Integrative Biology

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/ibiology

Insight, innovation, integration

The identification and validation of effective metrics for protein-protein interaction (PPI) predictions and mainly an increase in the coverage of the interaction network, our methodology has the potential to efficiently predict PPI in an organism. This will allow a comparison of features at networks level and a better knowledge about the target organism, thereby, driving new biological postulations and new experiments. A validated computational method to predict PPI, allows the selection of specific interactions of our interest, reducing costs and increasing success rate in the future experimental results. Likewise, identifying the contribution of each metric for each individual public database and removing the inefficient metrics is important to prevent misuse in PPI network predictions.

An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage

Edson Luiz Folador[a*], Syed Shah Hassan[a], Ney Lemke[b], Debmalya Barh[c], Artur Silva[d], Rafaela Salgado Ferreira[e#], Vasco Azevedo[a#]

[a]Department of General Biology, Instituto de Ciências Biológicas (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
[b]Laboratory of Bioinformatic and Computational Biofisic, Instituto de Biociência, Universidade Estadual de São Paulo (UNESP), Botucatu, São Paulo, Brazil
[c]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India
[d]Instituto de Ciências Biológicas, Universidade Federal do Para, Belém, PA, Brazil.
[e]Department of Biochemistry and Immunology, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

*Corresponding author E-mail: vasco@icb.ufmg.br; Tel: +55 31 3409 2610

#The autor share senior authorship.

Abstract

Automated and efficient methods that map ortholog interactions from several organisms and public databases (pDB) are needed to identify new interactions in an organism of interest (interolog mapping). When computational methods are applied to predict interactions, it is important that these methods be validated and their efficiency proven. In this study, we compare six Blast+ metrics over three dataset to identify the best metric for protein-protein interaction predictions. Using Blast+ to align the protein pairs, the ortholog interactions from DIP were mapped to String, Intact and Psibase pDBs. For each interaction mapped to each pDBs, we retrieved the alignment score, *e-value*, bitscore, similarity, identity and coverage. We evaluated these Blast+ values, and combinations thereof, with the Receiver Operating Characteristic (ROC) curves and computed the Area Under Curve (AUC). To validate these predictions, we used a subset of the Database of Interacting Proteins (DIP) composed of experimental interactions curated by the International Molecular Exchange (IMEx). The cut-off point for each metric/pDB was computed aiming to identify the best on that separates the true and false predicted interactions. In contrast to other methods that only compute the first Blast hit, we considered the first 20 hits, thus increasing the number of predicted interaction pairs. In addition, we identified the contribution of each individual pDB, as well as their combined contribution to the prediction. The best metric had an AUC of 0.96 for a single pDB and AUC of 0.93 for combined pDBs. Compared to other studies, with a cut-off point of 0.70 representing a specificity of 0.95 and sensitivity of 0.90 for individual pDB, our method efficiently predicts protein-protein interactions.

Keywords:

Computational method, Protein-protein interaction prediction, Interaction network, Interolog Mapping, Orthologous interactions

1 Introduction

Understanding the dynamic nature of activities that take place inside the cell of a living organism is necessary at systems biology level. To achieve this, it is necessary to know how the elements of cells such as the genes, transcripts, proteins and various other cellular molecules interact each other and with the outer environment to facilitate the biological functions[1-5]. In this aspect, proteins and their interactions plays an important role and therefore, understanding of protein-protein interactions (PPI) is an important aspect to reveal the molecular mechanism of cell at systems level[6, 7]. Analysis of PPI helps in better understanding of the biology of phylogenetically close and even the distance organisms. PPI networks form complex systems and when such networks are computationally depicted in a graphical form; the nodes represent proteins and non-directional lines connecting these nodes represent the interactions between the proteins[8, 9]. Computationally analyzed PPI helps in developing new hypotheses about an organism and to design the laboratory experiments driven by the hypotheses[10, 11]. In case of infectious microorganisms, studying PPI networks offer identification of pathogenic proteins and therefore offers new opportunities for developing novel drug and vaccines[12-14]. The interactions of proteins within a cell depend on several biological or physico-chemical factors[15] and the PPI can be physical interactions, regulatory associations, genetic interactions, structural interactions, functional similarity associations among others. Such associations are not mutually exclusive and may occur simultaneously[8]. Several methods have been developed for studying PPI that can be categorized as genetic, biochemical, biophysical, high throughput, and computational approaches[16]. Several methods have been developed for studying PPI that can be categorized as genetic, biochemical, biophysical, high throughput, and computational approaches[16]. The important experimental methods include yeast-two-hybrid (Y2H)[17], protein chip, tandem affinity purification followed by mass spectrometry (TAP-MS)[18], atomic force microscopy (AFM)[4, 8, 9, 19, 20] and analytical ultracentrifugation (UC)[6]. Each approach has its advantages and disadvantages and therefore more than one technique may required to eliminate the false positives[16]. Computational methods can handle entire proteome interactions but generates false-positives interactions similar to the high throughput techniques[3, 8, 21]. Computational prediction of PPI and their analysis can be done using machine learning techniques[11, 22-26], protein sequence homology or interolog mapping[27-29], three-dimensional protein structure analysis[30-33], docking studies[34], domains interactions[35], text mining[36-39], protein co-evolution approaches[20, 23, 40], Mirror tree method[41], phylogenetic profile analysis[20] or a combination of these methods[42], which have also been described and reviewed in other works[43-46]. Computational methods, individually or in combination, have been used to develop and analysis of PPI interaction networks in several organisms such as *Drosophila melanogaster*[28], *Arabidopsis thaliana*[29], *Leishmania brasiliensis, Leishmania major and Leishmania infantum*[2, 27], yeast[17], *Saccharomyces cerevisiae*[47], *Xanthomonas oryzae*[48], *Helicobacter pylori*[49] and *Human*[50]. When the interaction network is predicted using sequence homology or interolog mapping, it is assumed that, if a pair of proteins interact in a particular organism, the ortholog proteins in another organism will interact as a similar pattern[3, 16] and is used to identify the conservation of protein interactions between two organisms when there is high similarity in the sequence of proteins[51] and transfer annotations between genomes[52]. But the prediction efficiency of interolog mapping is not yet satisfactory as compared to other computational methods[33]. This may be due to the use of only the first Blast hit[53]. Therefore there is scope of improving the method for its efficacy and accuracy in predicting and analyzing the PPI. Here, using publicly available PPI databases (pDB) both individually and collectively and less stringent criterion for Blast+; we tried to increase the efficacy and sensitivity of interolog mapping based PPI with minimal false-positive and false-negative interactions.

