Positome: A Method for Improving Protein-Protein Interaction Quality and Prediction Accuracy

Kevin Dick[1], Frank Dehne[2], Ashkan Golshani[3], James R. Green[1]
[1]Department of Systems and Computer Engineering
[2]School of Computer Science
[3]Department of Biology, Institute of Biochemistry

Carleton University
Ottawa, Canada
kevin.dick@carleton.ca, jrgreen@sce.carleton.ca

***Abstract*—The progressive elucidation of positive protein-protein interactions (PPIs) as wet-lab techniques continue to improve in both throughput and precision has increased the number and quality of known PPIs across the spectrum of life. Creating high quality datasets of positive PPIs is critical for training PPI prediction algorithms and for assessing the performance of PPI detection efforts. We present the Positome, a web service to acquire sets of positive PPIs based on user-defined criteria pertaining to data provenance including interaction type, throughput level, and detection method selection in addition to filtration by multiple lines of evidence (*i.e.* PPIs reported by independent research groups). The Positome provides a tunable interface to obtain a specified subset of interacting PPIs from the BioGRID database. Both intra- and inter-species PPIs are supported. Using a number of model organisms, we demonstrate the trade-off between data quality and quantity, and the benefit of higher data quality on PPI prediction precision and recall. Available at http://bioinf.sce.carleton.ca/POSITOME/.**

Keywords—protein-protein interaction prediction; data quality; datasets; data provenance; machine learning

I. Introduction

Life is enabled by the cellular dynamics of protein-protein interactions (PPIs) as they govern fundamental biological processes such as cellular division, protein transport, and signal transduction. The elucidation of PPIs is critical to understand fundamental biology across the tree of life, and resolve human disease and infection. Since the number of putative PPIs grows as the square of the number of proteins (>250M pairs in human), and experimental validation techniques are resource-intensive and error-prone [1], computational approaches are leveraged to better resolve the interactome and supplement our current knowledge base. Of these putative pairs, only a small proportion are expected to physically interact, making this a very difficult prediction problem with high class imbalance.

Machine learning has seen effective application to a range of problems in molecular biology, including PPI prediction. A broad range of PPI prediction paradigms exist including sequence-based [2], structure-based [3], network-based [4] (which leverages graph topology to make inferences), ontology-based [5], and evolution-based methods [6]. Additionally, the merging of prediction paradigms is increasingly popular [7]–[9]. Common to each is the requisite dataset upon which to train and test the predictive performance and the appropriate definition of what constitutes a positive PPI must be enforced. However, establishing consistent performance evaluations is challenging given the variety of machine learning algorithms, evaluation methods, and data processing procedures [10][11]. The data used to inform predictive methods, such as the training dataset of positively or negatively interacting PPIs, lack standardization as these data are obtained from a multitude of sources often collected using a broad range of experimental techniques using different approaches (*i.e. in vitro*, *in vivo*, *in silico*) [12]. Machine learning performance on biological data is subject to the quality and quantity of the data used to develop the predictive model [13]. Noisy datasets have produced unexpected associations, prompting the development of statistical measures to “de-noise” resulting networks [14]. Additionally, in an effort to augment datasets, the “pooling” of data from disparate data sources can result in data duplication, resulting in over-representation, [15] or introduction of noise from conflicting data definitions.

Independent datasets amass their content using various approaches such as manual curation and data deposits [16]–[18], data mining methods [19], or some combination of the two [20]. Increasingly, certain public repositories have merged their data in an effort to provide a consolidated resource for the scientific community [21]. General-purpose databases will amalgamate information from several sources, notably databases dedicated to the study of particular organisms or a subset of taxonomies [22]–[28], in an effort to provide increasingly complete coverage.

IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2017), Manchester, UK, 23-25 August 2017

