D750–D754 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 15 October 2010*
doi:10.1093/nar/gkq943

KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions

Xue-wen Chen[1,2,*], Jong Cheol Jeong[2] and Patrick Dermyer[2]

[1]Bioinformatics and Computational Life Sciences Laboratory, Information and Telecommunication Technology Center and [2]Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA

Received August 16, 2010; Revised September 28, 2010; Accepted September 30, 2010

ABSTRACT

KUPS (The University of Kansas Proteomics Service) provides high-quality protein–protein interaction (PPI) data for researchers developing and evaluating computational models for predicting PPIs by allowing users to construct ready-to-use data sets of interacting protein pairs (IPPs), non-interacting protein pairs (NIPs) and associated features. Multiple filters and options allow the user to control the make-up of the IPPs and NIPs as well as the quality of the resultant data sets. Each data set is built from the overall database, which includes 185 446 IPPs and ∼1.5 billion NIPs from five primary databases: IntAct, HPRD, MINT, UniProt and the Gene Ontology. The IPP set can be set to specific model organisms, interaction types and experimental evidence. The NIP set can be generated using four different strategies, which can alleviate biased estimation problems. Lastly, multiple features can be provided for all of the IPP and NIP pairs. Additionally, KUPS provides two benchmark data sets to help researchers compare their algorithms to existing approaches. KUPS is freely available at http://www.ittc.ku.edu/chenlab.

INTRODUCTION

Proteins are the fundamental units of life, which play significant roles in cellular processes such as composing cellular structure and promoting chemical reactions. The multiplicity of functions that proteins execute in most cellular processes and biochemical events is attributed to their interactions with other proteins. Thus, understanding protein–protein interactions (PPIs) will help to elucidate cellular processes and functions, identify pharmacological targets and design drugs.

While advanced high-throughput technologies are providing a large amount of new PPI data, the number of discovered PPIs is still far from complete. Therefore, there is a critical need to build *in silico* models that can predict PPIs existing *in vivo*. To develop and evaluate computational models for PPI prediction, two essential components are required: (i) positive data (interacting protein pairs or IPP) and negative data (non-interacting protein pairs or NIP) and (ii) features (attributes) for each data point (instance). While IPP data can be downloaded from some existing databases, generating NIP data is not a trivial task. Furthermore, there is no existing database that provides users with the second component, the features.

Negatome (1) is the first database that provides users with experimentally supported mammalian NIPs. These NIPs are derived from manually curated literature and protein complexes. Negatome is limited in developing PPI prediction algorithms because of the lack of IPPs and insufficient number of NIPs. To build a reliable PPI prediction model with generalization capabilities, the number of negatives is expected to be several orders of magnitude higher than the number of positives (2,3). For example, in yeast, ∼6000 proteins allow for ∼18 million potential interacting pairs; among them, <100 000 pairs are estimated to form actual interactions (2,4–6). With <2000 NIPs and zero IPPs, negatome data cannot be used as a sole source for training PPI prediction algorithms.

GRIP (7) is a web-based system that provides both IPPs and NIPs in *Saccharomyces cerevisiae*. In the GRIP, IPPs are extracted from the MIPS Comprehensive Yeast Genome Database (8), which includes only proteins in the same complex. NIPs are created by using uniform random sampling of proteins belonging to different subcellular locations. Although both IPPs and NIPs are available for users, GRIP database is limited in that the data are extracted from a single species (yeast) and are biased due to the mechanisms used in negative data

[*] To whom correspondence should be addressed. Tel: +1 785 864 4559; Fax: +1 785 864 3226; Email: xwchen@ku.edu

generation [studies show that choosing negative data in terms of different subcellular localizations leads to biased estimates of system performance (9)].

