HINT-KB: The Human Interactome Knowledge Base

Konstantinos Theofilatos[1], Christos Dimitrakopoulos[1], Dimitrios Kleftogiannis[2], Charalampos Moschopoulos[3], Stergios Papadimitriou[4], Spiros Likothanassis[1], Seferina Mavroudi[1,5]

[1] Department of Computer Engineering and Informatics, University of Patras, Greece
[2] Math. & Computer Sciences & Engineering King Abdullah Univ. of Science and Technology, Saudi Arabia
[3]Department of Electrical Engineering-ESAT, SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 300, Leuven, Belgium and IBBT Future Health Department, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 300, Leuven, Belgium
[4]Department of Information Management, Technological Institute of Kavala, Greece
[5]Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Patras, Greece

{theofilk, dimitrakop, kleftogi, likothan, mavroudi}@ceid.upatras.gr,
cmoschop@esat.kuleuven.be, sterg@teikav.edu.gr

Abstract. Proteins and their interactions are considered to play a significant role in many cellular processes. The identification of Protein-Protein interactions (PPIs) in human is an open research area. Many Databases, which contain information about experimentally and computationally detected human PPIs as well as their corresponding annotation data, have been developed. However, these databases contain many false positive interactions, are partial and only a few of them incorporate data from various sources. To overcome these limitations, we have developed HINT-KB (http://150.140.142.24:84/Default.aspx) which is a knowledge base that integrates data from various sources, provides a user-friendly interface for their retrieval, estimates a set of features of interest and computes a confidence score for every candidate protein interaction using a modern computational hybrid methodology.

Keywords: Protein-Protein Interactions, Human, PPI scoring methods, Genetic Algorithms, Kalman Filters, Knowledge Base

1 Introduction

Among the numerous participants in molecular interactions, proteins are probably the most important ones. In particular, proteins transmit regulatory signals throughout the cell, catalyze a huge number of chemical reactions, and are critical for the stability of numerous cellular structures. The total number of possible interactions within the cell is astronomically large and the full identification of all true PPIs is a very challenging task.

Many high throughput methodologies have been developed for the experimental prediction of PPIs, with the prevailing ones among them been the yeast two-hybrid

(Y2H) system, mass spectrometry (MS), protein microarrays, and Tandem Affinity purification (TAP) [1] . These methodologies have raised the interactome's coverage of known human PPIs, but they include many false positive PPIs and until now they do not include all the existing ones.

In order to expand the existing knowledge about PPIs and to score the existing ones so as to exclude false positives, many computational approaches have been developed. All computational methods use protein and protein-interaction data which are located in public databases and most of them are supervised machine learning classifiers. These classifiers use as inputs genome-scale, sequence based, structure based, network based and functional based features. The main machine learning methods that have already been applied in the computational prediction of PPIs are Bayesian classifiers, Artificial Neural Networks, Support Vector Machines and Random Forests [1]. All these approaches fail to achieve both high classification performance and interpretability. Moreover, all existing computational approaches face the class imbalance problem, as well as the problem of missing values and datasets incompleteness. For these reasons many hybrid modern methodologies have been proposed recently including methodologies proposed by our research group [2,3].

The computational approaches for the prediction and scoring of PPIs are not standalone applications. They need high quality datasets to train their models in order to produce better predictors. An enormous variety of databases containing information about proteins and protein interactions have been developed and they will be described briefly in the present manuscript. Up to our knowledge, there does not exist any database to integrate the available experimental, functional, structural and sequential information about PPIs. The absence of such a database was the initial motivation for the present work, which will incorporate and automatically curate a large amount of the available information about the PPIs. HINT-KB is not just a simple integrative database, but combines primitive information to produce new knowledge. Specifically, an integrated scoring method, produces confidence scores for every possible PPI and allows for creation of reliable negative PPI examples which could be used for the training of supervised classification techniques. Furthermore, HINT-KB offers some simple data preprocessing tools concerning normalization and missing values estimation. The content of HINT-KB may be accessed through a user friendly interface which provides forms for the retrieval of positive PPIs, negative PPIs, complete datasets and interactions of simple proteins.