978-1-4673-8988-4/17/$31.00 ©2017 IEEE

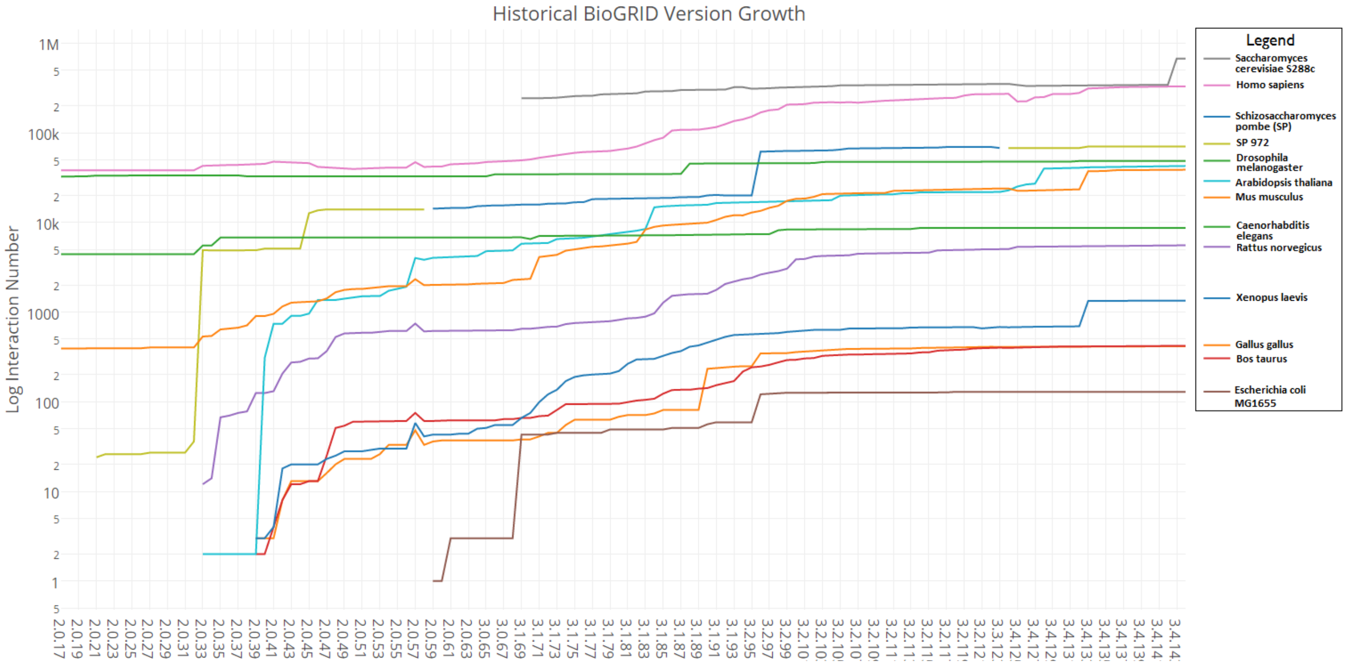

Figure 1 - Logarithm Interaction Number Growth of 12 Model Organism over the Entire BioGRID Release History. We note that *Schizosaccharomyces pombe* appears to be represented interchangeably under two unique taxonomy ids.

The large class imbalance between the expected number of protein pairs versus the expected number of physically interacting proteins leads to a difficult challenge in the elucidation of interactomes. In this case, a lenient false positive rate would result in an excessive number of false positives overwhelming any novel discoveries; the utility of such computational approaches is thereby lost. Adding to this difficulty is the fact that many reported PPIs are, in fact, false. This is particularly true of databases that include PPIs predicted through previous computational methods that are known to be imperfect, or data arising from early high-throughput PPI elucidation studies that made use of wet-lab techniques that we now know to be noisy and error-prone. For example, Yeast Two-Hybrid (Y2H) experiments have been found to have a false discovery rate of 9.9% and a false negative rate of 51% [29], implying that approximately one in every ten PPIs reported by Y2H have been mislabelled as a positive and that many true positive PPIs are being missed. By including mislabelled data in the training of PPI predictors, the effectiveness of these methods is dramatically reduced, and the estimated performance will be inaccurate.

We here propose a method of systematically engineering datasets of known PPIs for the purpose of training and evaluating PPI prediction methods. Several *ad hoc* approaches to improving the quality of positive PPI training data have been reported during the development of specific methods [13], [30], [31]. For example, some groups have restricted positive PPIs to only those physical interactions elucidated by means of high confidence detection methods and which have been independently reported in the literature on multiple occasions [32], [33]. Similarly, other groups have attempted to create standardized sets of negative non-interacting protein pairs, most notably the Negatome resource [19]. We here develop a systematic method to create high quality positive PPI datasets in a manner which is independent of the subsequent PPI prediction method. Our approach is integrative and allows the specification of several parameters to generate highly tailored datasets using a web-based interface not offered in other frameworks or PPI database "advanced search" functions.

Typically in machine learning, the inclusion of an increased amount of data will result in improved performance, since the model can learn about natural variation in the data. However, many machine learning methods are susceptible to noisy data. The incorporation of noisy data will typically result in performance loss. Filtering available reported PPIs can result in higher quality training data, but this will necessarily also lead to reduced quantity of data available for developing the machine learning method. This quality-quantity trade-off in the data leads to a commensurate precision-recall trade-off in the resulting prediction method.

Data provenance presents a considerable challenge in computational biology, as the majority of information contained in public databases results from a series of interpretations and transformations originating from experimental observation. Data provenance considers the question of "where did the data come from?" in the attempt to establish some amount of confidence in the published results. By grouping evidence from multiple sources pertaining to a given interaction, scoring metrics can be derived to quantify the strength of evidence for that PPI [34]. However, the scientific community has not yet adopted a standardized scoring measure for molecular interactions [35]. Without a definitive measurement of PPI data quality, researchers should be provided a flexible mechanism to build quality-assured datasets.

Here, we develop a web service enabling researchers to generate customized PPI datasets based on a number of tunable parameters to specify the inclusion criteria for the resulting positive PPI set. These PPI filtration parameters examine data provenance in limiting the types of included interactions in addition to ensuring that the data is supported by multiple independent lines of enquiry (if specified). The service can create both intra- and inter-species datasets for any organisms listed in the Biological General Repository for Interaction Datasets (BioGRID) [36]. We propose that the selection of a subset of high quality positive PPIs for the development of machine learning models will result in improved performance, particularly where the false positive rate must be conservatively limited. The Positome web service is available at http://bioinf.sce.carleton.ca/POSITOME/.