1.2 Materials and methods

1.2.1 Databases used

In this work, we have used four pDB: Database of Interacting Proteins (DIP)[54], String[55], Intact[56], and Psibase[57] (**Error! Reference source not found.**~~Supplementary material S1~~). Since the DIP contains experimental and curated data[58] for PPIs, it was used as the gold standard to evaluate our prediction. Aiming to increase the coverage of the interaction network prediction while also reducing the false negatives and false positives, we mapped the ortholog interactions and conducted the prediction of those interaction pairs found in the DIP database by comparing against three other pDBs instead of only one[20].

1.2.2 Blast+

The BLASTp program from the Blast+ package[53] was used to align and map de ortholog proteins between the databases. All the six alignment values of BLASTp: score, *e-value*, bit score, similarity, identity and coverage were considered to compose the metrics that will be evaluated. Aiming to validate a methodology that is able to classify non-orthologous and orthologous proteins, we run the Blast+ with the *e-value* parameter set to 0.1, all other parameters at their default value. To compare the metrics and how much each pDB contributes to the prediction of interaction pairs, we ran Blast+ to generate two distinct datasets: the first contains only the first Blast+ hit (num_alignments 1) and the second contains the first 20 Blast+ hits (num_alignments 20).

1.2.3 Interolog mapping

To map the ortholog proteins between pDBs using Blast+, we first used the DIP proteome as the query and the proteomes of the other pDBs (String, Intact and Psibase) as the subject. We then inverted this process, using the latter pDBs as the query and the DIP proteome as the subject. For the interaction analysis, only those proteins that had a reciprocal hit (RH), i.e., when protein "a" from DIP align to protein "A" from the pDB and protein "A" from the pDB align to protein "a" from the DIP were considered. Specific datasets and metrics were generated for each pDB versus DIP combination. For each identified RH, we extracted six values from the Blast+ alignment results as mentioned before. For each reciprocal hit, the minimum value of its metric was calculated using the following formula:

$$RH(a) = \min(BlastValue\ (a \rightarrow A), BlastValue\ (a \leftarrow A))$$

Here, “BlastValue” represents each of the six values extracted from the Blast+ alignment that will be evaluated, "a" represents the protein in our gold standard (DIP), and "A" represents the pDB protein. The reciprocal hit (RH) is represented by both “a→A”, indicating that the protein “a” in the DIP was used as the query and was aligned against the protein “A” in the pDB, and by “a←A”, indicating that the protein “A” in the pDB was used as the query and was aligned against the protein “a” in the DIP. The following thus represent an interaction pair:

$$RH(a), RH(b)$$

Here, the proteins "a" and "b" are reciprocal hits of proteins "A" and "B", respectively. Moreover, "A" and "B" are the identifiers of the interaction pairs found in the pDBs and were used to map the interaction pairs "a" and "b" in our gold standard DIP. The metric about each predicted interaction pairs were assessed by two distinct manners: using the average metric value and using the smallest metric value, which were respectively denoted by the following formulas:

$$avg(ab) = (RH(a) + RH(b))/2$$

$$min(ab) = \min(RH(a),RH(b))$$

Moreover, each pDB has its own confidence score that was also evaluated both individually and in combination with the other metrics extracted from the RHs. In addition, we have evaluated the contribution of each pDB to the interaction pair, for which we combined the other metrics with the number of times that the interaction pair was predicted in the pDBs (qt_pDB), giving greater weight to interaction pairs predicted by different pDBs.

1.2.4 Validation and precision prediction

To assess the efficiency of our predictions, in addition to a positive set of interactions, a set of negative interactions is also necessary. Because the DIP database contains only positive interactions, the negative interaction pairs were randomly generated from the DIP protein identifiers through an in-house script at a ratio of five times the number of positive interactions. This negative dataset is composed of protein interaction pairs that are not found in the set of known interactions[59]. We created metrics with each value extracted from Blast+, with the pDB score, with the number of databases in which the interaction was predicted (qt_pDB), or by combining these values. These metrics were validated for each pDB both individually and collectively, seeking to identify which metric variation versus pDB best represents the set of positive and negative interactions found in our gold standard (DIP). To validate the metrics and their combinations, we used the Receiver Operating Characteristic (ROC) curve plots and calculated the Area Under Curve (AUC) for each metric using the software package ROCR[60]. For metrics with a better AUC value, when seeking to identify a cut-off point that best represented the positive and negative sets of predicted interactions, we tested values from zero to one as cut-off points and compute the sensitivity, specificity and precision by the following formulas:

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Precision = TP / (TP+FP)

The best cut-off point was chosen using the formula

Sensitivity x Specificity

because, aside from being easy to implement, its result is equivalent to the Matthews Correlation Coefficient (MCC)[41]. The entire method is represented in Supplementary material S2.

2 Results and discussion

2.1 Comparison of predictions based on different numbers of blast alignments

One motivation for this study was the hypothesis that, when only the first hit returned by Blast+ is considered, important results might be disregarded. To test this hypothesis, we performed the analysis using two datasets: one containing only the first Blast+ hit (num_alignments 1) and another containing the first 20 Blast+ hits (num_alignments 20). We compared these two datasets and observed a general 16.95-fold increase in the number of alignments and a 5.10-fold increase in the number of distinct predicted interaction pairs. Proportionally, there was a larger increase in the number of alignments than in the number of interaction pairs. This fact is explained by comparing, especially in the case of the String pDB, the total number of interaction pairs (25,343,169) with the number of distinct interaction pairs (5,382,086), becoming evident the number of repeated interaction pairs (Table 1). When we used 20 Blast+ alignments, it is natural to expect that, if there are homolog proteins among the pDBs, these will be aligned against the same sequence in the DIP, thus mapping the same DIP identifier. Consequently, it reduces the number of distinct DIP interaction pairs identified in relation to the number of Blast+ alignments.