Another limitation existing in both the negatome and GRIP is that they both simply provide a list of protein pairs without features or attributions for each example. In algorithm development, feature extraction is one of the most important components that can significantly affect the performance of computational models. To facilitate the development and evaluation of new computational methods for PPI prediction, we develop KUPS (The University of Kansas Proteome Service) database, which addresses the existing issues by providing a significantly enlarged and enriched user-friendly database. Compared with the negatome and GRIP, KUPS allows users to create databases with larger scale (in terms of the number of IPPs and NIPs) and larger coverage (in terms of species involved). Furthermore, KUPS offers new options that are not available in both the negatome and GRIP: (i) users can construct ready-to-use data sets including both IPPs and NIPs with the associated features, rather than a simple list of protein pairs; (ii) users can construct data sets with examples of various levels of quality (e.g. high-throughput experiments, biochemical methods or inference) and different types (e.g. direct interaction and enzymatic reaction); and (iii) users can construct NIPs using four different strategies (discussed next) to alleviate the biased estimation problems. In addition, KUPS creates two benchmark data sets for researchers to develop and test their algorithms: one data set with balanced IPPs and NIPs and the other with imbalanced IPPs and NIPs, which reflects the observed sparsity of protein interactions. Table 1 summarizes the three databases. The KUPS is freely available through our website (http://www.ittc.ku.edu/chenlab/).

THE DATABASE

Overview

KUPS is built by integrating five databases [MINT (10), IntAct (11), HPRD (12), Gene Ontology (GO) (13) and UniProt (14)] into a primary database and two other databases [AAindex (15) and PSSM (16)] into a secondary database. Powerful filtering interfaces are provided to effectively handle the information from multiple databases and users. The structure of KUPS is shown in Figure 1.

Through front-end interfaces, users can specify three different groups of filters to effectively control the distributions and quality of positive (IPP) and negative (NIP) interacting protein pairs and the types of features (FEATURE) to use in the final output.

The primary database controls the population and distributions of examples. It integrates five independent databases: the three PPI databases (IntAct, MINT and HPRD) are used for extracting and generating IPPs and NIPs; GO is used for calculating the distance of annotations for generating NIPs and UniProt is used for referring annotations of individual proteins and their sequences. AAindex and PSSM are considered the secondary database in KUPS because they are not directly involved to define the list of final data set. However, they are used for generating features of the final PPI data. Finally, the

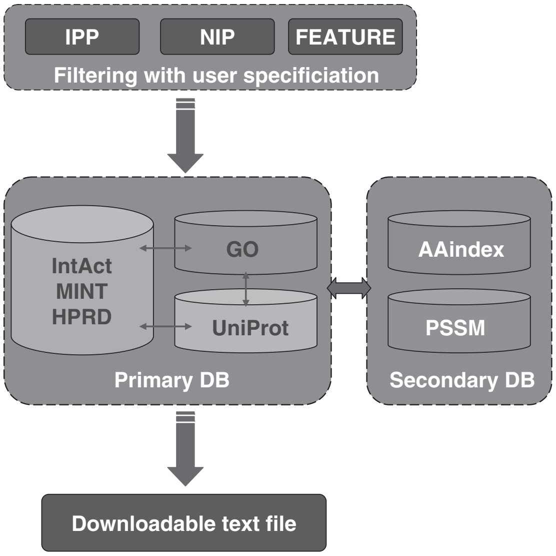

Figure 1. Structure of KUPS.

Table 1. Comparing three databases

	Databases		
Functions	Negatome	GRIP	KUPS
Users create IPPs?	No	Yes	Yes
Maximum no. of possible IPPs	0	10 994	185 446
Maximum no. of possible NIPs	1892	319 855	1.5 billion
Un-biased NIPs?	Yes	No	Yes
No. of model organisms	NA[a]	1	8
Choice of data quality and interaction types	No	No	Yes
Methods to create NIPs	Literature curation and structural information	Sub-cellular localization	Four strategies
Benchmarks	No	No	Yes
Feature extraction	No	No	Yes
Ready-to-use?	No	No	Yes

[a]Negative PPIs obtained from literature curation are extracted from mammalian proteins (most from human data); and negative PPIs obtained from PDB are extracted from mammalians (47%) and other species (53%).

generated IPP/NIP data set is written into a text file downloadable by users.

In addition, KUPS provides a data repository that includes benchmark data sets with evaluation results of commonly used learning models and testing data sets used in other studies.

Integrated databases

Five databases are integrated into KUPS as the primary database to generate both IPPs and NIPs. IPPs are extracted from three PPI databases (IntAct, MINT and HPRD) downloaded with a PSI-MI format. Essential information, such as species, gene name, interaction type and detection method, is extracted using our own parsing program. During the parsing process, a pair of interacting proteins is omitted if any one of the required information (i.e. protein names, interaction types and detection methods) is missed. Only unique pairs are stored in KUPS by avoiding redundancy among the three databases. After parsing, KUPS has stored ~185 446 PPI pairs from various organisms (the distribution for eight main model organisms is shown in Figure 2).