II. Methods

A. Acquision of Protein-Protein Interactions

The BioGRID group curates protein-protein interactions from primary biomedical literature [36]. This effort is to report the results of published experiments, noting that they cannot guarantee that any one PPI truly interacts, is well-established, or adheres to consensus in the scientific community. In this way, users of the BioGRID must perform their own quality assessment of this dataset. Since its inception in 2006, the dataset has steadily grown to now contain 1,418,871 interactions drawn from 48,312 publications (as of time of writing; BioGRID version 3.4.146). As a compendium of literature-derived PPIs, updated on a monthly basis, the BioGRID is an ideal dataset for the purposes of training and evaluating machine learning PPI predictors for various tasks such as genome-wide prediction [37], protein function prediction [38], or interactome-wide analyses [39][32]. The Positome tool described here provides a quality filtration layer to compliment the BioGRID curation efforts to improve PPI quality and prediction accuracy.

B. Database Interaction Number Historical Growth

The BioGRID database provides an archive of historically released datasets. To intuitively capture the historical growth of the database across the species it currently supports since its inception, prior archives were analyzed for intra-species interactions. The interaction count is illustrated in Fig. 1 on a log scale for a subset of 12 model organism (an interactive figure with the inclusion of all 61 organisms can be found at http://bioinf.sce.carleton.ca/POSITOME/publication/figures/biogrid_versions_plot_all.html).

C. Positome Pipeline

The Positome filters the current BioGRID dataset of positive interactions based on the specified user-defined parameters. Two files are generated: *protein_pairs.txt* and *protein_sequences.txt,* where the pairs file contains the filtered binary interactions resulting from the user-defined parameters and the sequences file contains the primary amino acid sequence for all proteins occurring in the pairs file. A FASTA formatted alternative is also provided.

The Positome web service allows users to specify whether they wish to include interactions acquired using each of 28 different experimental methods. Additionally, a user can specify the inclusion of either ‘physical’, ‘genetic’, or both interaction types, and the experimental throughput level of either ‘high’, ‘low’, or both. Finally, the user can specify whether they would like to apply “Multiple Lines of Evidence”

Algorithm 1 The Positome Algorithm

Input: User-Defined Parameters **as** config
Output: protein_pairs.txt, protein_sequences.txt

```
BioGRID Filtration :
PPI_list = empty list
$det$ = config.detectionMethodList
$inter$ = config.interactionType
$thru$ = config.throughputLevel
$data$ = open(current BioGRID version)
for PPI in $data$ do
    $A$ = PPI.firstProtein
    $B$ = PPI.secondProtein
    if ($A$ == config.taxA && $B$ == config.taxB) or
        ($A$ == config.taxB && $B$ == config.taxA) then
        $d$ = PPI.dectectionMethodList
        $i$ = PPI.interactionType
        $t$ = PPI.throughputLevel
        if $d$ in $det$ && $i$ in $inter$ && $t$ == $thru$ then
            PPI_list.append(PPI)
        end if
    end if
end for
if config.multipleLinesOfEvidence then
    return duplicates(PPI_list)
end if
return unique(PPI_list)

Mapping to Uniprot Accession Number :
map_cache = load previously mapped proteins
mapped_PPIs = empty list
for PPI in PPI_list do
    uniprot_A = map_cache(PPI.A)
    uniprot_B = map_cache(PPI.B)
    uniprot_PPI.A = uniprot_A
    uniprot_PPI.B = uniprot_B
    if uniprot_PPI not in mapped_PPIs &&
        reverse(uniprot_PPI) not in mapped_PPIs then
        mapped_PPIs.append(uniprot_PPI)
    end if
end for

Sequence Acquisition :
sequence_cache = load previously acquired sequences
unique_proteins = empty list
for PPI in mapped_PPIs do
    if PPI.A not in unique_proteins then
        seq_A = sequence_cache(PPI.A)
        unique_proteins.append(PPI.A, seq_A)
    end if
    if PPI.B not in unique_proteins then
        seq_B = sequence_cache(PPI.B)
        unique_proteins.append(PPI.B, seq_B)
    end if
end for
write(protein_pairs.txt, mapped_PPIs)
write(protein_sequences.txt, unique_proteins)
```

(MLoE) filtration, which conserves only those interactions that are reported by multiple independent research groups.

Algorithm 1 defines the various steps in the generation of the resulting files, capturing relevant information about the replication of PPI findings in independent studies to inform the multiple lines of evidence filtration. Lines 1-21 describe the preliminary filtration of the dataset according the user-defined parameters in a single pass. Hashing the interactions in a dictionary as a key-value pair with their count allows single-pass detection of duplicate PPIs (where ||PPI_list[key]|| ≥ 2; see line 19). Simply returning the dictionary keys produces a list of unique PPIs. Hashing is done in such a way where both the forward, AB, and reverse, BA, PPIs hash to the same key.

Lines 22-32 describe the conversion process from Entrez identifier to Uniprot Accession Number using a locally cached mapping, a resource made available to the community to link users back to the source database. Here, duplicates may arise given that multiple Entrez identifiers might map to the same Uniprot Accession Number thereby potentially generating a duplicate PPI. Again, both forward and reverse binary interactions are considered to produce the final non-redundant set of unique PPIs.

Finally, lines 33-46 obtain the sequences corresponding to the unique proteins appearing in the binary interaction pairs. To ensure usability and timely generation of results, the majority of this work is cached and unique occurrences of the proteins and their sequences are extracted. This data are then written to file and made available for download.

D. *Positome Features*

1) Provides both Inter- and Intra-species PPIs

The BioGRID dataset curates PPIs both within a given organism (intra-species PPI) and between two organisms (inter-species PPI). The Positome allows users to specify whether they wish to obtain inter- or intra-species interactions.