Table 1 – Quantification of the alignments and interaction pairs comparing 1 and 20 blast hits dataset

	Blast+ output alignment hits			Interaction pairs mapped from the pDBs			
pDB	1 hit	20 hits	Proportion	1 hit	20 hits	20 hits(*)	Proportion(*)
String	44,660	853,234	19.10	1,651,858	25,343,169	5,382,086	3.25
Intact	41,846	450,308	10.76	101,439	5,023,022	3,518,501	34.6
Psibase	9,392	322,272	34.31	112	314,280	47,951	428.13
Total	95,898	1,625,814	16.95	1,753,409	30,680,471	8,948,538	5.10

1 hit: corresponds to reciprocal hits from Blast+ running with the parameter num_alignments set to 1. 20 hits: corresponds to reciprocal hits from Blast+ running with the parameter num_alignments set to 20. Proportion(*): Proportion of the quantity of interaction revealed by Blast+ with num_alignments 20 had over num_alignments 1 (20 hits(*) / 1 hit). Hits were counted in both the a->A and a<-A directions. (*) Represents the number of distinct interaction pairs for Blast+ 20 hits.

Consideration of first 20 Blast+ alignments generates a large number of repeated interaction pairs. But we were able to increase the number of distinct interaction pairs five times more with an aim to increase >5 times the network coverage for a more informative interactions. After significant increase in the number of distinct interaction pairs generated by Blast+ (num_alignments 20), we investigated the amount of said alignments in relation to the number of hits that Blast+ returned after each run. It was done to identify how much distinctiveness is actually contributed by increasing the parameter num_alignments to 20. From the total 812,907 alignments returned by Blast+ for the three pDBs, 71.8% had 20 hits, indicating that an even higher cut-off value for num_alignments, may be 30 or 40, could be considered (Supplementary material S3). In addition, we investigated the quality of these alignments because better alignments have a greater chance of participating in positive interactions. We then considered only those hits with > 80% identity versus coverage ratio. Most Blast+ alignments (41.4%) had exactly 20 hits indicating that num_alignments to a value above 20 might return significant alignments too (Supplementary material S3). Considering that these Blast+ alignment results are not homologous proteins, which would map identical identifiers in the DIP, they certainly should contribute to the identification of new interaction pairs. Hence, we investigated the number of distinct identifiers mapped to the DIP that would be returned when the Blast+ parameter num_alignments is set to values between 1 and 20. For this analysis, we considered that identifiers found with num_alignments 2 were unique. This was done successively until num_alignments was set to 20, and only the unique identifiers that were not found in identifier sets for num_alignments below 20 were considered. As expected, most distinct DIP identifiers were found when num_alignments was set to 1 (76.65%) and only 1.4% when num_alignments 20. Of the total 23,680 distinct identifiers present in the DIP, 23,280 were found with the Blast+ parameter num_alignments set to 20, achieving a total identifier coverage of 98%. Comparing the use of num_alignment set to 1 and 20, there was an increase of approximately 23% in the number of distinct identifiers (Supplementary material S3). Although it is small, this increase may contribute to increase the number of predicted interacting pairs therefore may increase the network coverage.

2.2 Analysis of interaction pairs

In our gold standard database DIP, there are positive and negative interaction pairs. The positive set consists of experimental interactions curated by the IMEX consortium[58], whereas the negative set was randomly generated at a proportion of five times the number of positive interactions. In the DIP, all predicted interaction pairs can not be mapped. Therefore, it is impossible to assess whether these predicted interactions are true or false. To avoid the doubtful inference, we considered only those interaction pairs predicted in the pDBs that were also mapped in the DIP to analyze our metrics. Given the difference in the number of Blast+ hits when comparing the two datasets generated with num_alignments set to 1 and 20, we studied the pattern of each metric in the interactions generated by each dataset. To do this, we predicted the PPI pairs, generated ROC curves and computed the respective AUC values for both the datasets: num_alignments 1 (Table 2) and num_alignments 20 (Table 3). For both the datasets, we used the metric avg(ab) to compute the six proposed blast values; score, bitscore, conserved, identity, expected and pdb_score, in addition to a combination of two other metrics. For the first dataset, the score, bitscore, conserved, identity and expected blast values displayed a random behavior with an AUC close to 0.50. Therefore, it was not possible to distinguish between positive and negative interactions. In contrast, the pDB_score metric showed considerable improvement for the String (AUC 0.70) and Intact (AUC 0.72) pDBs individually. However, when these pDBs were combined the AUC value became 0.69. We then tested the Combined I metric (pDB score * qt_pDB), which showed considerable improvement for the pDB combination (0.80) and for the String pDB (AUC 0.82), whereas the result was poorer for the Intact pDB (0.58). After observing the behavior of the metrics, we combined the best metric of each individual pDB (pDB score*qt_pDB for String and pDB score*3 for Intact) to compose the Combined II metric. This approach yielded the best result for each pDB individually (AUC of 0.82 for String and 0.72 for Intact) as well as the best result for the combined pDBs (AUC 0.90). We evaluated all metrics for the Psibase pDB in an identical manner, but only a small number of positive interactions were mapped without a set of negative interactions as required to generate an ROC curve (Table 2). In all ROC curves, “All pDB” corresponds to the union of the data from all the other pDBs which, in theory, would be expected to contain a value close to the average AUC of the individual pDBs. However, in some cases, the AUC value was below the average. This suggested that joining the data from distinct pDBs and assessing them using the same metric will not always improve prediction and that this condition should be carefully tested. We can improve predictions by combining these metrics (Table 2 - Combined I). Still, if the best metrics of each individual pDB are normalized, they may collectively produce better results than if they are individually analyzed (Table 2 - Combined II).

Table 2 – AUC values relating to metrics from dataset created with Blast+ parameter num_alignments set to 1 and average interaction pair metric value (avg(ab)).

