Classes C1, C2 and C3 contained 1825, 2046 and 842 PPIs for human and 1636, 1663 and 746 for yeast. We applied the same procedure to the time difference test, with the full redundancy reduced HumanHQ as the training set and test sets C1–C3 created from HumanHQ_new (C1: 392, C2: 580 and C3: 218 PPIs).

2.1.5 Negative interactions

Given a positive training set (known PPIs), our definitions of C1–C3 classified each protein pair AB in an organism (none (C3), one (C2) or both (C1) of the proteins A and B have HVAL > 20 to proteins in the training set). Hence, we sampled negatives (non-interactions) randomly for each class C1–C3 from all protein pairs in that class. For example, all the C1 negatives of one training test set combination were sampled from all pairs that were classified as C1 with respect to this particular training set (See Supplementary Material S1.1). All proteins of an organism were given by the EBI Reference Proteomes (Dessimoz *et al.*, 2012) and filtered so that no protein was shorter than 50 or longer than 5000 residues. The number of negatives was set to be 10 times higher than the number of positives in each training and test set. We always removed negatives listed as positive in the full Hippie 1.2 database, i.e. including low-quality experimental evidence.

2.1.6 Data set modifications

The setup described so far was the default for training and comparing different PPI prediction methods. In the following, we describe modifications. *Modification 1: No sequence similarity between negative training and testing*. We tested the effect on performance in class C3 when sequence-similar negatives were removed between training and test sets. To this end, we sampled negatives as before from all C3 pairs in an organism, but made sure that no protein of a negative training interaction had HVAL > 20 to a protein of a negative test interaction. *Modification 2: Bringing redundancy back*. PPIs can be redundant on the level of interaction partners and on the level of sequences. We explored the impact of both types in the training set (keeping test and negative training sets the same; see Supplementary Material S1.2). Redundant interactions were added from the full HQ data set (HumanHQ or YeastHQ).

For redundancy on the level of interaction partners, we took the proteins of all positive PPIs in a redundancy-reduced training set and added all PPIs between them ($2.5 \pm 0.1x$ more PPIs for human, $3.1 \pm 0.1x$ for yeast). Sequence redundancy was added by including all PPIs in the training set that did not violate the sequence similarity constraints to the respective test sets C1–C3 ($4.9 \pm 0.2x$ more PPIs for human, $4.3 \pm 0.1x$ for yeast). We obtained similar results for both types of redundancy. For simplicity, we only show those for interaction partner redundancy in the main text (See Supplementary Material S4.2 for others).

2.1.7 Evaluation

All prediction methods that we assessed can be retrained with custom PPIs and provide a score for each test PPI. We could therefore calculate standard recall-precision curves. To minimize sampling noise, we repeated every experiment above ten times from the beginning (e.g. we performed ten times 10-fold CVs) and averaged over 10 curves. The curves in the Results correspond to the difference between two such average curves, because we measured the change in precision when applying the modifications described before. They were calculated by subtracting the precision values of one curve from those of another curve.

2.1.8 Prediction methods

Park and Marcotte (Park and Marcotte, 2012) established PIPE2 (Pitre *et al.*, 2008) and SigProd (Martin *et al.*, 2005) as the top two methods that predict PPIs from sequence alone and that have already been published. We used those two along with the lower performing but methodologically distinct AutoCorrelation (Guo *et al.*, 2008) and a new in-house method named PPI-PK (Protein-Protein Interaction Profile Kernel; unpublished). The latter encodes a PPI as pairs of k-mers that are conserved in the evolutionary profiles of the two sequences and then uses SVMs for prediction. PIPE2 was the only method that did not use negative PPIs during training. The other details did not matter with respect to the results we report, because we confirmed similar findings for several, very different methods developed in-house over the last years (data not shown).