2) Recommended Parameters

While the Positome interface permits great flexibility in selecting the data quality filtering options, two recommended parameter sets are provided: "Conservative" and "Permissive" which we here use to generate various datasets to validate the Positome. The permissive level sets all filters to their most inclusive settings to ensure the maximum number of PPIs are included. The conservative setting limits the results to physical interactions, both high and low throughput levels, applies MLoE filtration, and includes a subset of experimental detection methods which we postulate to provide a higher confidence in the reported physical interaction (Table I).

For our work, we are interested in how these two definitions relate to the change in data quality over time. Fig. 2 illustrates the historical change in data quality for *Homo sapiens* when these two definitions are applied retroactively to the data at each historical release.

3) Timely Generation of Results

In order to deliver timely results, the majority of the data processing and conversion tasks are cached locally. The number of interactions in various organisms ranges between less than ten, to over half a million. To ensure that data requests are returned in a reasonable time frame, the Positome pipeline is optimized to operate with a time complexity of O(n), where n is the number of known PPIs.

4) Job Completion Notification/Recovery

The Positome assigns a unique "job id" for every request. This ensures that multiple queries are handled independently and can be later retrieved by individual researchers through a results retrieval interface. An email notification is optionally sent when the job is complete.

TABLE I. DATASET INCLUSION PARAMETERS

Inclusion Parameters	**Datasets**	
	Permissive	***Conservative***
Interaction Type	Both Physical and Genetic	Physical only
Detection Methods	All	Two-hybrid, Affinity Capture-MS, Affinity Capture-Western, Reconstituted Complex, Affinity Capture-Luminescence, Co-crystal Structure, Far Western, FRET, Protein-peptide, Co-localization, Affinity Capture-RNA, Co-purification
Throughput Level	Both High and Low	Both High and Low
Multiple Lines of Evidence	No	Yes

5) Pre-Computed Datasets

A number of pre-computed datasets are made available for researchers for a number of species. These datasets are defined based on the permissive and conservative parameters defined in Table I. These are meant to serve as starting points for researchers looking to build their own customized datasets.

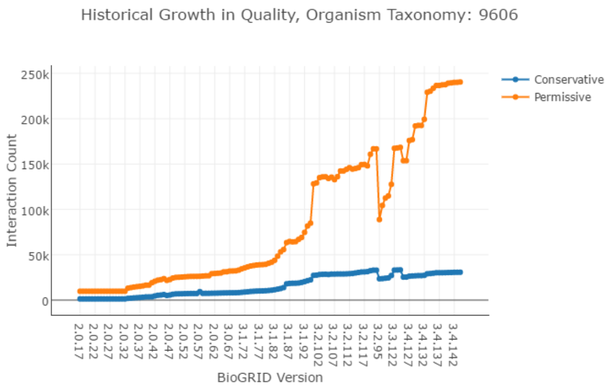

Figure 2 - Historical Growth in Data Quality in *Homo sapiens*. A substantial difference between the permissive and conservative dataset sizes can be observed across releases. Growth in PPI data quality is not keeping pace with growth in quantity of reported PPIs.

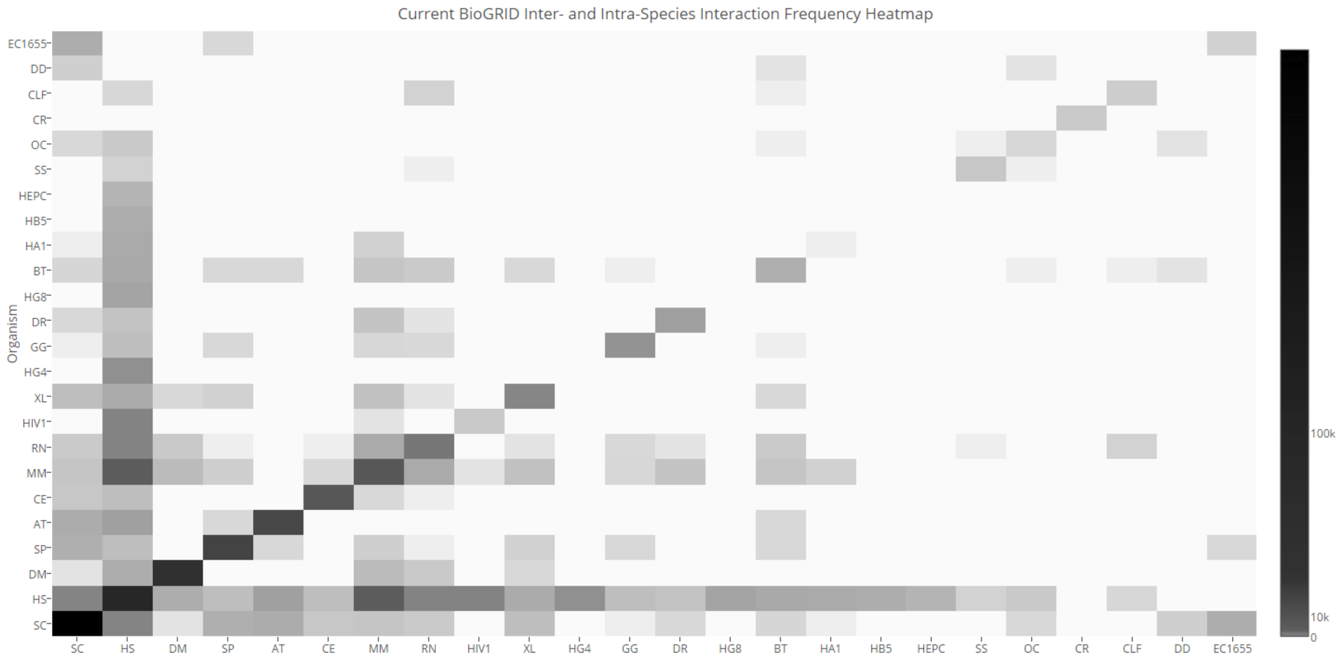

Figure 2 - Heatmap Representation of Inter- and Intra-Species PPIs in the Current BioGRID Release. Organisms with at least eight intra-species interactions were included. Abbreviated organism codes can be found at **http://bioinf.sce.carleton.ca/POSITOME/supplemental.html**.