AUC Metric	pDB Intact	pDB String	pDB Psibase	All pDB
Score	0.44	0.52	?	0.51
Bitscore	0.44	0.52	?	0.51
Conserved	0.46	0.49	?	0.49
Identity	0.46	0.49	?	0.49
Expected	0.47	0.50	?	0.50
pDB_score	0.72	0.70	?	0.69
Combined I	0.58	0.82	?	0.80
Combined II	0.72	0.82	?	0.90

All pDB: contains the combined data of Intact, String and Psibase pDBs. The values ? of pDB Psibase column could not be computed. The ROC curves related to the AUC values are detailed in Supplementary material S4.

Other combinations of values may generate better metrics for predicting interactions in these datasets (num_alignments 1). Our priority, however, was to perform larger analyses for the dataset generated with the Blast+ parameter num_alignments set to 20 (Table 3).

This parameter value is justified by the increased number of predicted interaction pairs, the improvement in the ROC curves and the AUC values together making this dataset more biologically relevant for analysis. Because it contains more interaction pairs, it was possible to generate the plots for the Psibase pDB, even though the AUC values for this pDB were not good. For the String and Psibase pDBs, the AUC values showed considerable improvement for all metrics. The Conserved and Identity metrics yielded the best AUC values for each individual pDB, especially for Intact, with AUC of 0.95 and 0.96, respectively. The Identity metric was used to compose the Combined II metric, which yielded the best AUC value for this dataset, both for the individual pDBs and for their combination (AUC 0.92 - Table 3). To improve the AUC values obtained with avg(ab) metrics (Table 3), we also computed the min(ab) metrics to the interaction pair (Table 4). The comparison of the plots generated for the ROC curves shows that both the metrics obtained from the average value for the interaction pair (Table 3) and those obtained from the minimum value (Table 4) yielded good results, indicating that, these two metrics are similar in predicting interaction networks. A considerable improvement is observed for the Psibase pDB when the metric is computed using the minimum value of each interaction pair. In both datasets analyzed in this study, the AUC value for the Combined II metric (0.92 - Table 4) obtained by joining all pDBs was very close to that was found in another study[27], where an AUC equal to 0.94 was obtained.

Table 3 – AUC values relating to metrics from dataset created with Blast+ parameter num_alignments set to 20 and average interaction pair metric value (avg(ab)).

AUC Metric	pDB Intact	pDB String	pDB Psibase	All pDB
Score	0.83	0.60	0.58	0.68
Bitscore	0.83	0.60	0.58	0.68
Conserved	0.95	0.73	0.67	0.80
Identity	0.96	0.74	0.68	0.81
Expected	0.88	0.61	0.60	0.71
pDB_score	0.57	0.72	0.50	0.65
Combined I	0.79	0.84	0.50	0.80
Combined II	0.96	0.91	0.72	0.92

All pDB: contains the combined data of Intact, String and Psibase pDBs. The ROC curves related to the AUC values are detailed in Supplementary material S5.

By analyzing the pDBs individually, we identified their individual contribution to the composition of the general AUC value of all pDBs. The largest contribution was from the Intact pDB (0.96), followed by the String (0.90) and Psibase pDBs (0.79) (Table 4 -

Combined II). Each pDB gave a different AUC for each metric, contributing in different ways to the composition of the general AUC value. Distinct pDB combinations can also contribute differently to prediction, a fact observed when analyzing the ROC curve generated using only, both the String and Intact pDB. Without the Psibase pDB, the ROC curve yielded a better general AUC (0.93 - Figure 2 - Combined II).

Table 4 – AUC values relating to metrics from dataset created with Blast+ parameter num_alignments set to 20 and minimum interaction pair metric value (min(ab)).

AUC Metric	pDB Intact	pDB String	pDB Psibase	All pDB
Score	0.88	0.61	0.73	0.71
Bitscore	0.88	0.61	0.73	0.71
Conserved	0.95	0.74	0.74	0.80
Identity	0.96	0.74	0.77	0.81
Expected	0.89	0.61	0.73	0.71
pDB_score	0.57	0.72	0.50	0.65
Combined I	0.79	0.84	0.50	0.80
Combined II	0.96	0.90	0.79	0.92

All pDB: contains the combined data of Intact, String and Psibase pDB. The ROC curves related to the AUC values are detailed in Supplementary material S6.

Independently from using the average (avg(ab)) or minimum (min(ab)) value in the metrics, the individual values extracted from Blast+ that were most effective in predicting interaction pairs were Coverage and Identity. When an interaction pair is predicted by more than one pDB, the chances of this interaction being true are higher. We used this premise to improve the ROC curves of the String and Psibase pDBs by giving greater weight to interactions that were predicted in more than one pDB (qt_pDB in Combined II). For the Psibase pDB, this change did not improve the curve; however, it significantly improved for the combination of all pDBs (0.92) and for the String+Intact pDB combination (0.93). Individually, the Intact pDB had the best AUC value (0.96) (Figure 2 – Supplementary material S6).

For the best ROC curves, we assessed several cut-off points to choose the one having the best relationship between sensitivity and specificity. We tested cut-off points for the Combined II metric in relation to the Intact pDB on its own (Figure 3) and for the union of the String and Intact pDBs (Figure 4). For both the tested sets, the sensitivity and specificity were inversely correlated, which made it difficult to choose the best suited cut-off point. We also tested the sensitivity to specificity ratio, a measure that is equivalent to the Matthews Correlation Coefficient (MCC), which has been used to predict interaction networks[41]. For both the Intact pDB dataset and the String+Intact combination, the best cut-off point of the Combined II metric was at 0.70, representing the highest sensitivity to specificity ratio (Figure 3 and Figure 4). The cut-off point at 0.70 corresponded to a sensitivity of 0.90 and a specificity of 0.95 for the Intact pDB and to a sensitivity of 0.83 and specificity of 0.95 for the String+Intact pDB (Table 5). This cut-off point was more specific than sensitive, which, in practice, means that less interaction pairs would be selected (0.90-0.83). However, the generated results have a higher probability of being true (0.95).

Table 5 – Summary of Roc curve obtained by applying the Combined II metric

Data	AUC	Cut-Off	Sensitivity	Specificity	Sens. * Spec.	Precision
Intact	0.96	0.70	0.90	0.95	0.86	0.99
String+Intact	0.93	0.70	0.83	0.95	0.79	0.99

The following formulas were used to compute the values in this table: Sensitivity = TP / (TP + FN); Specificity = TN / (TN + FP); Precision: TP / (TP+FP).