2 Existing Human Interaction Databases

The most important existing public available databases containing information about human protein-protein interactions are the following:

- **MIPS**: The Munich Information Center for Protein Sequences, is a resource of high quality experimental protein interaction data in mammals, including Homo Sapiens. It is designed in order to favor quality over completeness including mainly published experimental evidence from individual experiments instead of large scale surveys. For every protein interaction it provides sequence, structure and some functional annotations.

- **MINT**: The Molecular Interaction database, is a relational database designed to store data on functional and direct interactions between proteins. It focuses on experimentally verified protein-protein interactions providing, whenever available, information about kinetic and binding constants as well as the domains participating in the interaction. MINT provides also information about complexes located in the PPI networks.
- **IntAct**: The molecular InterAction database, is an open-source, open data molecular interaction database emphasizing on protein-protein interactions of 275 species. It follows a deep curation model, capturing a high level of detail from the experimental reports on the full text of their publications. Binary PPIs can be found using IntAct, and additionally annotations about their binding domains, tags or stoichiometry may be derived in a user friendly way.
- **DIP**: The Database of Interacting Proteins, is a database that documents experimentally determined protein-protein interactions. It provides information about the proteins that take part in the interactions and for every interaction it gives information about the protein domains involved, when known. Beyond cataloging details of protein-protein interactions the DIP is useful for understanding protein function and protein-protein relationships, studying the properties of networks of interacting proteins, benchmarking predictions of protein-protein interactions and studying the evolution of protein-protein interactions.
- **BIND**: The Biomolecular Interaction Network Database is a Database designed to store full descriptions of interactions, molecular complexes and pathways. About protein-protein interactions it stores a text description, cellular place of the interaction, experimental conditions used to observe binding, a comment on evolutionary conserved biological sequence, binding sites of proteins and how they are connected, chemical action including kinetic and thermodynamic data and chemical state of the proteins involved. Furthermore, BIND includes detailed information about post-translational modifications from mass spectrometry experiments.
- **HPRD**: Human Protein Reference Database is a database of curated proteomic information pertaining to human proteins. It provides information about 38194 protein-protein interactions, 16972 post translational modifications and annotations of one or more sites of subcellular localization for 8620 proteins. Recently, they have developed PhosphoMotif Finder, which allows users to find the presence of over 320 experimentally verified phosphorylation motifs in proteins of interest.
- **BioGRID**: Biological General Repository for Protein interaction Datasets is a database that maintains a collection of protein and genetic interactions from major model organism species. Specifically, it contains over 198.000 interactions from six different species which are derived from both high-throughput studies and conventional focused studies.
- **STRING:** is an online database resource search tool for the retrieval of interacting genes and proteins. It provides uniquely comprehensive coverage and access to both experimental and predicted interaction information. STRING covers more than 1100 organisms varying from bacteria to humans. It also provides a confidence score giving guidance to users who want to balance different levels of coverage and accuracy.

- **PIPS** [4] is an online database source which includes computationally predicted protein-protein interactions which have been predicted using a Bayesian Approach and combining various features, such as homology, domain-domain interactions, gene expression profiles similarity, post translational modifications co-occurrence, protein disorder, localization and a network topology based feature.

The continuing growth in scientific community needs for convenient extraction of PPI data led to some efforts for creating databases that integrate information from other existing databases. A striking example in this direction is iRefIndex [5] which incorporates data from the most famous databases like BIND, BioGrid, DIP, HPRD etc. For every PPI recorded in iRefIndex, this database provides the number of references about it, the identification method used to detect it, the databases referencing it and some other annotations about the PPIs. The main disadvantage of iRefindex, which HINT-DB tries to surpass, is the fact that it does not incorporate much functional or structural information about the PPIs. Another integrative database is IPA (http://www.ingenuity.- com/products/pathways_analysis.html).

Except for the basic information about the PPIs, which are derived from databases like the ones described previously, researchers are interested in supplementary functional and structural information. Next, we present the basic databases which provide important functional or structural information about proteins. The following databases are used by researchers in order to compute the various features – inputs needed by computational PPI predictors.

- **Inparanoid**: An eukaryotic ortholog database containing pairwise ortholog groups between 17 whole genomes. Information about ortholog proteins is very important because if the orthologs of two proteins interact in one eukaryotic organism it is possible that they will interact in another eukaryotic organism as well.
- **InterPro**: The integrative protein signature database integrates together predictive models or 'signatures' representing protein domains, families and functional sites from multiple, diverse source databases. In the problem of PPIs' identification with computational methods, InterPro is mainly used to find the co-occurrence of domains in the protein pairs.
- **PDB**: The Protein Data Bank is the single worldwide archive of structural data of biological macromolecules. It contains information about every known 3D protein structure. Except for the coordinates of the 3D models, it maintains annotations about the method that was used for the structures determination.
- **Pfam** is a database of protein families that currently contains 7973 entries. A recent development of Pfam has enabled the grouping of related families into clans. If two protein pairs belong to the same clan, then this is an indication that they probably interact.
- **GO**: The Gene Ontology database provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. In the problem of PPIs' identification, GO database is used for categorizing proteins in functional categories described in it and for extracting information

about the participation of a query protein in the metabolic or signaling pathways.

- **SCOP**: The Structural Classification Of Proteins is a database which comprehensively orders all proteins of known structure due to their evolutionary and structural relationships. Protein domains in SCOP are hierarchically classified into families, superfamilies, fold and classes.
- **CATH** is a database containing information about the classification of up to 86151 structural protein domains. The classification is achieved through an automatic procedure. CATH is also available for classifying remote folds and homologues.

To the best of our knowledge, there does not exist any database incorporating all the functional, structural and sequential information available in public databases about the protein-protein interactions.

3 HINT-KB: The Human INTeractome Knowledge Base

In order to provide the highest PPI coverage for the Human Interactome, HINT-KB is based on PPIs which are included in the integrative database IrefIndex [5]. Human protein interactions were downloaded from iRefindex (264923 interactions) and from this dataset 20845 unique proteins were identified filtering out proteins for which information about the amino-acid sequence is not available. A set of 500.000 negative-candidate PPIs is produced by randomly combining and filtering out protein interactions included in iRefindex's positive PPIs dataset.

3.1 Calculated Protein-Protein Interaction Features and Related Sources

The computational prediction of PPIs is based on a set of features which contain information about the proteins consisting the PPI and about their combination. The most important features, used as inputs by computational models for the prediction of PPIs, are stored in HINT-KB. These features are briefly described in the subsequent list:

- **Gene Ontology (Co-function, Co-process, Co-localization)**.(3 features) Gene Ontology database contains information of three types that includes:
  - Molecular function of a gene product,
  - Biological process in which the gene product participates,
  - Cellular component where the gene product acts

  For each protein pair, the number of common GO-terms shared by the two proteins is computed for each one of the GO types.
- **Sequence Similarity**. (1 feature) This feature was obtained by using the NCBI-BLAST+ standalone's executable BLASTP to perform the human to human protein alignment. All BLASTP hits obtained with the default parameters and

the sequence alignment E-values of each protein-protein pair have been used to fill in the specific feature.

- **Homology Based PPI**. (1 feature) For each human protein pair we have found its homology pairs in yeast by using the 0.001 cutoff on the E-value taken when similarity tests are conducted using BLASTP. If at least one of the resulting homology pairs is referred as interacting in the Yeast organism's DIP dataset, a value equal to 1 is assigned to this feature, otherwise "0".

- **Gene Expression Profile Similarity**. (15 features) Fifteen gene expression datasets were used to estimate 15 gene expression profile similarity features. The datasets selected are fifteen datasets from NCBI Gene Expression Omnibus (GDS531, GDS534, GDS596, GDS651, GDS806, GDS807, GDS843, GDS987, GDS1085, GDS2855, GDS1402, GDS181, GDS1088, GDS841, GDS3257) and they consist the most large-scaled ones about human genome. For each dataset, the expression profile of each protein of every protein-pair were extracted and the Pearson correlation between the two profiles was estimated. This correlation estimation is the actual value for this feature.

- **Co-localization.** (1 feature) PLST tool was used to predict the local cellular compartments in which each protein is probable to function. For every pair of proteins, a feature is calculated by taking the value of 1 if the two proteins share at least one local compartment and the value of 0 if they do not have any common local cellular compartment.

- **Domain-Domain Interactions.** (1 feature) The known domains for human proteins were downloaded from the InterPro Database. All possible domain pairs were evaluated using the training set and the hypergeometric distribution to locate pairs that are interacting with high probability (p-values less than 0.05). Then for every protein pair, their domain combinations are computed and the number of the interacting ones is measured and used as the feature's value.

Our algorithm does not assume independence of the features. Hence, we incorporate several features in our dataset, which might share similar but not identical information. Nevertheless, our algorithm is able to detect close to optimal combinations between the features by searching for the best combination of mathematical terms.

For the training and testing process of our classifier we used 1000 positive interactions referred in HPRD and 1000 negative protein interaction samples to train the classifier. We assume HPRD database to be highly reliable as it contains protein interactions that are supported by low and high throughput experimental evidence. The negative samples were created randomly and uniformly from the unique identities of the whole set of proteins. The features are normalized in the range [0,1] before they are submitted to classification. Moreover, missing values are filled in using the kNN-impute methodology. If a specific feature is missing for a specific sample, then its k nearest neighbors are been searched and their average of the missing feature for the subjected sample replaces the missing value.

3.2 PPI Scoring methodology

The main idea of the classification methodology that has been performed is to find a simple mathematical equation that governs the best classifier enabling the extraction of biological knowledge. The algorithm [2,3] combines a state-of-the-art adaptive filtering technique named Kalman Filtering with an adaptive genetic algorithm, one of the most contemporary heuristic methods which are based in natural selection process. The classification method uses a genetic algorithm to find the best subset of terms in order to build the mathematical model for our predictor and then apply Extended Kalman Filters to find its optimal parameters. The optimal mathematical model obtained from this methodology is used to score all available protein-protein interactions in the HINT-KB database.

The problem of finding the optimal subset of terms to include in our mathematical model is computationally very costly. Having 341 candidate terms the search space of the problem is 2^{341}. This is why we needed Genetic Algorithms capable of performing well in large search spaces with many local optima. Genetic Algorithms can deal with large search spaces and do not get trapped in local optimal solutions like other search algorithms. The 341 mathematical terms that were used in our method were taken from 3 known nonlinear mathematical models that are described below:

- Volterra Series model which contains 231 mathematical terms.
- Exponential model which contains 44 mathematical terms.
- Polynomial model which contains 88 mathematical terms.

Some mathematical terms exist in more than one mathematical models. Counting all individual terms once, gives us the number of 341 possible mathematical terms (22 linear terms exist in this set). In our approach, we used a simple Genetic Algorithm, where each chromosome comprises of genes that encode the best subset of mathematical terms to be used in our classifier. For the selection step we used roulette selection including elitism to accelerate the evolution of the population.

The evaluation process that was used in our method is described in the following steps: For every individual of the population, the Extended Kalman Filter Method is deployed (using the training dataset) to compute the best parameters for the subset of terms that arises from the individual’s genes. Then the mathematical model, found at the previous step, is used to compute the fitness of the classifier in the validation set. For every individual the output of the following function is computed:

$$\frac{P\_pos(m_{pos}-m_0)+P\_neg(m_{neg}-m_0)}{P\_pos*\sigma_{pos}{}^2+P\_neg*\sigma_{neg}{}^2}+d*\#\,features \qquad (1)$$

The individuals are ranked from the one having the lower fitness value to the one having the bigger fitness value. Every individual is assigned with a fitness value equal to their ranking number. For example, the worst individual will take fitness equal to 1, the next fitness equal to 2 etc.

During the training we used 10-fold external and 9-fold internal cross validation. Specifically, the external folds vary the subset of data used for testing the mathematical models (10 non-overlapping different subsets) after the execution of the genetic algorithm, whereas the internal folds vary the subset of data used for validating the mathematical models (9 non-overlapping different subsets) during the execution of the genetic algorithm. We keep the optimal model of each implementation of the 9 internal folds. Then, the average of the measure for the 10 optimal models is been computed (Table 1.1).

One problem found during experimentations is that there are many subsets of terms that give classifiers which have approximately the same F value. Using the evaluation procedure described above, better scaled fitness scores were achieved. After experimentation in the training set, we concluded in crossover probability equal to 0.9. Furthermore, a self-adaptation mechanism of a single mutation rate per individual was deployed. The basic idea is that better parameter values lead to better individuals and these parameter values will survive in the population since they belong to the surviving individuals [6]. The mutation of the mutation rate value gives the new mutation rate through the equation:

$$p' = (1 + \frac{1-p}{p} * \exp(-\gamma * N(0,1)))^{-1} \qquad (2)$$

The size of the initial population was set equal to 30 chromosomes and the termination criterion is the maximum number of 100 generations to be reached combined with a termination method that stops the evolution when the population is deemed as converged. The population is deemed as converged when the average fitness across the current population is less than 5% away from the best fitness of the current population.

The optimal mathematical model obtained by the implementation of the above algorithm contains 162 of the possible 364 mathematical terms and is highly predictive for both positive and negative PPI samples (the table 1.1 contains the corresponding measures). This mathematical model has been used as a scoring assignment mechanism to the protein interactions. The measures achieved in average by the 10 optimal mathematical models are: 0.87 (± 0.051) Accuracy, 0.85 (± 0.043) Sensitivity and 0.89 (± 0.056) Specificity.

3.3 HINT-KB Database

HINT-KB supports a relational database, in order to store the produced information and allow user to ask queries and download data with efficiency and speed. The database is organized in four tables: proteins, protein pairs, positive interaction and negative interactions. For each protein, its identifier and entry name in Uniprot [7], its amino-acid sequence and the Entrez gene id which is an identifier showing the gene that is transcribed to this protein are stored. For every protein pair the following information is stored: the calculated 22 features normalized and in their initial form, if there exists evidence about this interaction in HPRD database, a binary string of length 22 showing if a feature is missing or not and the score predicted for this pair by our computational approach. The tables of positive and negative interactions include the identifiers of the protein pairs which have been predicted as positive or negative respectively.

3.4 User Interface

The HINT-KB uses a three layered user accessibility approach. Administrator users, are able to call through an interface a combination of python and matlab scripts to update the dataset, provided that a new version of a source is available and downloaded. The new sources versions should be downloaded manually and uploaded through the interface in HINT-KB.

Simple users may access the knowledge base using four available modules. The first and the second ones are the modules for acquiring positive and negative PPIs with score over and under a user defined threshold. The third module enables users to download a complete dataset providing information about the number of positive interactions they need, the positive to negative ratio for their dataset and the values for the 22 features, normalized or not, with the missing values being computed or not. Finally, the forth module enables users to insert a protein Uniprot identifier of interest plus a score threshold and returns as output the interactions which are associated with this protein and their score exceeds the threshold value.

Registered users may access the same modules as simple users. Moreover, they have the privilege to download the results of their queries as tab delimited text files. The registration process is simple and fast. Simple users may upgrade themselves to registered users, by filling up a registration form.

4 Conclusions

HINT-KB is an integrative knowledge base for Human PPIs. We use the PPIs reported in IrefIndex database as a reference, as this database contains positive PPIs which are included in a high number of existing human PPI databases. Furthermore, for each protein pair it provides an adequate set of annotation features which could be used as inputs for computational methods for the prediction of PPIs. HINT-KB also includes a modern accurate and interpretable classification and scoring methodology, which assigns to each pair of proteins a confidence score. This confidence score is used to filter out existing false positives and to predict novel candidate PPIs which should be checked experimentally. A user friendly web interface enables researchers to download PPI data stored in HINT-KB and could be used to produce reference datasets for the comparison of PPI computational prediction methods. Moreover, users are able to specify their own classification threshold for the retrieval of positive and negative datasets, tuning this way the specificity and the sensitivity of the classification algorithm. Except for the issue of classifying protein interactions and scoring their confidence, HINT-KB database supports the downloading of the most important features that characterize protein interactions (functional, structural and sequential features). Furthermore, the mathematical model which was extracted for the computational prediction of PPIs could be further analyzed to uncover the complex mechanisms which control whether a pair of proteins interacts in a human cell or not.

Our future agenda for the HINT-KB improvements contains the incorporation of computational methodologies for the visualization of protein interaction networks [8], the computational prediction of protein complexes [9] and the computational prediction of interactions between biological processes in human [10]. Hence, HINT-KB is intended to incorporate different kind of interacting information of the human cellular level, addressing in this way the issue of storing all types of Human Interactome data in a single database.

Acknowledgments. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. Theofilatos, K.A., Dimitrakopoulos, C.M., Tsakalidis, A.K., Likothanassis, S.D., Papadimitriou, S.T., Mavroudi, S.P. Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey, Current Bioinformatics, Vol. 6, Number 4, pp. 398-414 (2011)
2. Theofilatos, K.A.; Dimitrakopoulos, C.M.; Tsakalidis, A.K.; Likothanassis, S.D.; Papadimitriou, S.T.; Mavroudi, S.P. A new hybrid method for predicting protein interactions using Genetic Algorithms and Extended Kalman Filters, In: Proceedings of the IEEE/EMBS Region 8 International Conference on Information Technology Applications in Biomedicine (ITAB) art. no. 5687765, doi : 10.1109/ITAB.2010.5687765 (2010)
3. Dimitrakopoulos, C.M., Theofilatos, K.A., Georgopoulos, E.F., Likothanassis, S.D., Tsakalidis, A.K., Mavroudi, S.P., Efficient Computational Construction of Weighted Protein-Protein Interaction Networks Using Adaptive Filtering Techniques Combined with Natural-Selection Based Heuristic Algorithms, International Journal of Systems Biology and Biomedical Technologies (IJSBBT), vol.1, Issue 2, pp. 20-34 (2011)
4. Scott M. and Barton G., Probabilistic prediction and ranking of human protein-protein interactions, BMC Bioinformatics, vol. 8:239 (2007).
5. Razick S., Magklaras G., Donaldson I.M.: iRefIndex: A consolidated protein interaction database with provenance, BMC Bioinformatics, vol. 9(Issue 1):405 (2008).
6. Breukelaar R., Baeck T., Self-adaptive mutation rates in genetic algorithm for inverse design of cellular automata, In: Proceedings of the 10th annual conference on Genetic and evolutionary computation, July 12-16, Atlanta, GA, USA [doi>10.1145/1389095.1389298] (2008).
7. The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. vol. 40: D71-D75 (2012).
8. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. vol. 13, pp.2498–2504 (2003)
9. Moschopoulos C.N., Pavlopoulos G.A., Schneider R., Likothanassis S.D., Kossida S.: GIBA: a clustering tool for detecting protein complexes. BMC Bioinformatics, 10(Suppl 6):S11 (2009).
10. Dotan-Cohen D., Letovsky S., Melkman A.A., Kasif S. Biological Process Linkage Networks. PLoS ONE 4(4): e5313. doi:10.1371/journal.pone.0005313 (2009)