E. Evalution of Datasets

We here demonstrate that the improved quality of Positome datasets leads to increased performance of PPI prediction methods. The increase in performance was established by comparing the results of a PPI prediction tool trained using both the "permissive" and "conservative" datasets for five model organisms: *Homo sapiens*, *Saccharomyces cerevisea*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Mus musculus*.

The Protein-protein Interaction Prediction Engine (PIPE) developed by the Carleton University Bioinformatics Research Group [32], [39]–[41] was used to train and evaluate the quality of the dataset. Prevalence-Corrected Precision-Recall (PR) curves are used to quantify prediction accuracy.

For each organism, the "permissive" and "conservative" datasets were obtained and a PIPE model trained on each, with negative PPIs defined as in [40]. Leave-one-out cross validation (LOOCV) was performed on each dataset to compare performance when training the PIPE binary PPI classifier. Given the class imbalance in the datasets, precision-recall (PR) curves were used in favour of receiver-operator characteristic (ROC) curves, as recommended in [42]. The prevalence-corrected precision is used to account for the class imbalance when limited negative test data are used. A conservative prevalence of 100 negative PPIs per positive PPI was selected, based on previous work on the development of interactome-scale predictive models and reported estimates [43]–[45]. While the entire PR curve is informative, only a portion of the curve is of relevance when users demand a minimum level of precision. For example, a predictor that is incorrect most of the time (Pr < 50%) is not a useful predictor. Here, the recall (*i.e.* sensitivity) at a prevalence-corrected precision of 50% (Re@Pr50) was used to determine superior performance between the conservative and permissive datasets for each organism.

The permissive method was able to leverage significantly more training data than the conservative set, while the conservative method benefitted from higher training data quality. However, to ensure that both the conservative and permissive methods are evaluated in a consistent way, only performance over PPIs appearing in both sets were used for evaluating the methods (computing PR curves). We therefore define a modified LOOCV for the evaluation of the permissive model. Here, we perform the LOOCV using only those proteins within the intersection of the permissive and conservative sets (*i.e.* given conservative set, C, and permissive set, P, where $C \subseteq P$). Each known PPI from the intersection set is removed, the remaining PPIs are used to retrain the respective models, and the removed PPI is predicted. In maintaining equal testing set conditions, the two datasets are appropriately compared using the PR curves and Re@Pr50.

III. Results and Discussion

Machine learning requires consistent data, free of noise or mislabeled data, to generate models capable for meaningful prediction tasks. To address challenges in data provenance a flexible method to specify inclusion parameters allows researchers to build quality-assured datasets. Here we introduce a scientific tool as a layer between a dataset of mixed quality and specific machine learning methods to enable the