GO database consists of three categories of a gene product: molecular function, biological process and cellular component. The terms in the GO database are used to annotate database entries like UniProt to describe the roles of genes and gene products in any organism. GO also provides 'Evidence Codes' that are controlled vocabulary describing the nature of evidence to support a particular association. UniProt provides manually curated stable, comprehensive, freely accessible central resource on protein sequences and functional annotation and is often used as cross-references to other databases. UniProt is essential in KUPS by providing the links between PPI databases and GO database because some PPI databases do not provide annotations or related protein sequences.

KUPS provides not only IPPs and NIPs but also enriched information about each protein such as protein sequence, PSSM and converted real value features. This additional information of each unique protein is gathering from secondary database through the request of external users. In the secondary database, AAindex provides numerical scales of physicochemical and biochemical properties of amino acids by manually curating amino acid scales from research papers. Many studies (17–20) have used physicochemical and biochemical properties to elucidate cellular functions or processes; therefore, KUPS adopts AAindex to define features, which can be used for user's own purpose on designing *in silico* methods for PPI predictions. Another database, PSSM, was first introduced for detecting distantly related proteins based on the probability of amino acid appearance at each position by using previously aligned group of protein sequences or structures. Generating PSSMs is time consuming and normally requires a high performance machine and background knowledge about operating the software and machine. Therefore, KUPS provides PSSMs for each protein in the list of final data set by using PSI-BLAST (21). To create PSSM profiles in KUPS, the e-value and number of iterations are set to be 0.001 and 3, respectively. We use the default values set by PSI-BLAST for other parameters.

Filters with users' specification

KUPS provides various filtering processes for users to specify their needs. Figure 3 shows the workflow diagram with three filters: IPP filter for specifying positive interacting protein pairs; NIP filter for specifying non-interacting protein pairs and FUNC filter for specifying features to extract.

There are four adjustable parameters in IPP filter: number of IPPs, species, interaction types and detection

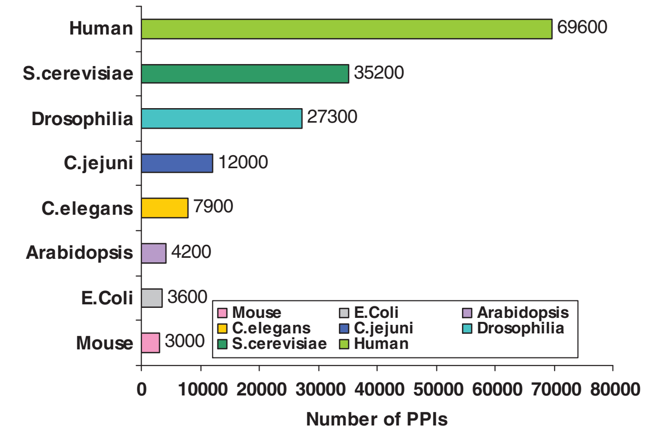

Figure 2. KUPS PPI distributions for model organisms.

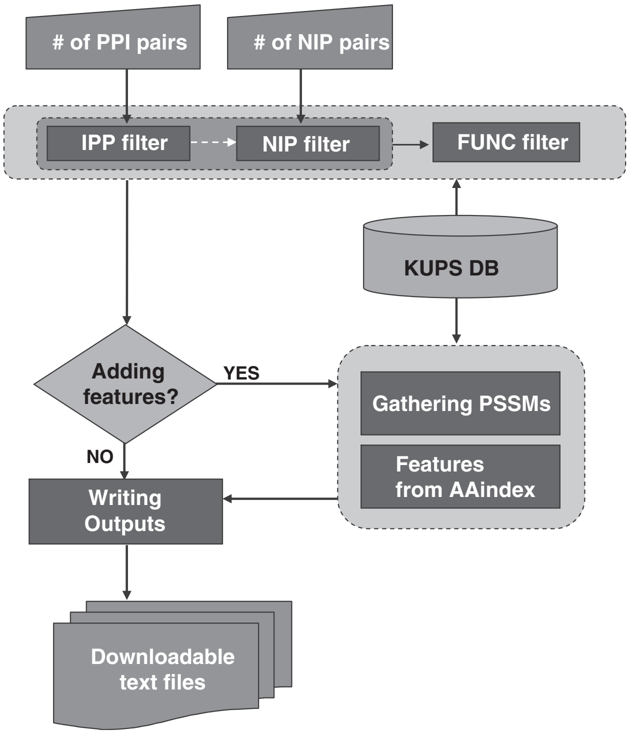

Figure 3. Workflow diagram in KUPS.