The Combined II metric consists of the identity and coverage values extracted from Blast+. The cut-off point is a ratio of these two values, e.g., equivalent to a coverage of 0.837 and an identity of 0.837 or a combinations of these values for which the product is 0.70. This cut-off point was higher than those were recommended (0.30 for identity and 0.80 for coverage) to avoid the identification of false positives using the method of homolog interaction mapping [16]. The value corresponding to the score of each pDB itself (pDB score) used in the Combined I metric (Table 4) considerably improved the individual prediction for the String pDB. Thus, the pDB score could be used in combination with other values extracted from Blast+ to further improve the ROC curve of the String pDB individually or together with other pDBs. The use of the pDB score, even if justified by improvements in the ROC curve, would lead us to use different metrics for each pDB in the same ROC curve. Because this practice is not reported in the literature, we adopted a conservative posture and did not add this value for the String pDB. Each pDB sets its own criteria to classify the interactions as true, and as a consequence, the use of different metrics for each pDB may normalize these criteria and improve the prediction of interaction networks when several pDBs are used. In addition to the values extracted from the Blast+ alignments and the pDB score, the way we use the negative interaction set of the gold standard to evaluate metrics can also influence the final results (Supplementary material S7 – The negative dataset).

2.3 Comparison to similar studies

Several other methods and metrics have been developed and have shown themselves viable when applied to the prediction of interaction networks (Table 6). A comparison of the metrics found in other studies with the presented herein, considering the different methods, techniques and datasets used by each, has shown our method to be effective: it obtained an AUC of 0.93 for the String+Intact pDB combination and an AUC of 0.96 for the Intact pDB individually. The prediction of interactions using the interolog mapping method was shown to be viable for application, due to both the results presented in this study and the comparison to other studies (Table 6).

Table 6 - Comparison of the AUC value of our methodology against other methods

Method	AUC Value	Reference
Structure	Not informed	33
Support Vector Machine (SVM)	0.69	24
Support Vector Machine (SVM)	Not informed	26
Text-Mining (*)	0.91	37
Interolog Mapping	0.71	28
Mirrortree	0.73	41
Interolog Mapping (**)	0.94	27
Interolog Mapping (***)	0.96 and 0.93	This study

* Organism-specific method that makes predictions only for annotated genes
** Using only a single first hit of the Blast[61] program and only 702 interactions as positive gold standard dataset.
*** Using the first 20 Blast+ [53] hits for prediction

Finally, we used to evaluate our work a data set consisting of 70.630 experimental and cured interactions as the gold standard[54, 58]. Considering the different metrics used to measure the efficiency of the prediction methods and the cut-off point of 0.70, we obtained a precision of 0.99 for both metrics, a value higher than the precision of 0.74 obtained with a method based on text mining[38]. In addition, comparing the results from our methodology obtained here with the methodology using Support Vector Machine (SVM) and 1.500 protein interactions, though the specificity (0.98) and precision (0.8) values are approximate in both works, the sensitivity value (0.15 and 0.28)[26] was much lower than the obtained value in this study (0.83 and 0.90, Table 5). These results, thus, reinforce the efficiency of our metrics and the good ratio between sensitivity and specificity.

3 Conclusions

This is the first study that uses the first 20 Blast+ hits to compare the combinations of values extracted from alignments for the prediction of PPIs using ortholog interaction mapping and, in addition, evaluates these values for each pDB individually and in combination. Based on our observations in this study, we concluded that each pDB contributes differently to the prediction of interactions, and when used in combinations, the results must be carefully analyzed because adding another pDB does not necessarily improve prediction. This study contributes to the scientific community the good AUC values obtained from the pDB Intact (0.96) and pDB Intact + String (0.93). Most importantly, it also contributes to the possibility of increasing the coverage of a predicted interaction network for an organism by using the first 20 Blast+ hits instead of only the single first hit, thus maintaining a decent performance. In addition, despite identifying the metrics that yield good AUC values, we also identified the metrics that are not adequate for predicting PPI using the interolog-mapping method. The blast values such as *e-value*, score and bit score are good metrics for indicating the best alignments for one query protein against a group, but they fail to generally differ true and false homology for all query proteins of a group. In this way, it becomes difficult to identify a cut-off point to distinguish true homologous proteins. This phenomenon is explained by the bias that these metrics are due to the size of the subject database (*e-value*) or even due to the length of the amino acid sequence (score and bit score). After all, two small proteins with good alignments receive a lower score than two larger proteins with good alignments. The combination of the coverage and identity metrics was effective to mapping orthologous interactions. It joins in a single metric, both the quality (identity) and quantity (coverage) of an alignment between two proteins. In this case, the database size do not influence these metrics and, the percentage values act as normalizers for the protein size. With the results obtained in this study, we intend to use and apply our methodology to predict the *pan-interactome* of fifteen strains of the gram-positive bacterium *Corynebacterium pseudotuberculosis,* a pathogen of great veterinary and economic importance. In addition, we will use the properties of the predicted interaction network to improve the functional annotation of *C. pseudotuberculosis* genes[7, 52]. Likewise, we hope that the scientific community will also make use of the *in silico* methodology that we have validated here, to predict the interaction networks of their organisms of interest. The approach we have followed can be reproduced using public-domain computer programs and databases that are freely available.

Author Contributions

Conceived and designed the experiments: ELF. Performed the experiments: ELF. Analyzed the data: ELF, SSH, RSF, NL, DB. Wrote the paper: ELF. Participated in revising the draft: ALL. Contributed materials/analysis tools: AS, VA.

Acknowledgements

This study was conducted with support from the CENAPAD-MG. Funding: Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), Conselho Nacional de Pesquisa (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig).

Bibliography

1. L. Garma, S. Mukherjee, P. Mitra and Y. Zhang, *PloS one*, 2012, **7**, e38913.
2. A. Flórez, D. Park, J. Bhak, B. C. Kim, A. Kuchinsky, J. Morris, J. Espinosa and C. Muskus, *BMC bioinformatics*, 2010, **11**, 484.
3. R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp and T. Ideker, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 1974-1979.
4. A. L. Barabási and Z. N. Oltvai, *Nature Reviews Genetics*, 2004, **5**, 101-113.
5. M. W. Gonzalez and M. G. Kann, *PLoS computational biology*, 2012, **8**, e1002819.
6. N. Wetie, G. Armand, I. Sokolowska, A. G. Woods, U. Roy, J. A. Loo and C. C. Darie, *Proteomics*, 2013.
7. W. Peng, J. Wang, J. Cai, L. Chen, M. Li and F.-X. Wu, *BMC systems biology*, 2014, **8**, 35.
8. J. De Las Rivas and C. Fontanillo, *Briefings in Functional Genomics*, 2012.
9. J. Wang, M. Li, Y. Deng and Y. Pan, *BMC genomics*, 2010, **11**, S10.
10. P. Braun and A. C. Gingras, *Proteomics*, 2012, **12**, 1478-1498.
11. X. Zhang, J. Xu and W.-x. Xiao, *PloS one*, 2013, **8**, e58763.
12. B. Andreopoulos and D. Labudde.
13. H. Li, V. Kasam, C. S. Tautermann, D. Seeliger and N. Vaidehi, 2014.
14. K. Lage, *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 2014.
15. J. Luo, Y. Guo, Y. Zhong, D. Ma, W. Li and M. Li, *Journal of Computer-Aided Molecular Design*, 2014, 1-11.

16. A. G. N. Wetie, I. Sokolowska, A. G. Woods, U. Roy, K. Deinhardt and C. C. Darie, *Cellular and Molecular Life Sciences*, 2013, 1-24.

17. H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li and N. Simonis, *Science*, 2008, **322**, 104-110.

18. X. Sun, P. Hong, M. Kulkarni, Y. Kwon and N. Perrimon, Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, 2012.

19. F. S. Kao, W. Ger, Y. R. Pan, H. C. Yu, R. Q. Hsu and H. M. Chen, *Biotechnology and Bioengineering*, 2012.

20. E. D. Harrington, L. J. Jensen and P. Bork, *FEBS letters*, 2008, **582**, 1251-1258.

21. R. Mrowka, A. Patzak and H. Herzel, *Genome research*, 2001, **11**, 1971-1973.

22. B. Q. Li, K. Y. Feng, L. Chen, T. Huang and Y. D. Cai, *PloS one*, 2012, **7**, e43927.

23. R. A. Craig and L. Liao, *BMC bioinformatics*, 2007, **8**, 6.

24. L. Li, P. Zhang, T. Zheng, H. Zhang, Z. Jiang and D. Huang, *PloS one*, 2014, **9**, e91898.

25. S.-W. Zhang, L.-Y. Hao and T.-H. Zhang, *International journal of molecular sciences*, 2014, **15**, 3220-3233.

26. H. Kumar, S. Srivastava and P. Varadwaj, *International Journal for Computational Biology (IJCB)*, 2014, **3**, 37-43.

27. A. M. Rezende, E. L. Folador, D. M. Resende and J. C. Ruiz, *PloS one*, 2012, **7**, e51304.

28. G. Gallone, T. I. Simpson, J. D. Armstrong and A. P. Jarman, *BMC bioinformatics*, 2011, **12**, 289.

29. J. Geisler-Lee, N. O'Toole, R. Ammar, N. J. Provart, A. H. Millar and M. Geisler, *Plant physiology*, 2007, **145**, 317-329.

30. W. Zhou, H. Yan, X. Fan and Q. Hao, *Current Bioinformatics*, 2013, **8**, 3-8.

31. Q. C. Zhang, D. Petrey, J. I. Garzón, L. Deng and B. Honig, *Nucleic Acids Research*, 2013, **41**, D828-D833.

32. R. A. Jordan, E. L. M. Yasser, D. Dobbs and V. Honavar, *BMC bioinformatics*, 2012, **13**, 41.

33. Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili and T. Hunter, *Nature*, 2012, **490**, 556-560.

34. M. Ohue, Y. Matsuzaki, T. Shimoda, T. Ishida and Y. Akiyama, *Highly precise protein-protein interaction prediction by integrating template-based and template-free protein docking*, 2013.

35. W. H. Jang, S. H. Jung and D. S. Han, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2012, **9**, 1081-1090.

36. M. Krallinger, F. Leitner, M. Vazquez, D. Salgado, C. Marcelle, M. Tyers, A. Valencia and A. Chatr-Aryamontri, *Database: the journal of biological databases and curation*, 2012, **2012**.

37. Z. Xiang, T. Qin, Z. S. Qin and Y. He, *BMC Systems Biology*, 2013, **7**, S9.

38. J. Köster, E. Zamir and S. Rahmann, *Integrative Biology*, 2012, **4**, 805-812.

39. J. Czarnecki and A. J. Shepherd, in *Biomedical Literature Mining*, Springer, Editon edn., 2014, pp. 135-145.

40. T. Sato, Y. Yamanishi, M. Kanehisa and H. Toh, *Bioinformatics*, 2005, **21**, 3482-3489.

41. H. Zhou and E. Jakobsson, *PloS one*, 2013, **8**, e81100.

42. C. Saccà, S. Teso, M. Diligenti and A. Passerini, *BMC Bioinformatics*, 2014, **15**, 103.

43. A. Valencia and F. Pazos, *Current opinion in structural biology*, 2002, **12**, 368-373.

44. L. Skrabanek, H. K. Saini, G. D. Bader and A. J. Enright, *Molecular biotechnology*, 2008, **38**, 1-17.

45. V. S. Rao, K. Srinivas, G. Sujini and G. Kumar, *International Journal of Proteomics*, 2014, **2014**.

46. J. Zahiri, J. Hannon Bozorgmehr and A. Masoudi-Nejad, *Current genomics*, 2013, **14**, 397-414.

47. T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred and A. Tong, *Journal of biology*, 2006, **5**, 11.

48. J. G. Kim, D. Park, B. C. Kim, S. W. Cho, Y. T. Kim, Y. J. Park, H. J. Cho, H. Park, K. B. Kim and K. O. Yoon, *BMC bioinformatics*, 2008, **9**, 41.