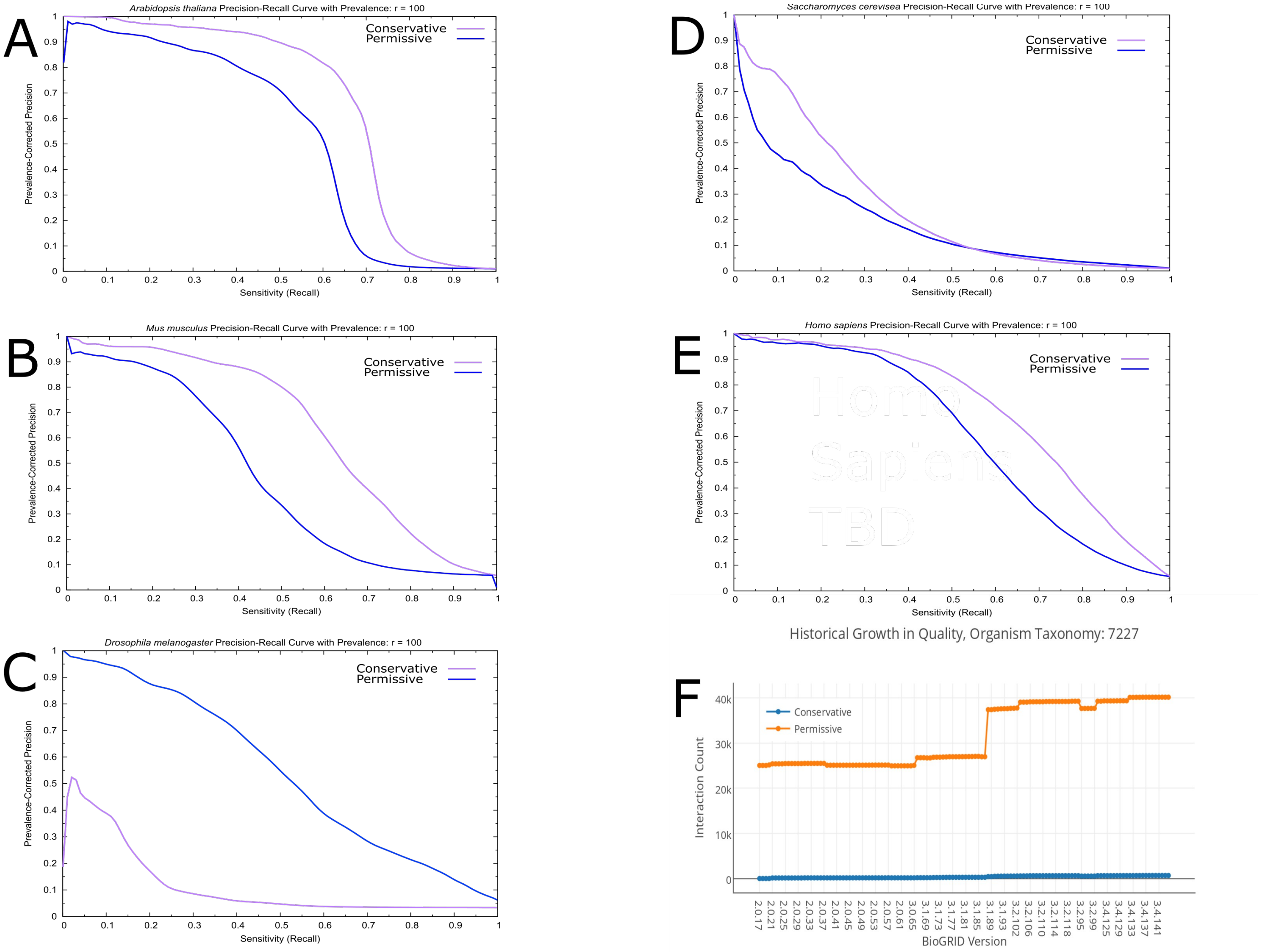

Figure 3 - Experimental Validation Results for Five Organisms. A-E are Prevalence-Corrected Precision-Recall curves comparing the LOOCV results of the conservative and permissive datasets when used to create PIPE PPI predictors. Figure F depicts the historical growth of dataset sizes across the BioGRID history for *Drosophila melanogaster.*

scientific community to produce customizable and consistent data sets suited to the task of prediction.

A. BioGRID Source Data Quality vs. Quantity over Time

Fig. 1 illustrates the change in BioGRID composition across 12 model organisms over its archived history and notable trends emerge. First, we remark that some model organisms (*e.g. H. sapiens, S. cerevisea)*, as expected, are overrepresented compared relative to the other organisms. Interestingly, we observe the periodic introduction of new organisms in addition to substantial increases in interaction count at various points, presumably the result of high throughput experiments being published for a given organism. It is important to note that this figure is plotted on a logarithmic scale; the small variances in the upper regions corresponding to increases or decreases of thousands or tens of thousands of interactions. With the active study of intra- and inter-species interactions, we anticipate a gradual increase in both the quality and quantity of interactions in the BioGRID database. As the cost of high-throughput methods continue to decline, more studies on non-model organisms also become feasible.

The current composition of the BioGRID dataset (version 3.4.146), is illustrated in Fig. 3 using a logarithmic heatmap. The diagonal captures intra-species interactions while the remaining cells illustrate inter-species interactions. The matrix is sparsely populated indicative that the majority of research has been on intra-species interactions and certain bands with greater representation indicate the bias towards studying intra-species interactions with a small subset of organisms (*e.g. Homo sapiens*). Interestingly, some inter-species combinations appear to be strictly *in vitro*, such as between *H. sapiens* and *M. musculus*.

B. Evalution of Positome Datasets for PPI Prediction

LOOCV tests were used to evaluate the impact of Positome data quality filtering for the task of binary classification of PPIs. PIPE trained on the conservative dataset outperforms the permissive dataset across four of the five organisms considered (Fig. 4A-E, Table II). We concern ourselves primarily with the upper half of the PR curves, where the prevalence-corrected precision is greater than 50%, as this pertains to use cases where the number of predicted positives are actually true in the majority of predictions. In reality, we would limit ourselves to a more conservative margin depending on the cost associated

TABLE I. Performance Evaluation Between Conservative and PErmissive Datasets in Five Organisms

Organisms	Recall at 50% Prevalence-Corrected Precision	
	Permissive	***Conservative***
M. musculus	0.417	0.640
A. thaliana	0.608	0.708
S. cerevisea	0.066	0.211
D. melanogaster	0.535	0.032
H. sapiens	0.596	0.741

to predicting a false positive. As we can see in *M. musculus*, *A. thaliana*, *S. cerevisea*, and *H. sapiens*, the conservative set consistently outperforms the permissive set within this range, Pr = [0.5, 1]. Interestingly, the *H. sapiens* performance converges in both sets for precision values in the range [0.9, 1] indicative that the quality-quantity trade-off in this range is less well-defined (Fig. 4E). These findings support the hypothesis that a smaller number of high quality interactions offer superior prediction performance over the use of larger quantities of lower quality data. It is primordial that researchers concern themselves with data provenance so as to use datasets consistent with their efforts.