3 Results and Discussion

Sequence similarity has not played a crucial role in the evaluation of PPIs predictions. One reason may be that homology-based inference of PPIs works reliably only for high levels of similarity (Mika and Rost, 2006). Park and Marcotte discovered that the accuracy of predicting a new target interaction between proteins A and B depends crucially on whether both A and B (C1), any of the two (C2) or neither (C3) have been used for training (Park and Marcotte, 2012). Here, we have investigated three important questions that arose from this discovery. Firstly, can we improve methods by using non-interacting pairs (negative PPIs) that are sequence-similar between training and testing? Secondly, how does data redundancy in the training set affect the three classes (C1–C3)? Thirdly, which prediction method is best and can we actually determine real-world prediction accuracy?

3.1 Sampling of negatives crucial for optimal performance

The work by Park and Marcotte suggests that higher sequence similarity between training and test sets improves performance. This might also pertain to negatives (non-interacting protein pairs), which have exclusively been sampled from proteins of the positive

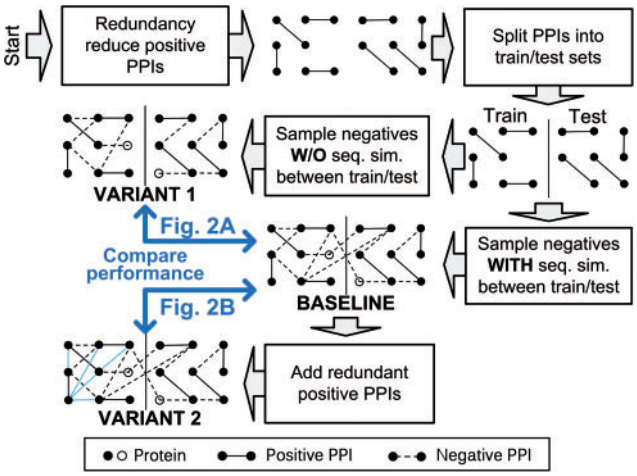

Fig 1. *Flow chart of data preparation.* Starting with all redundant positive PPIs (upper left), we created two data sets (*Variant 1* and *Variant 2*) which differ from a third *Baseline* data set in terms of redundancy amongst positive training PPIs and sequence similarity between negative training and test PPIs. We compared the PPI prediction performance achieved with each variant to the performance of the baseline (Fig. 2)

training interactions in (Park and Marcotte, 2012). We increased similarity by sampling negative PPIs separately for each class C1–C3 from all the respective class-specific protein pairs in an organism (See Section 2). This procedure made negative test PPIs highly sequence-similar to negative training PPIs and sampled all negatives from the same population (See Supplementary Material S1.1).

We measured the effect of this increased sequence similarity with high-quality data sets and several prediction methods. We trained and evaluated models twice: First, all negatives were chosen randomly from all C3 pairs of an organism. Then we did the same, but proteins had to be sequence-dissimilar between training and test sets (See Section 2; Fig. 1). For all data sets, sequence similarity between negative PPIs yielded higher performance for three of the four methods tested (human CV: Fig. 2A, other results and details in Supplementary Material S4.1). PIPE2 (Pitre *et al.*, 2008) was the exception as it did not use negatives for training. Its differences resulted from a different protein background population from which negative PPIs were sampled. This might indicate that the test set was more difficult, thus emphasizing the performance gain of the other methods. The same conclusions hold for yeast and new human PPIs, but the effect of the background population was weaker. AutoCorrelation performed only slightly better than random in C3, so that the differences were not as pronounced as for the other methods.

3.2 Challenging predictions not improved by sampling many PPIs per protein

In machine learning, good training data is equally important to choosing the right learning method. One frequent problem is redundancy. It may carry important signals, but also 'mislead the learner' to focus on less relevant aspects. PPI data sets can have two types of redundancy: overrepresented protein families and overrepresented interaction partners. For example even if all proteins are pairwise dissimilar, one protein may still have many more interaction partners than others. On the other hand, proteins may still be sequence similar even if every protein has only one interaction partner. Most developers have not redundancy reduced PPI data sets at all; very

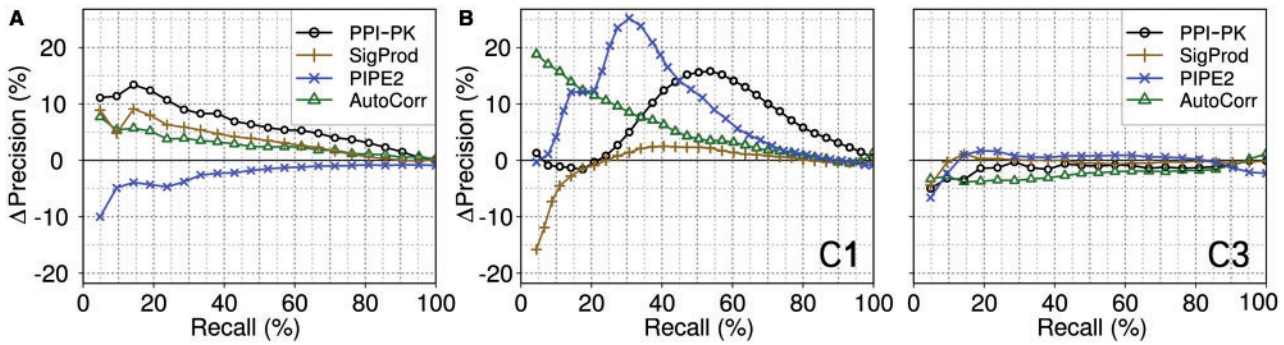

Fig. 2. *Cross-validation results on human data.* (**A**) Increase in precision when allowing sequence similarity between negative training and test interactions. We tested four PPI prediction methods (PPI-PK, SigProd, PIPE2 and AutoCorrelation). Positive values on the *y*-axis indicate that precision increased by that amount when proteins of negative interactions were similar between training and test sets. Negative values indicate decrease in precision. All test cases belonged to difficulty class C3. (**B**) Increase in precision when allowing more than one interaction partner per protein in the positive training sets. The left plot shows results for C1, the right plot for C3 test cases. The meaning of the *y*-axis is analogous to (**A**)

few did it on the level of sequences. Supposedly, interaction modes are so diverse that additional PPIs always add knowledge. Also, different proteins can use the same mechanism of interaction (e.g. SH2/3 domains). Thus, more PPIs should also help predicting interactions between so far unseen proteins.

We put this hypothesis to the test. First, we performed CVs with the non-redundant data sets used before. We considered test classes C1–C3. Then we compared these results to those obtained after adding all interactions back to the training sets and repeating the experiments (See Section 2; Fig. 1). For both yeast and human, we found significant differences in performance for classes C1 and C2 in 15 out of 16 experiments, but virtually no difference for C3 (Fig. 2B for human CV; others in Supplementary Material S4.2). For the lowest levels of recall, the non-redundant training sets even performed slightly better. Adding even more redundancy strengthened the trends (Supplementary Material S4.2).

Our observations were puzzling. A PPI may be more likely to be predicted if similar proteins in the training data have many interaction partners. This would explain some of the difference that we observed in C1–C2. However, it does not explain how a fraction of the input data can achieve the same or even better performance than the full data. We additionally suspect a distinction between signals that act on the level of particular family pairs (interologs) and those that are universally valid among all proteins. For C1 and C2, a new PPI can use the signals specific for at least one of its families, which may be learned if multiple interaction partners for each protein are allowed. For C3 pairs, neither protein can 'hook' to a family and universally valid motifs might dominate.

3.3 Best method lies in the eye of the beholder

If we zoom into the comparison of two methods, e.g. PIPE2 and SigProd, we see that test classes C1–C3 impact the methods differently and even change which method is perceived as better (Supplementary Material S4.3). For human C1, e.g. PIPE2 outperforms SigProd up to ∼51% recall and becomes worse than SigProd for higher recalls. The latter regime is relevant for finding many interactions with as few wet-lab experiments as possible. The precision of PIPE2 in C2 was reduced and lower than SigProd starting at much lower recall levels than in C1. In C3, SigProd consistently outperformed PIPE2. This was the same for all data sets (HumanHQ, YeastHQ and HumanHQ_new).

What is considered the best method to predict PPIs from sequence depends on the sequence similarity between training and test sets, on how many true PPIs should be found (recall) and hence also on the ratio of positives to negatives. The latter has been estimated for human and yeast several times, but keeps changing (Rajagopala *et al.*, 2014; Sambourg and Thierry-Mieg, 2010; Stumpf *et al.*, 2008; Venkatesan *et al.*, 2009). Its effect on precision can be estimated quite precisely (Supplementary Material S2), but classes C1–C3 further complicate matters. For C1, e.g. the ratio might be much lower than for C3 as we may have already found many PPIs experimentally (Supplementary Material S3). In practice, this could turn the intuition that C1 is 'easier' than C2 or C3 on its head. The best C1 predictions may actually contain fewer true interactions than the best C2 or C3 predictions. One may also challenge the biological relevance of the remaining C1 interactions.

4 Conclusions

Our results prove that the determination of performance of PPI prediction methods is even more complicated than anticipated. They also stress the importance of data preparation when machine-learning PPIs. Just as the choice of the method, the best preparation depends on the objective. The latter is largely defined by the sequence similarity between targets and templates (class C1 or C3?). The levels of similarity that are important depend on many parameters, including model organism and amount of experimental data. This is different for cross- and interspecies predictions. For optimal performance, we may also need to trade-off data quality with proteome coverage. Lastly, the sequence similarity threshold that defined classes C1–C3 was arbitrary (HVAL > 20). All trends that we observed here will most likely be reinforced when pushing it to the upper or lower limits. Clearly, much remains to be improved for sequence-based PPI predictions.

Acknowledgements

We thank Yung Ki Park (HJKRI) and Shawn Martin (Sandia) for method implementations and very helpful hints; to Tim Karl (TUM) for invaluable help with hardware and software; to Marlena Drabik (TUM) for administrative support. Thanks also to Miguel Andrade (MDC Berlin) and the whole Hippie team for providing such an excellent resource. Last, not least, thanks to Rolf Apweiler (UniProt, EBI, Hinxton), Amos Bairoch (CALIPHO, SIB, Geneva), Ioannis Xenarios (Swiss-Prot, SIB, Geneva), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

Funding

This work was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung).

Conflict of Interest: none declared.

References

Dessimoz,C. *et al.* (2012) Toward community standards in the quest for orthologs. *Bioinformatics*, **28**, 900–904.

Guo,Y. *et al.* (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18.

Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *IJCAI'95 Proceedings*. Morgan Kaufmann., Montreal, Canada, pp. 1137–1143.

Martin,S. *et al.* (2005) Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218–226.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Mika,S. and Rost,B. (2006) Protein–protein interactions more conserved within species than across species. *PLoS Comput. Biol.*, **2**, e79.

Park,Y. and Marcotte,E.M. (2012) Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods*, **9**, 1134–1136.

Pitre, S. *et al.* (2008) Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res.,* **36**, 4286–4294.

Rajagopala, S.V. *et al.* (2014) The binary protein-protein interaction landscape of Escherichia coli. *Nat. Biotechnol.,* **32**, 285–290.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol. Biol.*, **232**, 584–599.

Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.,* **32**, D449–D451.

Sambourg,L. and Thierry-Mieg,N. (2010) New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics*, **11**, 605.

Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schaefer,M.H. *et al.* (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*, **7**, e31826.

Stumpf,M.P. *et al.* (2008) Estimating the size of the human interactome. *PNAS*, **105**, 6959–6964.

Venkatesan,K. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat Methods*, **6**, 83–90.