method. Except for the number of IPPs, all other parameters have multiple choices and by default all choices for each parameter are used. In KUPs, IPPs can be extracted from 13 species with different interaction types (e.g. physical association) and detection methods (e.g. biochemical methods or two hybrid methods). After users choose the parameters, KUPS searches the entire database and returns PPI pairs that match the filtering parameters.

The NIP filter controls the populations and distributions of non-interacting protein pairs. Besides the number of NIPs to generate, it has five more parameters that allow for producing eight different negative sets. By checking the first parameter 'Restrict protein selection pool', the NIPs are restricted to the combinations of proteins appeared in the previously generated IPPs. If this parameter is unchecked, then the entire database is used for generating NIPs. Four different methods are used for creating NIPs: uniform random pairs; functionally dissimilar pairs; spatially separated pairs and non-interacting domains.

Uniform random pairs are often considered less biased than molecular processes-based selection methods. They create NIPs by randomly selecting possible interacting pairs that are not appeared in known PPI lists.

Functionally dissimilar pairs are also known to be less biased than spatially separated NIP selection methods (9). KUPS first calculates semantic similarity between two proteins according to Lord *et al.* (22,23). The assumption is that the most dissimilar annotation pair will have the smallest similarity score. Potential non-interacting protein pairs are sorted by their similarity scores.

Spatially separate pairs define NIPs based on annotations of cellular component by selecting protein pairs that do not have any overlapping cellular component annotations.

Non-interacting domain pairs create NIPs based on non-interacting domain pairs defined in the negatome database (1): all proteins in KUPS are first mapped into Pfam (24) to identify their domains; KUPS selects protein pairs with non-interacting domains as NIPs.

Features generation

Besides creating the list of interacting and non-interacting protein pairs, KUPS provides eight groups of features for algorithm design and evaluation. This is an important component that allows users to focus on designing and evaluating computational algorithms for PPI predictions. The first six feature sets are 'Interaction type', 'Detection method', 'Species', 'Locality' and 'GO annotation' that describe the data users generate. The other three groups of features are extracted from protein sequences as described next.

Studies have shown that amino acid sequences contain significant information for characterizing protein interactions (25–31). However, extracting features from protein sequences is often time consuming and requires large computational power. KUPS provides several sequence-based features for users to evaluate. The first group of features is amino acid sequences for each protein. The second feature set is AAindex-based sequence features (15), which include the scales of physicochemical or biochemical properties. PSSM profile is another set of features that gives the log-odds score of each amino acid in each position of a target protein sequence. It has shown great potential in protein secondary structure prediction (32) and function prediction (33). KUPS provides PSSM profiles created by PSI-BLAST for each protein.

With KUPS, users can choose all or combinations of these eight groups of features and create a data set of IPPs and NIPs for algorithm development.

TEST SET COLLECTION AND BENCHMARK

KUPS has a collection of test sets used in research publications on PPI inference for researchers intending to compare their algorithms to other existing methods. The test set collection has three categories: methods for predicting PPIs, predicting protein functions and predicting protein interface residues. Each category has information about the published papers and downloadable data sets. Data sets are collected directly from authors' websites or Supplementary Materials of the paper.

Two benchmark data sets generated by this service are also included for comparing algorithm performance to existing solutions. The first is balanced with equal interacting and non-interacting protein pairs in both the training and test sets. The second is an imbalanced data set, with significantly more non-interacting pairs than interacting pairs. For machine-learning approaches, correct classification is easier in balanced data sets; however, the imbalanced data set more closely reflects the observed sparsity of interacting protein pairs. The performance of learning models on the benchmarks is presented with seven common measurements: overall accuracy, specificity, recall, precision, f-measure, correlation coefficient and confusion matrix. Results from four common learning models (Naïve Bayes, decision tree, support vector machine and random forests) are already available. Researchers are encouraged to share their algorithm's results to the benchmarks for inclusion on the website. The goal is to provide as complete a comparison between existing algorithms as possible.