In the case of *Drosophila melanogaster*, training on the conservative dataset actually results in significantly poorer performance. We attribute this to the fact that the conservative parameters are excessively stringent for this organism. For *D. melanogaster*, only 568 PPIs pass the conservative filters whereas the permissive dataset contains 37,251; this represents approximately a 65-fold increase in data quantity (Table III). In fact, considering the historical growth in PPI number throughout the BioGRID version history, we remark that the conservative dataset remains consistently low relative to the permissive set (Fig. 4F). This illustrates the quality-quantity trade-off where overly-severely limiting a training dataset to a minute dataset of high-quality interactions is, in fact, more harmful than beneficial to predictive performance. Lessening the stringency of the conservative parameters by including the “Phenotypic Suppression” (2.79 fold increase in PPI number) in addition to “Phenotypic Enhancement” (4.29 fold increase in PPI number) detection methods, the two methods which contribute the largest number of PPIs in the organism, to the *Drosophila melanogaster* dataset, we observe a fold difference of 16.3 and 11.4 with the permissive set, respectively; such a fold-increase is more in line with the other four organisms (Table III). This is suggestive that obtaining a balance of appropriate filtration parameters for various species is necessary to obtain an appropriate quantity-quality trade-off. For this reason, beyond supplying the standard conservative and permissive filter sets, the Positome interface permits users to fine-tune the various filter settings to suit their specific application.

C. *Quantification of the Strength of Publication*

The Positome offers MLoE filtration which, when enabled, retains only those PPIs reported by two or more publications, as determined by PubMed IDs in BioGRID. Quantifying the strength of a publication and appropriately synthesizing evidence in support of a PPI remain as future work. The STRING database offers a “combined score” which leverages evidence from phylogenetic co-occurrence, experimentation, literature curation, and genomic context [20]. This is an actively debated problem and subject to further investigation in the effort to resolve data provenance challenges.

TABLE II. Dataset Magnitude in Five Organisms

Organism	||Permissive||	||Conservative||	Fold Difference
M. musculus	6,434	1,056	5.09
A. thaliana	13,673	1,328	9.30
S. cerevisea	192,490	12,175	14.8
D. melanogaster	37,251	568	64.6
H. sapiens	65,786	14,701	3.47

D. *Future Directions*

The Positome is designed to automatically update every month to use the latest BioGRID dataset. With monthly updates from BioGRID, the Positome will leverage the continual increase in known interactions. It is expected that the incorporation of PPIs from non-model organisms and from inter-species studies will progressively increase. Thereby, the Positome’s value as a research tool will increase with time. Future work will investigate the incorporation of data from additional sources in the effort to address data provenance challenges and the development of standard filter settings tuned to specific species for which the default conservative set is inappropriate. Additionally, a visualization of each stage of filtration will be offered to further clarify data provenance of the resulting PPI dataset. Finally, a machine-accessible (REST) interface, in combination with an application programming interface, will be developed.

IV. Conclusion

Creating high quality datasets of positive PPIs is critical for training PPI prediction algorithms to discover new meaningful interactions. In an effort to address data provenance challenges in PPI datasets, we developed the Positome, a web service to acquire sets of positive PPIs based on user-defined criteria supporting both intra- and inter-species interactions. Using a number of model organisms, we demonstrate the trade-off between data quality and quantity and the benefit of higher data quality on PPI prediction precision and recall. Cross-validation tests were used to detect overfitting and are the computational equivalent to wet-lab validation to confirm our results. Although the PIPE PPI prediction method is used to demonstrate the increase in prediction performance with increasing data quality, it should be understood that the Positome web service is in no way restricted to a single PPI prediction paradigm.

Acknowledgment

The authors would like to thank the Carleton University Bioinformatics Research Group for their support of this work.

References

[1] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright,

"Computational Prediction of Protein–Protein Interactions," *Mol. Biotechnol.*, vol. 38, no. 1, pp. 1–17, Jan. 2008.

[2] S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne, "Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions," *PLoS One*, vol. 4, no. 11, p. e7813, Nov. 2009.

[3] Q. C. Zhang, D. Petrey, J. I. Garzon, L. Deng, and B. Honig, "PrePPI: a structure-informed database of protein-protein interactions," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D828–D833, Jan. 2013.

[4] J. Wang *et al.*, "A protein interaction network for pluripotency of embryonic stem cells," *Nature*, vol. 444, no. 7117, pp. 364–368, Nov. 2006.

[5] X. Luo, H. Al-Mubaid, and S. Bettayeb, "Ontology based semantic similarity for protein interactions," in *Proceedings of BICOB-2013 Int'l Conf on Bioinformatics and Computational Biology*, 2013.

[6] Z.-W. Li, Z.-H. You, X. Chen, J. Gui, and R. Nie, "Highly Accurate Prediction of Protein-Protein Interactions via Incorporating Evolutionary Information and Physicochemical Characteristics," *Int. J. Mol. Sci.*, vol. 17, no. 9, p. 1396, Aug. 2016.

[7] Q. Zhong *et al.*, "An inter-species protein–protein interaction network across vast evolutionary distance," *Mol Syst Biol*, vol. 12, 2016.

[8] J. Yu, M. Vavrusa, J. Andreani, J. Rey, P. Tufféry, and R. Guerois, "InterEvDock: a docking server to predict the structure of protein–protein interactions using evolutionary information," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W542–W549, Jul. 2016.

[9] K. Dick and J. Green, "Comparison of sequence- and structure-based protein-protein interaction sites," in *2016 IEEE EMBS International Student Conference (ISC)*, 2016, pp. 1–4.