49. R. Häuser, A. Ceol, S. V. Rajagopala, R. Mosca, G. Siszler, N. Wermke, P. Sikorski, F. Schwarz, M. Schick and S. Wuchty, *Molecular & Cellular Proteomics*, 2014, **13**, 1318-1329.

50. X. Yu, A. Wallqvist and J. Reifman, *BMC bioinformatics*, 2012, **13**, 79.

51. A. C. F. Lewis, N. S. Jones, M. A. Porter and C. M. Deane, *PLOS Computational Biology*, 2012, **8**, e1002645.

52. H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal and M. Gerstein, *Genome research*, 2004, **14**, 1107-1118.

53. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. Madden, *BMC bioinformatics*, 2009, **10**, 421.

54. I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, *Nucleic acids research*, 2000, **28**, 289-291.

55. A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork and C. von Mering, *Nucleic acids research*, 2013, **41**, D808-D815.

56. H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff and A. Valencia, *Nucleic acids research*, 2004, **32**, D452-D455.

57. S. Gong, G. Yoon, I. Jang, D. Bolser, P. Dafas, M. Schroeder, H. Choi, Y. Cho, K. Han and S. Lee, *Bioinformatics*, 2005, **21**, 2541-2543.

58. S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman and G. Cesareni, *Nature methods*, 2012, **9**, 345-350.

59. V. Y. Muley and A. Ranjan, *PloS one*, 2012, **7**, e42057.

60. T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, *Bioinformatics*, 2005, **21**, 3940-3941.

61. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *Journal of molecular biology*, 1990, **215**, 403-410.

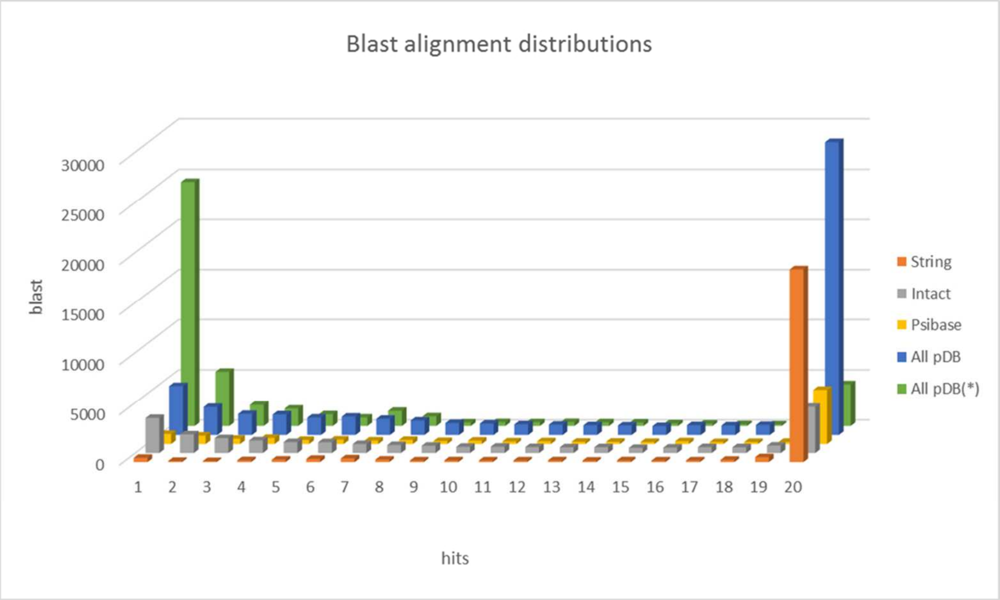
Figure legends

Figure 1 Distribution of Blast+ alignments grouped by number of hits. The alignments was generated with the Blast+ parameter num_alignments set to 20. All pDB: is the sum of String, Psibase and Intact. (*) Alignments in which the coverage to identity ratio is above 80%.

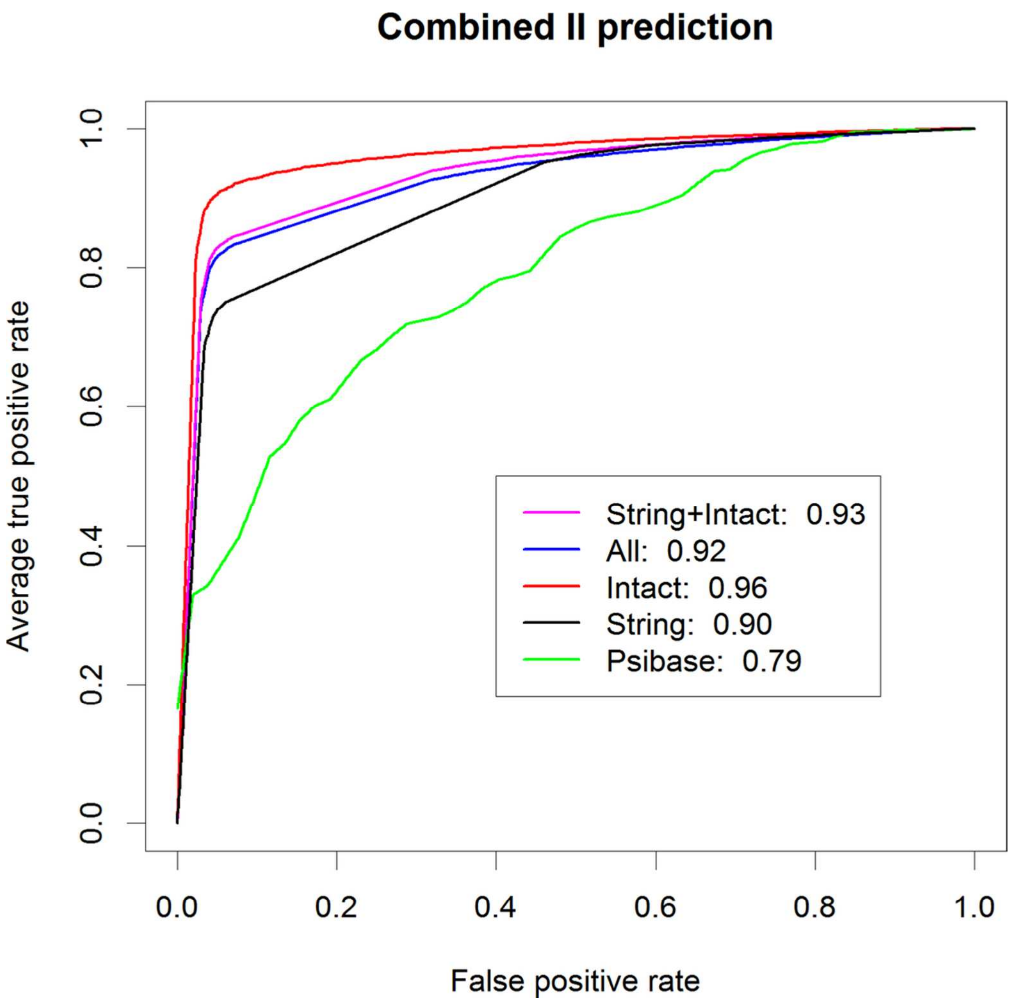
Figure 2 Combined II ROC curve. ROC curve corresponding to the metrics generated with the Blast+ parameter num_alignments set to 20 and minimum interaction pair metric value (min(ab)).