CONCLUSIONS

To the best of our knowledge, KUPS is the first database that provides ready-to-use data sets for researchers to use. Compared with the other two existing databases, KUPS offers larger coverage (number of organisms where PPIs are derived) and larger scale (total number of PPIs). It allows users not only to generate positive and negative PPI data but also to create associated features as well, while other databases often provide users with a list of positive and negative PPI pair names only. KUPS allows for a fair comparison and evaluation of different learning methods for PPI prediction, as different methods can be applied to the same data sets with the same features generated from KUPS. In addition, users can generate a highly imbalanced data set with significantly more negative examples than positive examples, which mirrors the observed sparsity of true protein interactions. Furthermore, users can generate negative examples with KUPS using four different strategies to avoid the bias problems in GRIP.

ACKNOWLEDGEMENTS

We thank Matthaw Mandelbaum who has provided valuable discussions.

FUNDING

National Science Foundation (IIS-0644366).

Conflict of interest statement. None declared.

REFERENCES

1. Smialowski,P., Pagel,P., Wong,P., Brauner,B., Dunger,I., Fobo,G., Frishman,G., Montrone,C., Rattei,T., Frishman,D. *et al.* (2009) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res.*, **38(Database issue)**, D540–D544.
2. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N., Ching,S., Emili,A., Snyder,M., Greenblatt,J. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
3. Jansen,R. and Gerstein,M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.*, **7**, 535–545.
4. Bader,G. and Hogue,C. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.
5. Grigoriev,A. (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.*, **31**, 4157–4161.
6. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
7. Browne,F., Wang,H., Zheng,H. and Azuaje,F. (2009) GRIP: A web-based system for constructing gold standard datasets for protein-protein interaction prediction. *Source Code Biol. Med.*, **4**, 2.
8. Mewes,H., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28–30.
9. Ben-Hur,A. and Noble,W. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformaticss*, **7(Suppl. 1)**, S1.
10. Ceol,A., Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38(Database issue)**, D532–D539.
11. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P. and Valencia,A. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32(Database issue)**, D452–D455.
12. Keshava Prasad,T., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicheria,D., Raju,R., Shafreen,B. and Venugopal,A. (2009) Human protein reference database0—2009 update. *Nucleic Acids Res.*, **37(Database issue)**, D767–D772.
13. The Gene Ontology Consortium. (2010) The Gene ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38(Suppl. 1)**, D331–D335.
14. The UniProt Consortium. (2010) The universal protein resource (UniProt)in 2010. *Nucleic Acids Res.*, **38(Suppl. 1)**, D142–D148.
15. Kawashima,S., Ogata,H. and Kanehisa,M. (1999) AAindex: amino acid index database. *Nucleic Acids Res.*, **27**, 368–369.
16. Gribskov,M., McLachlan,A. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
17. Garian,R. (2001) Prediction of quaternary structure from primary structure. *Bioinformatics*, **17**, 551–556.
18. Garg,A. and Raghava,G. (2008) A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.*, **8**, 129–140.
19. Bhasin,M. and Raghava,G. (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.*, **13**, 596–607.
20. Chen,X. and Jeong,J. (2009) Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, **25**, 585–591.
21. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
22. Lord,P., Stevens,R., Brass,A. and Goble,A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
23. Lord,P., Stevens,R., Brass,A. and Goble,A. (2003) Semantic similarity measures as tools for exploring the gene ontology. *Pac. Symp. Biocomput.*, **8**, 601–612.
24. Finn,R., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J., Gavin,L., Gunasekaran,P., Ceric,G. and Forslund,K. (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38(Suppl. 1)**, D211–D222.
25. Bock,J. and Gough,D. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.
26. Hopp,T. and Woods,K. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
27. Guo,Y., Yu,L., Wen,Z. and Li,M. (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.
28. Jones,S. and Thornton,J. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
29. Jones,S. and Thornton,J. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.
30. Jones,S. and Thornton,J. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.
31. Shen,J., Zhang,J., Luo,X., Zhu,W., Yu,K., Chen,K., Li,Y. and Jiang,H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.
32. Jones,D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
33. Jeong,J., Lin,X. and Chen,X. (2010) On position-specific scoring matrix for protein function prediciton. *IEEE/ACM Trans. Comput. Biol. Bioinform*, http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.93.