[10] T. Hamp and B. Rost, "More challenges for machine-learning protein interactions," *Bioinformatics*, vol. 31, no. 10, pp. 1521–1525, May 2015.

[11] P. Larranaga *et al.*, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, Feb. 2006.

[12] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "Protein-protein interaction detection: methods and analysis.," *Int. J. Proteomics*, vol. 2014, p. 147648, 2014.

[13] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins Struct. Funct. Bioinforma.*, vol. 63, no. 3, pp. 490–500, Jan. 2006.

[14] A. Yambartsev *et al.*, "Unexpected links reflect the noise in networks," *Biol. Direct*, vol. 11, no. 1, p. 52, Dec. 2016.

[15] J. J. Yang, O. Ursu, C. A. Lipinski, L. A. Sklar, T. I. Oprea, and C. G. Bologa, "Badapple: promiscuity patterns from noisy evidence," *J. Cheminform.*, vol. 8, no. 1, p. 29, Dec. 2016.

[16] S. Kerrien *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D841–D846, Jan. 2012.

[17] A. Chatr-aryamontri *et al.*, "MINT: the Molecular INTeraction database," *Nucleic Acids Res.*, vol. 35, no. Database, pp. D572–D574, Jan. 2007.

[18] A. Chatr-aryamontri *et al.*, "The BioGRID interaction database: 2017 update," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D369–D379, Jan. 2017.

[19] P. Blohm, G. Frishman, and P. Smialowski, "Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis," *Nucleic acids*, 2013.

[20] D. Szklarczyk *et al.*, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, Jan. 2017.

[21] S. Orchard *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D358–D363, Jan. 2014.

[22] T. Sheppard, B. Hitz, S. Engel, and G. Song, "The Saccharomyces genome database variant viewer," *Nucleic acids*, 2016.

[23] M. McDowall, M. Harris, and A. Lock, "PomBase 2015: updates to the fission yeast database," *Nucleic acids*, 2014.

[24] K. Howe, B. Bolt, S. Cain, J. Chan, and W. Chen, "WormBase 2016: expanding to enable helminth genomic research," *Nucleic acids*, 2015.

[25] H. Attrill, K. Falls, J. Goodman, and G. Millburn, "FlyBase: establishing a Gene Group resource for Drosophila melanogaster," *Nucleic acids*, 2015.

[26] C. Bult, J. Eppig, J. Blake, and J. Kadin, "Mouse genome database 2016," *Nucleic acids*, 2016.

[27] L. Ruzicka *et al.*, "ZFIN, The zebrafish model organism database: Updates and new directions," *genesis*, vol. 53, no. 8, pp. 498–509, Aug. 2015.

[28] T. Z. Berardini *et al.*, "The arabidopsis information resource: Making and mining the 'gold standard' annotated reference plant genome," *genesis*, vol. 53, no. 8, pp. 474–485, Aug. 2015.

[29] H. Huang and J. S. Bader, "Precision and recall estimates for two-hybrid screens.," *Bioinformatics*, vol. 25, no. 3, pp. 372–8, Feb. 2009.

[30] Y.-A. Huang, Z.-H. You, X. Chen, and G.-Y. Yan, "Improved protein-protein interactions prediction via weighted sparse representation model combining continuous wavelet descriptor and PseAA composition," *BMC Syst. Biol.*, pp. 3–5, 2016.

[31] S. Jain *et al.*, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC Bioinforma. 2010 111*, vol. 20, no. 7, pp. 203–205, 2010.

[32] A. Schoenrock, F. Dehne, J. R. Green, A. Golshani, and S. Pitre, "MP-PIPE," in *Proceedings of the international conference on Supercomputing - ICS '11*, 2011, p. 327.

[33] T. Hamp and B. Rost, "Evolutionary profiles improve protein–protein interaction prediction from sequence," *Bioinformatics*, vol. 31, no. 12, pp. 1945–1950, Jun. 2015.

[34] J. M. Villaveces *et al.*, "Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study," *Database*, vol. 2015, no. 0, p. bau131-bau131, Feb. 2015.

[35] B. Aranda *et al.*, "PSICQUIC and PSISCORE: accessing and scoring molecular interactions," *Nat. Methods*, vol. 8, no. 7, pp. 528–529, Jun. 2011.

[36] C. Stark, B. Breitkreutz, T. Reguly, and L. Boucher, "BioGRID: a general repository for interaction datasets," *Nucleic acids*, 2006.

[37] I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte, "A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans," *Nat. Genet.*, vol. 40, no. 2, pp. 181–188, Feb. 2008.

[38] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Mol. Syst. Biol.*, vol. 3, Mar. 2007.

[39] S. Pitre *et al.*, "Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps," *Sci. Rep.*, vol. 2, pp. 686–93, Jan. 2012.

[40] S. Pitre *et al.*, "PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC Bioinformatics*, vol. 7, no. 1, p. 365, 2006.

[41] A. Amos-Binks *et al.*, "Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences."

[42] T. Saito, M. Rehmsmeier, L. Hood, O. Franco, R. Pereira, and K. Wang, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015.

[43] A. Schoenrock *et al.*, "Efficient prediction of human protein-protein interactions at a global scale."

[44] C.-Y. Yu, L.-C. Chou, D. Tien, and -Hao Chang, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins," *BMC Bioinformatics*, vol. 11, 2010.

[45] Y. Park, "Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences."