HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2015 December 05.

Published in final edited form as:

J Proteome Res. 2014 December 5; 13(12): 5944–5955. doi:10.1021/pr5008416.

An Improved Algorithm and Web Application for Predicting Co-Complexed Proteins from Affinity Purification – Mass Spectrometry Data

Dennis Goldfarb[1,2], **Bridgid Hast**[2], **Wei Wang**[1], and **Michael B. Major**[2]

[1]Department of Computer Science, University of North Carolina at Chapel Hill, Box#3175, Brooks Computer Science Building, Chapel Hill, NC 27599, USA

[2]Department of Cell Biology and Physiology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill School of Medicine, Box#7295, Chapel Hill, NC 27599, USA

Abstract

Protein-protein interactions defined by affinity purification and mass spectrometry (APMS) approaches suffer from high false discovery rates. Consequently, the candidate interaction lists must be pruned of contaminants before network construction and interpretation, historically an expensive and time-intensive task. In recent years, numerous computational methods have been developed to identify genuine interactions from hundreds revealed by APMS experiments. Here, comparative analysis of several popular algorithms revealed complementarity in their classification accuracies, which is supported by their divergent scoring strategies. As such, we used two accurate and computationally efficient methods as features for machine learning using the Random Forest algorithm. Additionally, we developed novel mathematical models to include a variety of indirect data, such as mRNA co-expression, gene ontologies and homologous protein interactions as features within the classification problem. We show that our method, which we call Spotlite, outperforms existing methods on four diverse and public APMS datasets. Because implementation of existing APMS scoring methods requires computational expertise beyond many laboratories, we created a user-friendly and fast web application for APMS data scoring, analysis, annotation and network visualization, for use on new and existing data (http://152.19.87.94:8080/spotlite). The utility of Spotlite and its visualization platform for revealing physical, functional and disease-relevant characteristics within APMS data is established through a focused analysis of the KEAP1 E3 ubiquitin ligase.

INTRODUCTION

Mapping the global protein-protein interaction network and defining its dynamic reorganization during specific cell state changes will provide an invaluable and transformative knowledgebase for many scientific disciplines. Recent advancements in two-hybrid technologies and affinity purification – mass spectrometry (APMS) have

Author for correspondence: Michael B. Major, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill Box#7295, Chapel Hill, NC 27599, benmajor@med.unc.edu, Phone: 919-966-9258, Fax: 919-966-8212.

dramatically increased protein connectivity information, and therefore a proteome-wide interaction map may be realized in the not-so-distant future. Specifically, technological and computational advancements in mass spectrometry-based proteomics have increased sample throughput, detection sensitivity and mass accuracy, all with decreasing instrumentation costs. Consequently, to date over 2,200 human proteins have been analyzed by APMS, as estimated through BioGRID and data presented herein (1). Similarly, the generation of arrayed human clone sets has revealed binary interactions among approximately 13,000 proteins (HI-2012 Human Interactome, Center for Cancer Systems Biology). While both approaches detect direct protein interactions, only APMS can detect indirect interactions – though with limited ability to distinguish between the two types.

In general, APMS-based protein interaction experiments are performed by selectively purifying a specific protein, termed the bait, along with its associated proteins from a cell or tissue lysate. Mass spectrometry is then used to identify and more recently quantify the bait and all associated proteins within the affinity purified protein complex, collectively termed the prey. Though a prey's presence supports its existence within a complex, high numbers of non-specific contaminants—owing largely to technical artifacts during the biochemical purification—lead to false protein complex identifications and therefore significantly hamper data interpretation. As such, numerous computational methods have been developed to differentiate between genuine APMS protein complex interactions and false-positive discoveries.

These algorithms can be broadly categorized based on which features of the APMS data are included and how the resulting network is mapped. Methods such as SAI, Hart, Purification Enrichment scores and Dice Coefficients use the binary presence of the protein as evidence for an interaction (2–8). More recently, computational approaches employed by SAINT (9), MiST (10), CompPASS (11) and the HGSCore (12) achieved improved scoring accuracy by taking advantage of label free quantification using spectral counts, a reflection of the abundance of a protein after purification. Additionally, these algorithms can also be categorized by whether they use a spoke or matrix model to represent protein connectivity (4). The spoke model represents only bait-prey interactions, while the matrix model – used by the Hart (7) and HGSCore methods – additionally represents all prey-prey interactions, resulting in a quadratic number of potential interactions per experiment instead of linear, and therefore contain an order of magnitude more interactions to test. Though the matrix model can detect more true complex co-memberships, it has the added difficulty of filtering prey pairs that form distinct complexes with the bait. Each method has its merits and has been successfully applied to APMS data; however, their widespread utilization has been limited.

In addition to using direct features from APMS experiments to predict the validity of putative protein-protein interactions, success in the *de novo* prediction of protein interactions has been achieved through the analysis of indirect data (13–16). Specifically, mRNA co-expression has been shown to positively correlate with co-complexed proteins, and the Gene Ontology's (GO) biological process and cellular component annotations have proven to be useful for interaction prediction by utilizing semantic similarity (17–19). Both co-expression and GO co-annotation are also commonly used metrics for evaluating predicted interactions. Sequence and structural homology at the domain and whole-protein level have established

themselves as powerful predictors as well (20, 21). Though individually useful, integration of these indirect sources using machine learning techniques such as Support Vector Machines (22), Random Forests (23), Naïve Bayes (24), and Logistic Regression (25) have further increased prediction accuracy. APMS data has also been used as a discriminative feature, however it was encoded as a binary value representing an interaction's presence – far less powerful than the sophisticated APMS scoring methods now available (16).

Among the label free methods, only SAINT's software is available for public use and requires compilation and command line execution – limiting its use for research groups lacking computational expertise (9). CompPASS provides a public web interface to search its data, but no option to employ the algorithm on private datasets (11). Aside from APMS scoring methods, numerous web applications are available for *de novo* protein-protein interaction prediction (26, 27). These methods do not incorporate new APMS data, and therefore provide an insufficient resource for researchers wishing to integrate their own experiments into the predictions.

Given the independent successes of using direct and indirect data to predict and score protein-protein interactions, we created Spotlite, a Random Forest-based (28) classifier to identify genuine interactions from human APMS experiments, one which utilizes features taken from APMS data and a variety of indirect data. To foster its use within the proteomic community, we developed and speed-optimized a web application for Spotlite execution on existing and novel datasets. In addition to providing an integrated network visualization tool, because all features employed within the machine learning algorithm are provided, the resulting protein interactions are annotated for function, model organism phenotype and human disease relevance.

EXPERIMENTAL PROCEDURES

Data Collection

To develop a classification strategy capable of efficiently segregating false positive protein interactions from true interactions within APMS-derived data, we collected four publically available and well-diversified APMS datasets. These data were received directly from the authors of the respective publications and searched and filtered using the criteria described in their methods. The data contained spectral counts, baits, and preys for each experiment. For the purposes of establishing a classifier, we defined protein-protein interactions taken from the Human Interactome and BioGRID as known interactions. More specifically, protein-protein interactions were downloaded from BioGRID (1)(http://thebiogrid.org/ Release 3.1.89) and appended with the Human Interactome project's two-hybrid data from The Center for Cancer Systems Biology at the Dana-Farber Cancer Institute (http://interactome.dfci.harvard.edu/). Protein sequences and cross database accession mappings were downloaded from IPI (29)(http://www.ebi.ac.uk/IPI/ final releases) and UniProt/SwissProt (30)(http://www.uniprot.org/ Release 05/2012). Protein domains were determined with PfamScan (31)(http://pfam.sanger.ac.uk/ Release 26.0) using an e-value threshold of 0.05. Entrez Gene IDs, official symbols, aliases, and gene types were extracted from NCBI Gene's FTP site, http://www.ncbi.nlm.nih.gov/gene (gene_history.gz and gene_info.gz - downloaded 05/26/12). Gene homolog data was downloaded from NCBI's Homologene