Figure 3 Sensitivity and specificity analysis for the Combined II metric ROC curve of the Intact pDB.

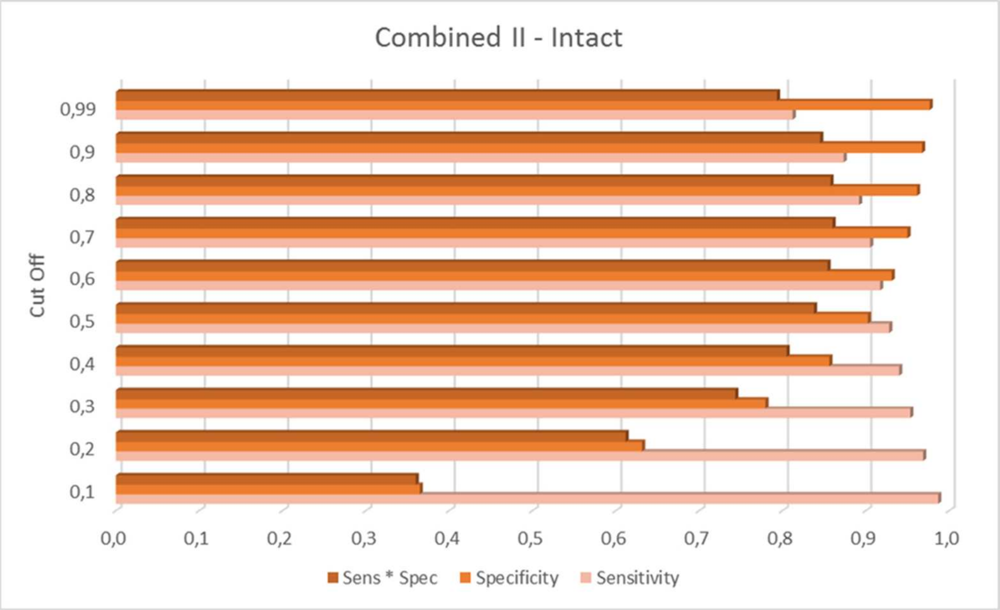
Figure 4 Sensitivity and specificity analysis for the Combined II metric ROC curve of the String+Intact pDB.



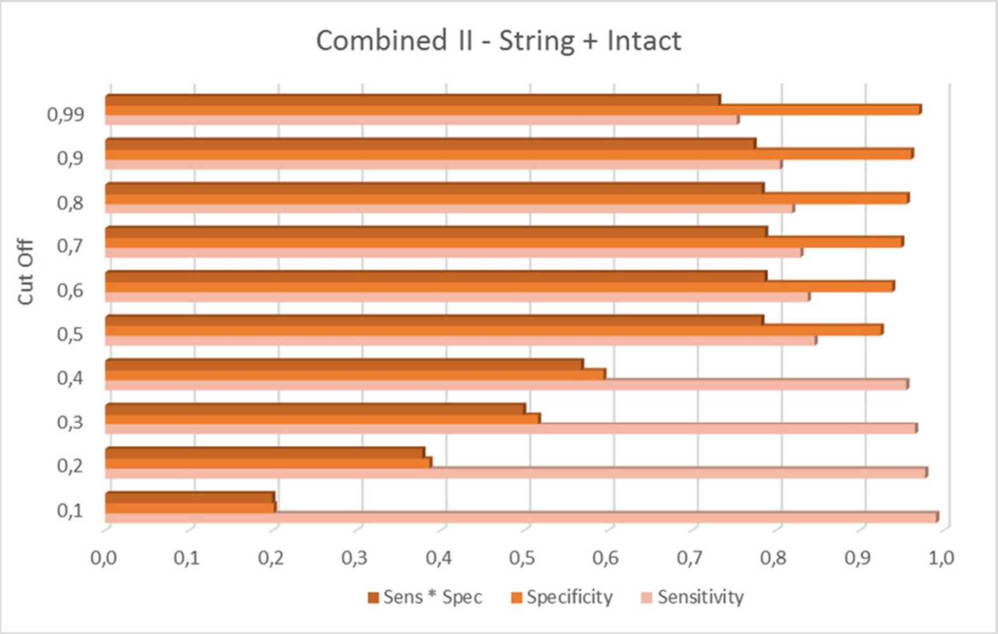
Distribution of Blast+ alignments grouped by number of hits. The alignments was generated with the Blast+ parameter num_alignments set to 20. All pDB: is the sum of String, Psibase and Intact. (*) Alignments in which the coverage to identity ratio is above 80%.
46x28mm (600 x 600 DPI)



Combined II ROC curve. ROC curve corresponding to the metrics generated with the Blast+ parameter num_alignments set to 20 and minimum interaction pair metric value (min(ab)).
51x50mm (600 x 600 DPI)



Sensitivity and specificity analysis for the Combined II metric ROC curve of the Intact pDB.
40x24mm (600 x 600 DPI)



Sensitivity and specificity analysis for the Combined II metric ROC curve of the String+Intact pDB.
40x25mm (600 x 600 DPI)