(http://www.ncbi.nlm.nih.gov/homologeneBuild66). Pearson correlation coefficients for co-expression data were downloaded from COXPRESdb (32) (http://coxpresdb.jp/) for *Homo sapiens* (version c3.1), *Mus musculus* (version c2.1), *Drosophila melanogaster* (version c1.0), *Caenorhabditis elegans* (version c1.0), *Danio rerio* (version c1.0), *Gallus gallus* (version c1.0), and *Rattus norvegicus* (version c2.0). Ontology hierarchies and annotations were downloaded on 05/26/12. The Gene Ontology supplied the biological process and cellular component ontology hierarchies, where the annotations were downloaded from NCBI Gene's FTP site (33). The Mammalian Phenotype Ontology (relevant organism: *Mus Musculus*) hierarchy and annotations were downloaded from Mouse Genome Informatics (34)(http://www.informatics.jax.org/). The Human Phenotype Ontology's hierarchy and annotations were downloaded from www.human-phenotype-ontology.org (35). The Disease Ontology annotations were taken from its associated publication's supplemental data (http://projects.bioinformatics.northwestern.edu/do_rif/) and the hierarchy from the OBO Foundry (36)(http://obofoundry.org/).

ID Mapping

IPI and UniProt accessions were directly mapped to Entrez Gene IDs using cross database accession mappings. If a direct mapping did not exist, indirect mappings were checked by finding accession to accession mappings (e.g. IPI to UniProt) and testing them for direct mappings. If multiple Gene IDs were identified as candidates, an arbitrary one was chosen. Similarly, Gene Symbols were directly mapped to Gene IDs. For Gene Symbols that did not match an Official Gene Symbol, gene aliases were queried and if multiple candidate IDs were found, one was chosen arbitrarily.

Machine Learning

We approached the probabilistic scoring of APMS protein-protein interactions as a binary classification problem in which the two classes are: 1) pairs of proteins that directly or indirectly form a complex together (positive class), and 2) pairs of proteins that are never members of the same complex (negative class). The classifier chosen was a Random Forest (28)(Weka v3.6.7 implementation, (37)), which was previously shown to demonstrate high classification accuracy for protein-protein interaction prediction (16). The features used to characterize a pair of putative co-complexed proteins were gene co-expression patterns in humans and their homologs in six different species, as well as semantic similarity scores for the pair's ontological annotations for biological processes, cellular components, diseases, mutant phenotypes, and mouse homologs' mutant phenotypes. Additionally, domain-domain binding affinities, homologous interactions, and label-free APMS scoring methods CompPASS and the HGSCore were used. Only spoke model interactions were tested, because CompPASS was not designed for the matrix model, as well as for computational efficiency. The classifier was trained specifically for human data, using a training set comprised of four published APMS datasets (Table 1). Ultimately, the Random Forest classifier outputs the probability a candidate APMS protein-protein interaction belongs to the positive or negative class of co-complexed proteins.

Feature Calculation for Random Forest

For classification, all putative APMS-derived protein-protein interactions were characterized by direct and indirect features. Direct features, which are derived from the output of the APMS experiment, included the HGSCore score and a modified CompPASS WD-score. The HGSCore is capable of testing matrix model interactions, however, for implementation within Spotlite, we restricted it to spoke model interactions. The HGSCore first converts spectral counts to normalized spectral abundance factors (NSAF (38)), then scales them by setting the smallest value to 1, and finally taking the integer value of the square root – resulting in a T_n value for each protein in an experiment. The HGSCore is then defined as:

$$P\left(\sum \min(T_N) > k | n, m, N\right) = \sum_{x=k}^{\min(n,m)} P_{hygeo}(x|n, m, N)$$

$$P_{hygeo}(x|n, m, N) = \frac{\binom{n}{x}\binom{N-n}{m-x}}{\binom{N}{m}}$$

$$k = \sum \min(T_N) \text{ for experiments with } T_{N;i} > 0 \text{ and } T_{N;j} > 0$$

$$n = \sum \min(T_N) \text{ for experiments with } T_{N;i} > 0$$

$$m = \sum \min(T_N) \text{ for experiments with } T_{N;j} > 0$$

$$N = \sum \min(T_N) \text{ for all experiments}$$

$$HGSCore_{ij} = -log(P_{hygeo}; i, j)$$

We computed a modified CompPASS score to account for increased variability within biological replicates. The original CompPASS WD-score equation was designed such that each bait protein would be analyzed by APMS two independent times. The equation was defined as:

$$WD_{i,j} = \sqrt{\left(\frac{k}{\sum_{i=1}^{i=k} f_{i,j}}\omega_j\right)^p X_{i,j}}$$

$$\omega_j = \left(\frac{\sigma_j}{\overline{X}_j}\right) \quad \overline{X}_j = \frac{\sum_{i=1,j=n}^{i=k} X_{ij}}{k}; n = 1, 2, .., m \quad \begin{array}{l} if\ \omega_j \leq 1 \rightarrow \omega_j = 1 \\ if\ \omega_j > 1 \rightarrow \omega_j = \omega_j \end{array}$$

$$f_{ij} = \begin{cases} 1; X_{ij} > 0 \\ X_{ij} \end{cases} \quad p = \begin{array}{c} \text{number of replicate runs in} \\ \text{which the interactor is present} \end{array} \quad X_{ij} = \begin{array}{c} \text{total spectral count} \\ \text{for prey j with bait i} \end{array}$$

To permit greater variability within the number of duplicate biological replicate experiments, as is required when analyzing disparate datasets originating from many independent laboratories, we modified the exponent to have a smaller effect on reproducibility:

$$WD_{i,j} = \sqrt{\left(\frac{k}{\sum_{i=1}^{i=k} f_{i,j}}\omega_j\right)^{\log_2\left(\frac{p}{\sqrt{n}}+1\right)} X_{i,j}}$$

Where n is the total number of replicates for bait *i*. Figure S1 shows the modification's improvement over the original WD-score for each of the four analyzed datasets. In cases

where both proteins of an interaction are tested as baits, the larger CompPASS score was taken.

In addition to these direct APMS-dependent features, indirect characteristics of a putative protein-protein interaction were also employed within the Random Forest classification. The correlation between mRNA expression patterns of two genes was quantified using the Pearson correlation coefficient (PCC). In total, seven co-expression features – one for each species mentioned in Data Collection – were added to the classification model. The human feature is the PCC for the pair of human genes to be classified. There often exist multiple homologs of a gene within a different species; therefore the co-expression features for genes *i* and *j*, in non-human species *k*, were defined as the maximum PCC among the set of homolog pairs for that species, H_{ijk}:

$$Coex_{ijk} = \max(PCC_{mn}); m, n \in H_{ijk}$$

A separate feature was used for each of the five ontologies: biological process, cellular component, human phenotype, mouse phenotype, and human disease. Semantic similarity scores were utilized to determine how similar two gene's sets of annotations, A and B, were to each other, and used as the feature value. We computed semantic similarity scores using Resnik's MAX method (17). This method searches for the set of nearest common ancestors, C, among all pairs of annotations between two genes and returns the maximum information content.

$$sim(A,B) = \max_{a_i, b_j \in A,B} r(a_i, b_j)$$
$$r(a,b) = \max_{c \in C}[-\ln(p(c))]$$
$$p(c) = \begin{matrix}\text{percentage of genes with annotation c} \\ \text{or its descendants in the ontology}\end{matrix}$$

We used the Maximum Likelihood Estimation (20) method to calculate the probability of each potential domain-domain interaction, λ_{mn}. This required all two-hybrid interactions for *Saccharomyces cerevisiae* from BioGRID. *S. cerevisiae* was chosen for its high interactome coverage, leading to a more accurate estimation of domain-domain interaction probabilities. A single protein sequence was used for each gene, with preference given to the longest UniProt/SwissProt sequence, followed by the longest IPI sequence. A false positive rate of 0.005 and a false negative rate of 0.37 were used, which are required parameters of the method. Our feature – The probability that two proteins interact via their domains – was calculated using the set of potentially interacting domains, D_{ij}, present within the proteins:

$$Pr(I_{ij}=1) = 1 - \prod_{m,n \in D_{ij}} (1 - \lambda_{mn})$$

The final feature used was the number of known distinct interactions among the homologs of the two proteins:

$$HI_{ij} = \sum_{m,n \in H_{ij}} I_{mn} \quad I_{mn} = \begin{cases} 1\text{:known interaction} \\ 0\text{:otherwise} \end{cases}$$

Co-expression and ontologies features are subject to missing values due to lack of microarray probes and lack of annotations, respectively, and were treated as null. The Random Forest algorithm is designed to handle these during training and classification. Null values for ontology features were assigned to candidate interactions where at least one of the proteins has no annotations aside from the root of the ontology. Null values for co-expression features were assigned when COXPRESdb did not have data for the feature.

Training Set Construction

To segregate false positive protein interactions from true interactions, we trained and tested candidate classifiers using a supervised learning approach on four published human APMS datasets. Specifically, the four datasets analyzed described protein complexes associated with unique biological functions — deubiquitination (DUB)(11), autophagy (AIN)(39), chromatin remodeling (40)(TIP49), and transcriptional regulation (Complexome)(41) (Table I). These datasets range extensively in their number of experiments, interaction network connectivity and purification technique, resulting in a diverse training set capable of testing the generalizability of our method. All bait and prey identifiers were mapped to Entrez Gene IDs using the method described in the ID Mapping section. Interactions tested in both directions in a dataset were included only once. Within each APMS dataset, the CompPASS and HGSCore feature was normalized to have a mean of zero and a standard deviation of one. Candidate interactions annotated as physical interactions in BioGRID were used the positive class, excluding interactions represented by a single publication employing CompPASS or the HGSCore, as this would create a bias towards one of the methods. To avoid an extremely unbalanced training set, the negative set was created by uniformly sampling the unknown interactions from each dataset to contain 10 times the number of the dataset's known interactions.

Model Training and Evaluation

The Random Forest classifier was trained by creating 100 decision trees and splitting from a subset of 4 randomly selected features at each node. The positive class was given a weight of 10, while the negative class received a weight of 1. For cross-validation, each full dataset was tested with a classifier trained on the three remaining sampled datasets. Some overlaps were present among datasets; therefore interactions present in the dataset being tested were removed from the training set, avoiding the mistake of testing on trained data. In the case where an interaction was in multiple training datasets, one of the instances was selected at random. The metric for success was the receiver operating characteristic (ROC) curve. We treat all unknown interactions as the negative set. Though real interactions exist in this set, the number of false interactions is expected to greatly exceed the number of false negatives, making the ROC curve an appropriate metric.

False Discovery Rate Calculation

We used the probabilistic method employed by ProteinProphet to compute false discovery rates (FDR). First, interaction probabilities calculated by the Random Forest are sorted in descending order. Next, the FDR is calculated for the top x interactions using the equation:

$$FDR(x) = 1 - \frac{\sum_{i=1}^{i=x} P_i}{x} \quad \text{for } x \in 1, 2, \ldots, n$$

Finally, the user selects a FDR threshold, and the corresponding minimum interaction probability to accept is outputted.

FLAG Affinity Purification and Western Blot Analyses

For FLAG affinity purification, HEK293T cells were lysed in 0.1% NP-40 lysis buffer (10% glycerol, 50mM HEPES, 150 mM NaCl, 2mM EDTA, 0.1% NP-40) containing protease inhibitor mixture (1861278, Thermo Scientific, Waltham, MA) and phosphatase inhibitor (78427, Thermo Scientific, Waltham, MA). Cell lysates were cleared by centrifugation and incubated with FLAG resin (F2426, Sigma-Aldrich Corporation, St. Louis, MO) before washing with lysis buffer and eluting with NuPAGE loading buffer (Life Technologies, Carlsbad, CA). Detection of proteins by Western blot was performed using the following antibodies: anti-FLAG M2 monoclonal (Sigma-Aldrich Corporation, St. Louis, MO), anti-FAM117b (21768, ProteinTech, Chicago, IL), anti-MAD2L1 (A300-301A, Bethyl Labs, Montgomery, TX), anti-MCM3 (A300-192A, Bethyl Labs, Montgomery, TX), anti-SLK (A300-499A, Bethyl Labs, Montgomery, TX), anti-βactin polyclonal (A2066, Sigma-Aldrich Corporation, St. Louis, MO), anti-KEAP1 polyclonal (ProteinTech. Chicago, IL), anti-DPP3 polyclonal (97437, Abcam, Cambridge, MA), and anti-VSV polyclonal (A190-131A, Bethyl Labs, Montgomery, TX).

Mass Spectrometry

For FLAG affinity purification of KEAP1, HEK293T cells stably expressing FLAG-KEAP1 were lysed in 0.1% NP-40 lysis buffer (10% glycerol, 50mM HEPES, 150 mM NaCl, 2mM EDTA, 0.1% NP-40) containing protease inhibitor mixture (1861278, Thermo Scientific, Waltham, MA) and phosphatase inhibitor (78427, Thermo Scientific, Waltham, MA). Approximately 50mg of whole cell lysate was cleared by centrifugation and incubated with 20μL of packed FLAG resin (F2426, Sigma-Aldrich Corporation, St. Louis, MO) before washing 5 times with lysis buffer. Following an on-beads digestion with FASP Protein Digestion Kit (Protein Discovery, San Diego, CA), tryptic peptides were cleaned up by C18 Spin Column (Thermo Scientific, Waltham, MA), then separated by reverse phase nano-HPLC using a nanoAquity UPLC system (Waters Corp, Milford, MA). Peptides were first trapped in a 2 cm trapping column (75 μm ID, Michrom Magic C18 beads of 5.0 μM particle size, 200 Å pore size) and then separated on a self-packed 25 cm column (75 μm ID, Michrom Magic C18 beads of 3.0μM particle size, 100 Å pore size) at room temperature. The flow rate was 200 nl/min over a gradient of 1% buffer B (0.1% formic acid in acetonitrile) to 30% buffer B in 180 min. Then a following wash raised buffer B to 70%. The identity of the eluted peptides was determined with an in-line LTQ-Orbitrap Velos mass

spectrometer (Thermo Scientific, Waltham, MA). The ion source was operated at 2.0–2.4 kV with ion transfer tube temperature set at 275 °C. Full MS scan (300–2000 m/z) was acquired in Orbitrap with 60,000 resolution setting, data-dependent MS^2 spectra were acquired in LTQ by collision induced dissociation (CID) using the 20 most intense ions. Precursor ions were selected based on charge states (1, 2 or 3) and intensity thresholds (above 2000) from the full scan, dynamic exclusion (one repeat during 30-s, a 60-s exclusion time window) was also taken into account. The polysiloxane lock mass of 445.120030 was used throughout spectral acquisition.

Mass Spectrometry Data Analysis

All raw data were converted to mzXML format before a search of the resultant spectra using Sorcerer™-SEQUEST® (build 4.0.4, Sage N Research, Milpitas, CA) and the Transproteomic Pipeline (42)(TPP v4.3.1). Data were searched against the human UniProtKB/Swiss-Prot sequence database (Release 2011_08) supplemented with common contaminants, i.e. porcine (Swiss-Prot P00761) and bovine (P00760) trypsin, and further concatenated with its reversed copy as a decoy (40,494 total sequences). Search parameters used were a precursor mass between 400 and 4500 amu, up to 2 missed cleavages, precursor-ion tolerance of 3 amu, accurate mass binning within PeptideProphet (43), semi-tryptic digestion, a static carbamidomethyl cysteine modification, and variable methionine oxidation. Two KEAP1 experiments labeled MA128 were searched as one experiment through Sorcerer's interface, as they constituted the same sample run through the mass spectrometer in two halves. False discovery rates (FDR) were determined by ProteinProphet (44) and minimum protein probability cutoffs resulting in a 1% FDR were selected individually for each experiment. ProteinProphet results for each APMS experiment were stored in a local Prohits database (45) and Cytoscape v2.8.2(46) was used for network visualization. Unsearched data is available through the ProteoCommon.org Tranche network using the following hash:

V3N2GaJ7UTihN2nz03Dp1o424ntyGD8DtTI11Af8kTpolRCL8k3+EYd52Mh1Yc4wp0iZk E9T XXz9jE6C49+4EKs6aUwAAAAAAAADSA==

RESULTS

Combining APMS Scoring Methods Enriches for Previously Reported Protein Interactions

Existing spectral count-based APMS scoring methods demonstrate a high level of accuracy in predicting protein complex co-membership, thus making them appealing features for classification. A direct comparison of three popular and fundamentally distinct scoring algorithms—HGSCore, CompPASS and SAINT—revealed overlapping and complementary prediction accuracies (Figure 1). Specifically, the three methods were separately applied to each dataset, and individual thresholds were determined to achieve a 5% false positive rate, treating all un-annotated interactions as the negative set. The unions and intersections of each method's set of accepted interactions were then compared (Figure 1). Although some methods performed better than others, each of the three approaches was capable of identifying known protein-protein interactions disjoint from the remaining two. As expected, no single method identified all of the previously annotated protein interactions. That said,

the intersection of the three datasets showed strong enrichment for validated protein interactions. As such, the HGSCore and CompPASS (with a modified WD score) were chosen as features to broaden and strengthen the confidence of selected interactions, and for their computational speed.

Spotlite Outperforms Previous APMS Scoring Methods

To further improve upon interaction predictions, we chose to include into Spotlite data outside of APMS that had previously been shown to correlate with co-complexed proteins. These indirect sources of evidence were mRNA co-expression patterns among seven species, GO annotation similarity, phenotypic similarity, domain-domain binding affinities, and homologous interactions. Each was encoded into a feature, and along with the HGSCore and CompPASS, describe a putative pair of interacting proteins. Then, using the Random Forest algorithm, these interactions were predicted to be genuine based on the values of their corresponding feature vector.

In order to benchmark Spotlite against previous methods, we performed a variation of cross-validation by training a Random Forest classifier on each combination of three datasets and then testing on the remaining fourth dataset (Figure 2). Spotlite consistently outperformed SAINT, CompPASS and HGSCore based on ROC curve analysis and partial area indexes, which demonstrates greater sensitivity and specificity toward known interactions in the BioGRID database. These data also demonstrate that the discriminatory patterns learned from each dataset were generally applicable, as classification accuracy was superior across all cross-validation instances. To generate our final classifier for use in the web application, all four of the datasets were used for training. Table II shows each feature's coverage within the Spotlite database and its respective importance based on the Random Forest's Gini Index. As expected, the HGSCore and CompPASS were the two most important features used for distinguishing between known and false or unknown co-complexed proteins. Additionally, to determine the increase in accuracy due to the inclusion of indirect data, we calculated the number of misclassified interactions while performing stratified five-fold cross-validation with all features included, and compared it to training with the APMS features only. Impressively, misclassified interactions occurred at a rate of 34.95% using only APMS-derived features, and decreased to 30.58% after inclusion of the indirect features.

Spotlite Web Application for Public Use

We have made Spotlite available to the research community through a user-friendly web application that follows a simple workflow (Figure 3). Users may upload a tab-delimited file containing each experiment, its associated purification technique, bait, prey, and each prey's spectral count. An option for using the publically available APMS data deposited within Spotlite allows the smaller datasets of individual researchers to take advantage of the existing larger collection (using only the data of the same purification technique), thereby improving the filtering of the common contaminants by the HGSCore and CompPASS features. Next, the indirect feature data, which has been pre-computed for every potential pair of genes, is retrieved from the database. Finally, the data are classified by the Random Forest classifier. The false discovery rates are calculated and users can then explore and

visualize their results through the website or export them to a spreadsheet. For our largest available dataset, the Complexome, the entire process takes approximately two minutes to complete after the initial file upload. To maintain privacy, all uploaded APMS data and results are deleted after 24 hours of upload, or destroyed on command by the user.

Spotlite Analysis of KEAP1 APMS Data

To demonstrate its utility, performance and ease in identifying true interacting proteins from APMS data, we affinity purified the KEAP1 E3 ubiquitin ligase from HEK293T cells and analyzed the resulting data within Spotlite (Table S1). Specifically, cells engineered to stably express FLAG-tagged KEAP1 were detergent solubilized and subjected to FLAG affinity purification and shotgun mass spectrometry. Using biological duplicate KEAP1 APMS experiments and a reference set of an additional 22 FLAG purifications performed on 18 different baits, the KEAP1 protein interaction network was scored and visualized with Spotlite. The unfiltered KEAP1 dataset contained 534 prey proteins, of which 24 were annotated in BioGRID as being previously identified as KEAP1 interactors; 18 through high-throughput experiments and 6 using low-throughput methods (Figure 4A). After application of Spotlite and a global 10% FDR threshold, the network reduced to 35 proteins, of which three were shown to be genuine through low-throughput experiments, nine through high-throughput methods and 23 putative novel interactors (Figure 4B). Next, we selected eight KEAP1 interacting proteins that passed Spotlite thresholding for further validation by immunoprecipitation and Western blot analysis: MCM3, DPP3, SLK, MCC, MCMBP, MAD2L1, SQSTM1 and FAM117B. Of these, five were previously annotated as high-throughput interactors, one by low-throughput assays and two were novel interactors (MCC and FAM117B). Impressively, all eight endogenously expressed proteins co-purified with FLAG-tagged KEAP1 (Figure 4C).

In addition to providing the Random Forest classification score, the Spotlite web application lists the following individual features for each protein pair: HGSCore (HGS), CompPASS, gene ontologies for biological process (BP) and cellular component (CC), gene co-expression for seven species (CXP), domain-domain binding score (Domain), number of homologous interactions (Homo int), shared phenotypes (Phen), shared human diseases (Disease) and whether the proteins have previously been shown to interact (Known?). As an example, Spotlite's visualization for the KEAP1-MAD2L1 interaction is provided in Figure 5. Both proteins affect growth and size in mice, specifically postnatal growth retardation with KEAP1 and decreased embryo size with MAD2L1. Additionally, both proteins are encoded by mRNAs which positively correlate across human tissues, and both proteins are strongly associated with oncogenesis.

DISCUSSION

Protein mass spectrometry is quickly becoming a staple technology in academic laboratories. The rapidly decreasing instrumentation costs, often pre-packaged and streamlined bioinformatic pipelines and enhanced mass accuracy and scan speeds are no doubt driving the recent explosion of protein mass spectrometry data. With similar advances in two-hybrid technologies, it is now economically feasible to pursue, and in fact achieve a

proteome-wide connectivity map. A key step in this endeavor, or in the comprehensive definition of any biomolecule—whether that be genomics, transcriptomics, metabolomics or proteomics—is the computational scoring and integration of the resulting datasets.

After performing hundreds of APMS experiments directed at mapping protein connectivity central to various signal transduction pathways, we and others quickly found the high rate of false-positive identification rate limiting and exceedingly expensive. Appreciating the need for an accessible and accurate APMS scoring algorithm, we developed Spotlite as a new computational tool capable of revealing true interactions from the contaminants within APMS data. Importantly, we deployed Spotlite through a web-based application which provides open access and full transparency to any interested scientist. The inclusion of indirect data as features within Spotlite's Random Forest classification not only provides 4.35% increased prediction accuracy but also yields valuable information regarding shared biological function, phenotype and disease relationships among protein pairs.

Given the success of established scoring approaches employed by CompPASS, HGSCore and SAINT, we initially set out to define their relative performance on various APMS datasets, and by doing so to identify the most accurate approach for implementation within a classification scheme. However, our analyses revealed valuable complementarity between the algorithms, which appeared partially dependent upon the network architecture of the analyzed APMS dataset. As such, we found greatest success by providing the Random Forest classifier with both CompPASS and HGSCore; given its relatively slow computation speed, inclusion of SAINT was not feasible for a web-based application. Though Spotlite's performance shows a marked improvement over existing methods, its success is governed by the small number of known protein interactions (positive dataset), the lack of validated non-interactions (negative dataset), and mislabeled instances used during training. Furthermore, many indirect features lacked high coverage, resulting in missing values. While these limitations may place a ceiling on current performance, data will continue to pour in and fill the gaps. We expect Spotlite to improve over time due to increased feature coverage, and re-training of the classifier as larger interaction networks become available.

A critical aspect of any supervised learning approach is the selection of a gold standard dataset containing accurately labeled examples that are representative of the future data to be classified. While many protein-protein interactions are annotated, proteins known not to interact are rare—the Negatome being the sole available resource and of prohibitively small size (47). In addition to our approach of using known interactions contained in the APMS data as our positive class and random samples from the APMS unknown interactions as the negative, we explored alternative strategies for defining our training set, namely the common practice for protein-protein interaction prediction of using all known interactions and a random sample for the noninteracting class. Classifiers were trained by including all known interactions, with missing values for the HGSCore and CompPASS, and also by including both known and randomly sampled unknowns, neither of which were able to match the accuracy of training on APMS interactions alone. A possible reason for this is a deemphasized role of the APMS scoring methods. It is also important to note that as we are predicting co-complexed proteins, our set of known interactions included both direct and co-complexed interactions.

Presently, Spotlite supports human APMS data exclusively; however, it can be extended to other species by compiling the data for their features. Aside from analyzing another species' data using the current workflow, we envision the possibility of using APMS from multiple species to improve predictions through homologous interactions, which is already a powerful feature in our implementation. Along these lines, merging datasets from various laboratories has the potential to further increase accuracy. While this is currently an optional feature within Spotlite, it should be done with great care as contaminants will vary due to differences in cell lines, mass spectrometers and protocols, leading to improperly high CompPASS and HGSCore values for mutually exclusive contaminants which now appear more unique. This combined analysis of datasets is an area of future research.

A major focus of our research focuses on the development of proteomic and functional genomic technologies to define the mechanics and disease contribution of the KEAP1. The KEAP1 protein functions as a CUL3-based E3 ubiquitin ligase, most well-known for its ubiquitination of the NFE2L2 transcription factor (48–50). Recently, somatic inactivating mutations in KEAP1 have been reported in a variety of solid human tumors, particularly in lung cancer (51–59). The leading model posits that KEAP1 inactivation results in constitutive NFE2L2 transcriptional activation of antioxidant and pro-survival genes (60, 61). APMS analysis of KEAP1 followed by Spotlite scoring and a 10% FDR filter revealed 35 associated proteins. Of the eight proteins validated to reside within KEAP1 protein complexes by IP/Western blot, the indirect data—as visualized through the Spotlite web application—drew attention to the KEAP1-MAD2L1 protein association. Specifically, the MAD2L1 protein is known to function pivotally within the spindle assembly checkpoint complex, which holds cells in metaphase until chromosome-spindle attachment is complete (62–64). Like KEAP1, MAD2L1 is strongly associated with cancer; its over-expression drives chromosomal instability and aneuploidy (65, 66). MAD2L2 is also known to be ubiquitinated, although the E3 ubiquitin ligase is unknown (67, 68). An intriguing possibility is that KEAP1 ubiquitinates MAD2L1 to control its activity and/or stability. Within cancer systems, somatic mutation of KEAP1 may coincide with elevated MAD2L1 activity, thus driving aneuploidy.

In conclusion, we have provided a user-friendly web application for predicting complex co-membership from APMS data. This web application employs a novel, Random Forest classifier that integrates existing, proven APMS scoring approaches, gene co-expression patterns, functional annotations, phenotypic observations, protein domains, and homologous interactions, which we have shown outperforms existing APMS scoring methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Wade Harper, Mathew E. Sowa, Alexey Nesvizhskii, Hyungwon Choi, Jun Qin and Anna Malovannaya for kindly providing APMS data in a format suitable for use in Spotlite. We also thank members of the Major lab for helpful comments and criticisms.

This work was supported in part by The Sidney Kimmel Foundation for Cancer Research Scholar Award, The NIH Directors New Innovator Award (DP2) and The National Science Foundation's Information and Intelligent Systems under grant IIS0812464.

ABBREVIATIONS

APMS	Affinity Purification – Mass Spectrometry
ROC	Receiver Operating Characteristic
FDR	False Discovery Rate
PCC	Pearson Correlation Coefficient
GO	Gene Ontology
IP	Immunoprecipitation

References

1. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M. The BioGRID Interaction Database: 2011 update. Nucleic acids research. 2011; 39:D698–704. [PubMed: 21071413]
2. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006; 440:631–636. [PubMed: 16429126]
3. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular & cellular proteomics : MCP. 2007; 6:439–450. [PubMed: 17200106]
4. Bader GD, Hogue CW. Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol. 2002; 20:991–997. [PubMed: 12355115]
5. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003; 4:2. [PubMed: 12525261]
6. Gilchrist MA, Salter LA, Wagner A. A statistical framework for combining and interpreting proteomic datasets. Bioinformatics. 2004; 20:689–700. [PubMed: 15033876]
7. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics. 2007; 8:236. [PubMed: 17605818]
8. Zhang B, Park BH, Karpinets T, Samatova NF. From pull-down data to protein interaction networks and complexes with biological relevance. Bioinformatics. 2008; 24:979–986. [PubMed: 18304937]
9. Choi H, Larsen B, Lin ZY, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras AC, Nesvizhskii AI. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nat Methods. 2011; 8:70–73. [PubMed: 21131968]
10. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, Shales M, Mercenne G, Pache L, Li K, Hernandez H, Jang GM, Roth SL, Akiva E, Marlett J, Stephens M, D'Orso I, Fernandes J, Fahey M, Mahon C, O'Donoghue AJ, Todorovic A, Morris JH, Maltby DA, Alber T, Cagney G, Bushman FD, Young JA, Chanda SK, Sundquist WI, Kortemme T, Hernandez RD, Craik CS, Burlingame A, Sali A, Frankel AD, Krogan NJ. Global landscape of HIV-human protein complexes. Nature. 2012; 481:365–370. [PubMed: 22190034]
11. Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell. 2009; 138:389–403. [PubMed: 19615732]
12. Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, McKillip E, Shah S, Stapleton M, Wan KH, Yu C, Parsa B, Carlson JW, Chen X,

Kapadia B, VijayRaghavan K, Gygi SP, Celniker SE, Obar RA, Artavanis-Tsakonas S. A protein complex network of Drosophila melanogaster. Cell. 2011; 147:690–703. [PubMed: 22036573]

13. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet. 2007; 8:699–710. [PubMed: 17703239]
14. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. Bioinformatics. 2007; 23:2322–2330. [PubMed: 17599939]
15. Qiu J, Noble WS. Predicting co-complexed protein pairs from heterogeneous data. PLoS Comput Biol. 2008; 4:e1000054. [PubMed: 18421371]
16. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins. 2006; 63:490–500. [PubMed: 16450363]
17. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI: Proceedings of the 14th international joint conference on artificial intelligence; San Francisco CA: Morgan Kaufmann Publishers, Inc; 1995.
18. Jain S, Bader GD. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC Bioinformatics. 2010; 11:562. [PubMed: 21078182]
19. Yang H, Nepusz T, Paccanaro A. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics. 2012; 28:1383–1389. [PubMed: 22522134]
20. Deng M, Mehta S, Sun F, Chen T. Inferring domain-domain interactions from protein-protein interactions. Genome research. 2002; 12:1540–1548. [PubMed: 12368246]
21. Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. Bioinformatics. 2005; 21(Suppl 1):i38–46. [PubMed: 15961482]
22. Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines. Protein engineering, design & selection : PEDS. 2004; 17:165–173.
23. Lin N, Wu B, Jansen R, Gerstein M, Zhao H. Information assessment on predicting protein-protein interactions. BMC Bioinformatics. 2004; 5:154. [PubMed: 15491499]
24. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003; 302:449–453. [PubMed: 14564010]
25. Bader JS, Chaudhuri A, Rothberg JM, Chant J. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol. 2004; 22:78–85. [PubMed: 14704708]
26. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P. STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic acids research. 2007; 35:D358–362. [PubMed: 17098935]
27. McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic acids research. 2009; 37:D651–656. [PubMed: 18988626]
28. Breiman L. Random Forests. Machine Learning. 2001; 45:28.
29. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. Proteomics. 2004; 4:1985–1988. [PubMed: 15221759]
30. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research. 2012; 40:D71–75. [PubMed: 22102590]
31. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic acids research. 2012; 40:D290–301. [PubMed: 22127870]
32. Obayashi T, Kinoshita K. COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic acids research. 2011; 39:D1016–1022. [PubMed: 21081562]
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]

34. Smith CL, Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley interdisciplinary reviews. Systems biology and medicine. 2009; 1:390–399. [PubMed: 20052305]
35. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics. 2008; 83:610–615. [PubMed: 18950739]
36. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu LJ, Danila MI, Feng G, Chisholm RL. Annotating the human genome with Disease Ontology. BMC Genomics. 2009; 10(Suppl 1):S6. [PubMed: 19594883]
37. Mark Hall EF, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter, Witten Ian H. The WEKA data mining software: an update. SIGKDD Explorations. 2009:11.
38. Zybailov BL, Florens L, Washburn MP. Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. Molecular bioSystems. 2007; 3:354–360. [PubMed: 17460794]
39. Behrends C, Sowa ME, Gygi SP, Harper JW. Network organization of the human autophagy system. Nature. 2010; 466:68–76. [PubMed: 20562859]
40. Sardiu ME, Cai Y, Jin J, Swanson SK, Conaway RC, Conaway JW, Florens L, Washburn MP. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proc Natl Acad Sci U S A. 2008; 105:1454–1459. [PubMed: 18218781]
41. Malovannaya A, Lanz RB, Jung SY, Bulynko Y, Le NT, Chan DW, Ding C, Shi Y, Yucer N, Krenciute G, Kim BJ, Li C, Chen R, Li W, Wang Y, O'Malley BW, Qin J. Analysis of the human endogenous coregulator complexome. Cell. 2011; 145:787–799. [PubMed: 21620140]
42. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010; 10:1150–1159. [PubMed: 20101611]
43. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical chemistry. 2002; 74:5383–5392. [PubMed: 12403597]
44. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Analytical chemistry. 2003; 75:4646–4658. [PubMed: 14632076]
45. Liu G, Zhang J, Larsen B, Stark C, Breitkreutz A, Lin ZY, Breitkreutz BJ, Ding Y, Colwill K, Pasculescu A, Pawson T, Wrana JL, Nesvizhskii AI, Raught B, Tyers M, Gingras AC. ProHits: integrated software for mass spectrometry-based interaction proteomics. Nat Biotechnol. 2010; 28:1015–1017. [PubMed: 20944583]
46. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011; 27:431–432. [PubMed: 21149340]
47. Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A. The Negatome database: a reference set of non-interacting protein pairs. Nucleic acids research. 2010; 38:D540–544. [PubMed: 19920129]
48. Cullinan SB, Gordan JD, Jin J, Harper JW, Diehl JA. The Keap1-BTB protein is an adaptor that bridges Nrf2 to a Cul3-based E3 ligase: oxidative stress sensing by a Cul3-Keap1 ligase. Mol Cell Biol. 2004; 24:8477–8486. [PubMed: 15367669]
49. Furukawa M, Xiong Y. BTB protein Keap1 targets antioxidant transcription factor Nrf2 for ubiquitination by the Cullin 3-Roc1 ligase. Mol Cell Biol. 2005; 25:162–171. [PubMed: 15601839]
50. Zhang DD, Lo SC, Cross JV, Templeton DJ, Hannink M. Keap1 is a redox-regulated substrate adaptor protein for a Cul3-dependent ubiquitin ligase complex. Mol Cell Biol. 2004; 24:10941–10953. [PubMed: 15572695]
51. Padmanabhan B, Tong KI, Ohta T, Nakamura Y, Scharlock M, Ohtsuji M, Kang MI, Kobayashi A, Yokoyama S, Yamamoto M. Structural basis for defects of Keap1 activity provoked by its point mutations in lung cancer. Mol Cell. 2006; 21:689–700. [PubMed: 16507366]
52. Singh A, Misra V, Thimmulappa RK, Lee H, Ames S, Hoque MO, Herman JG, Baylin SB, Sidransky D, Gabrielson E, Brock MV, Biswal S. Dysfunctional KEAP1-NRF2 interaction in non-small-cell lung cancer. PLoS Med. 2006; 3:e420. [PubMed: 17020408]

53. Ohta T, Iijima K, Miyamoto M, Nakahara I, Tanaka H, Ohtsuji M, Suzuki T, Kobayashi A, Yokota J, Sakiyama T, Shibata T, Yamamoto M, Hirohashi S. Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. Cancer Res. 2008; 68:1303–1309. [PubMed: 18316592]
54. Satoh H, Moriguchi T, Taguchi K, Takai J, Maher JM, Suzuki T, Winnard PT Jr, Raman V, Ebina M, Nukiwa T, Yamamoto M. Nrf2-deficiency creates a responsive microenvironment for metastasis to the lung. Carcinogenesis. 2010; 31:1833–1843. [PubMed: 20513672]
55. Solis LM, Behrens C, Dong W, Suraokar M, Ozburn NC, Moran CA, Corvalan AH, Biswal S, Swisher SG, Bekele BN, Minna JD, Stewart DJ, Wistuba II. Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. Clin Cancer Res. 2010; 16:3743–3753. [PubMed: 20534738]
56. Takahashi T, Sonobe M, Menju T, Nakayama E, Mino N, Iwakiri S, Nagai S, Sato K, Miyahara R, Okubo K, Hirata T, Date H, Wada H. Mutations in Keap1 are a potential prognostic factor in resected non-small cell lung cancer. J Surg Oncol. 2010; 101:500–506. [PubMed: 20213688]
57. Konstantinopoulos PA, Spentzos D, Fountzilas E, Francoeur N, Sanisetty S, Grammatikos AP, Hecht JL, Cannistra SA. Keap1 mutations and Nrf2 pathway activation in epithelial ovarian cancer. Cancer Res. 2011; 71:5081–5089. [PubMed: 21676886]
58. Li QK, Singh A, Biswal S, Askin F, Gabrielson E. KEAP1 gene mutations and NRF2 activation are common in pulmonary papillary adenocarcinoma. J Hum Genet. 2011; 56:230–234. [PubMed: 21248763]
59. Muscarella LA, Parrella P, D'Alessandro V, la Torre A, Barbano R, Fontana A, Tancredi A, Guarnieri V, Balsamo T, Coco M, Copetti M, Pellegrini F, De Bonis P, Bisceglia M, Scaramuzzi G, Maiello E, Valori VM, Merla G, Vendemiale G, Fazio VM. Frequent epigenetics inactivation of KEAP1 gene in non-small cell lung cancer. Epigenetics : official journal of the DNA Methylation Society. 2011; 6:710–719. [PubMed: 21610322]
60. Sykiotis GP, Bohmann D. Stress-activated cap'n'collar transcription factors in aging and human disease. Sci Signal. 2010; 3:re3. [PubMed: 20215646]
61. Ogura T, Tong KI, Mio K, Maruyama Y, Kurokawa H, Sato C, Yamamoto M. Keap1 is a forked-stem dimer structure with two large spheres enclosing the intervening, double glycine repeat, and C-terminal domains. Proc Natl Acad Sci U S A. 2010; 107:2842–2847. [PubMed: 20133743]
62. Hoyt MA, Totis L, Roberts BT. S. cerevisiae genes required for cell cycle arrest in response to loss of microtubule function. Cell. 1991; 66:507–517. [PubMed: 1651171]
63. Li R, Murray AW. Feedback control of mitosis in budding yeast. Cell. 1991; 66:519–531. [PubMed: 1651172]
64. Musacchio A, Salmon ED. The spindle-assembly checkpoint in space and time. Nat Rev Mol Cell Biol. 2007; 8:379–393. [PubMed: 17426725]
65. Sotillo R, Hernando E, Diaz-Rodriguez E, Teruya-Feldstein J, Cordon-Cardo C, Lowe SW, Benezra R. Mad2 overexpression promotes aneuploidy and tumorigenesis in mice. Cancer Cell. 2007; 11:9–23. [PubMed: 17189715]
66. Schvartzman JM, Duijf PH, Sotillo R, Coker C, Benezra R. Mad2 is a critical mediator of the chromosome instability observed upon Rb and p53 pathway inhibition. Cancer Cell. 2011; 19:701–714. [PubMed: 21665145]
67. Osmundson EC, Ray D, Moore FE, Gao Q, Thomsen GH, Kiyokawa H. The HECT E3 ligase Smurf2 is required for Mad2-dependent spindle assembly checkpoint. The Journal of cell biology. 2008; 183:267–277. [PubMed: 18852296]
68. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, Sowa ME, Rad R, Rush J, Comb MJ, Harper JW, Gygi SP. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell. 2011; 44:325–340. [PubMed: 21906983]

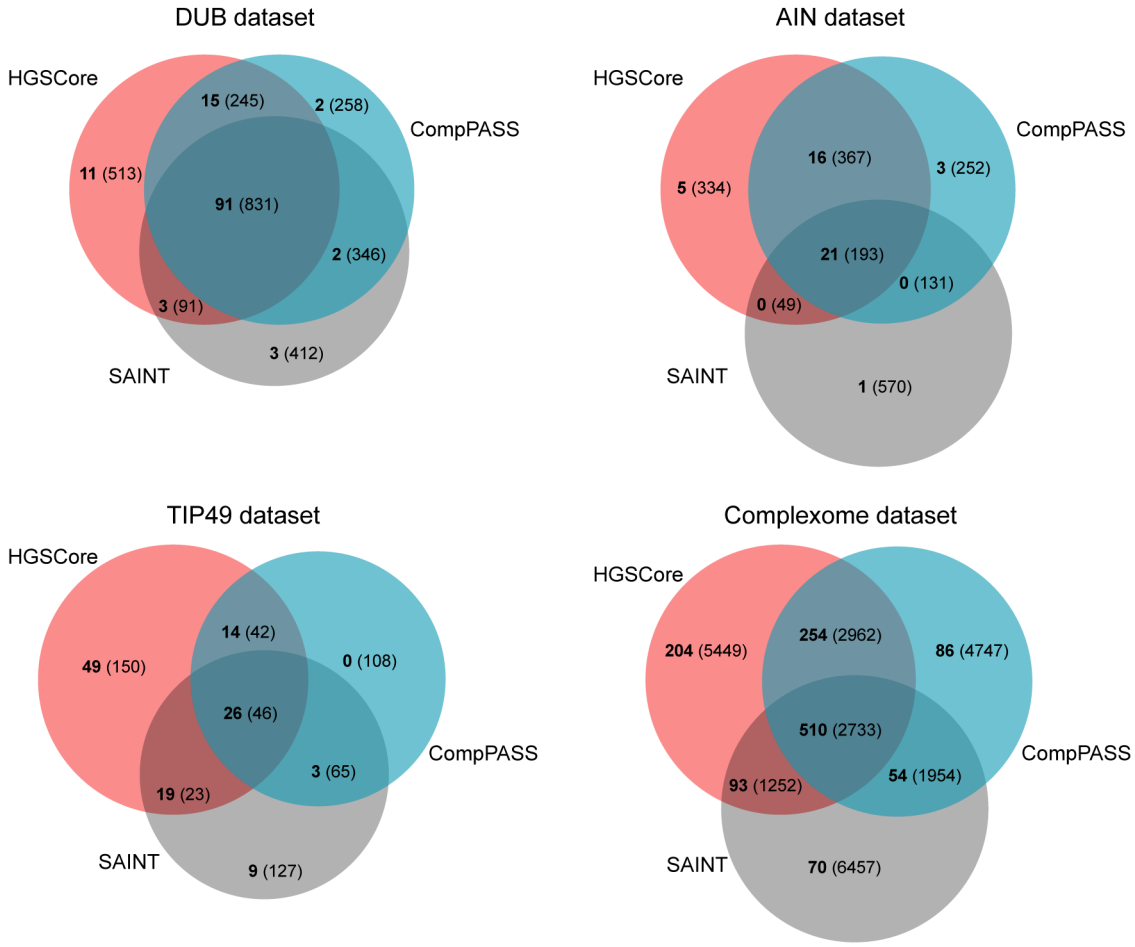

Figure 1. Comparison of accepted interactions using various APMS scoring methods

Venn diagrams show sets of previously established interactions using a 5% false positive rate threshold for each scoring method. The counts of known interactions are in bold and the counts of false or unknown interactions are in parentheses. Areas are proportional to the total number of interactions within their respective subsets.

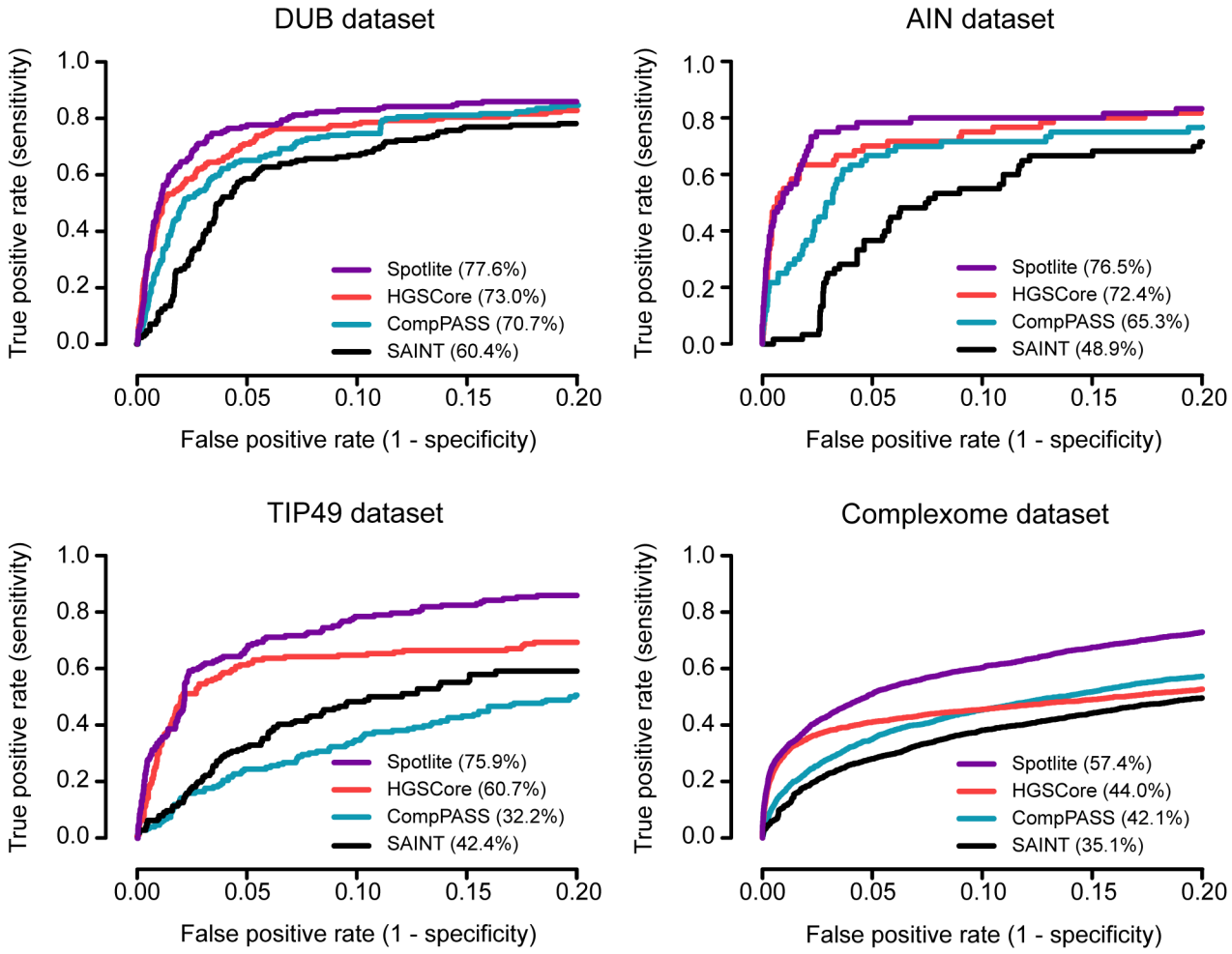

Figure 2. Random Forest cross-validation and comparison

Receiver operating characteristic curves for each dataset. Values in parentheses represent partial area indexes.

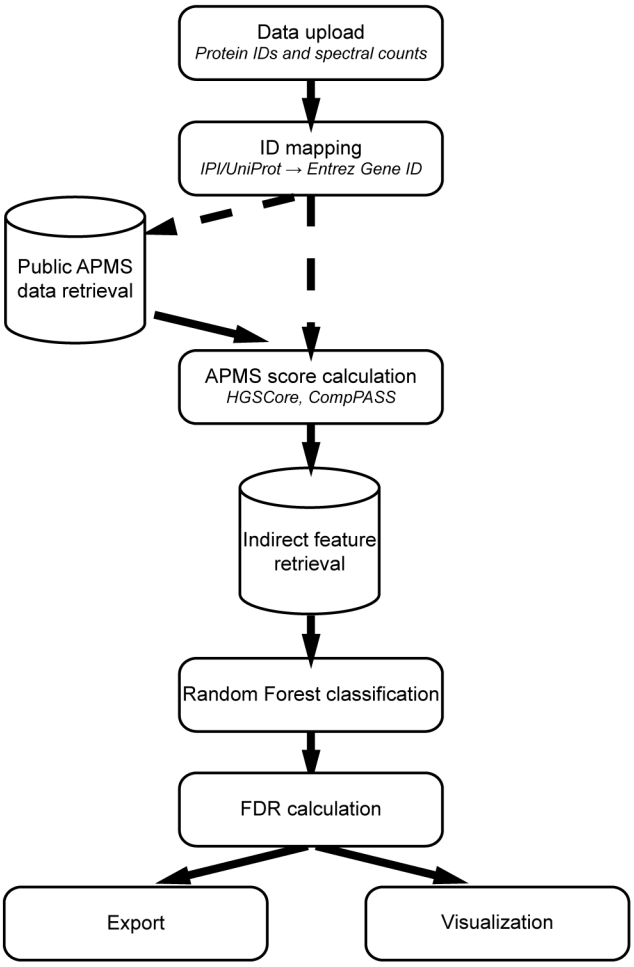

Figure 3. Schematic of Spotlite workflow

Dashed lines represent optional paths selected by the user. Upon uploading a new dataset, the user can choose to retrieve public APMS data of the same affinity purification technique before calculating scores for CompPASS and the HGSCore.

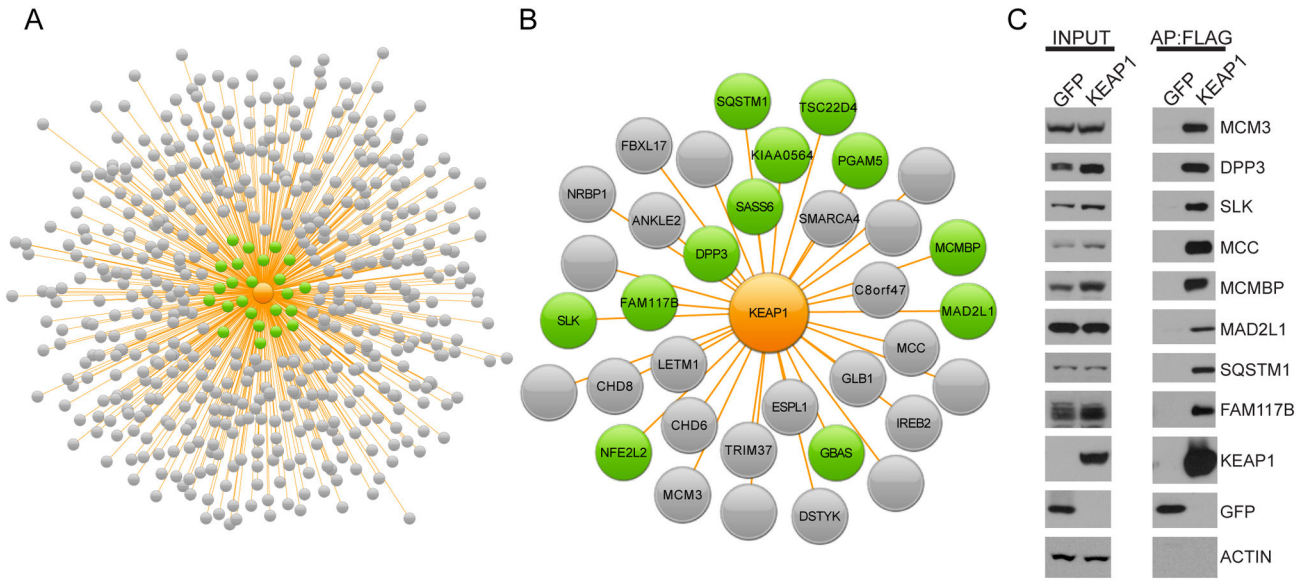

Figure 4. Spotlite application to KEAP1 APMS

(A) The unfiltered spoke model network of KEAP1 represents 534 nodes. Nodes were sized by total spectral counts, as indicated within the key. Green nodes represent proteins reported in BioGRID to form complexes with KEAP1. **(B)** Spotlite filtered spoke model network using a false discovery rate threshold of 10%. Eight unlabeled nodes represent possible contaminants. **(C)** FLAG affinity purified protein complexes from HEK293T cells stably expressing FLAG-GFP or FLAG-KEAP1 were analysed by Western blot for the indicated endogenously expressed proteins.

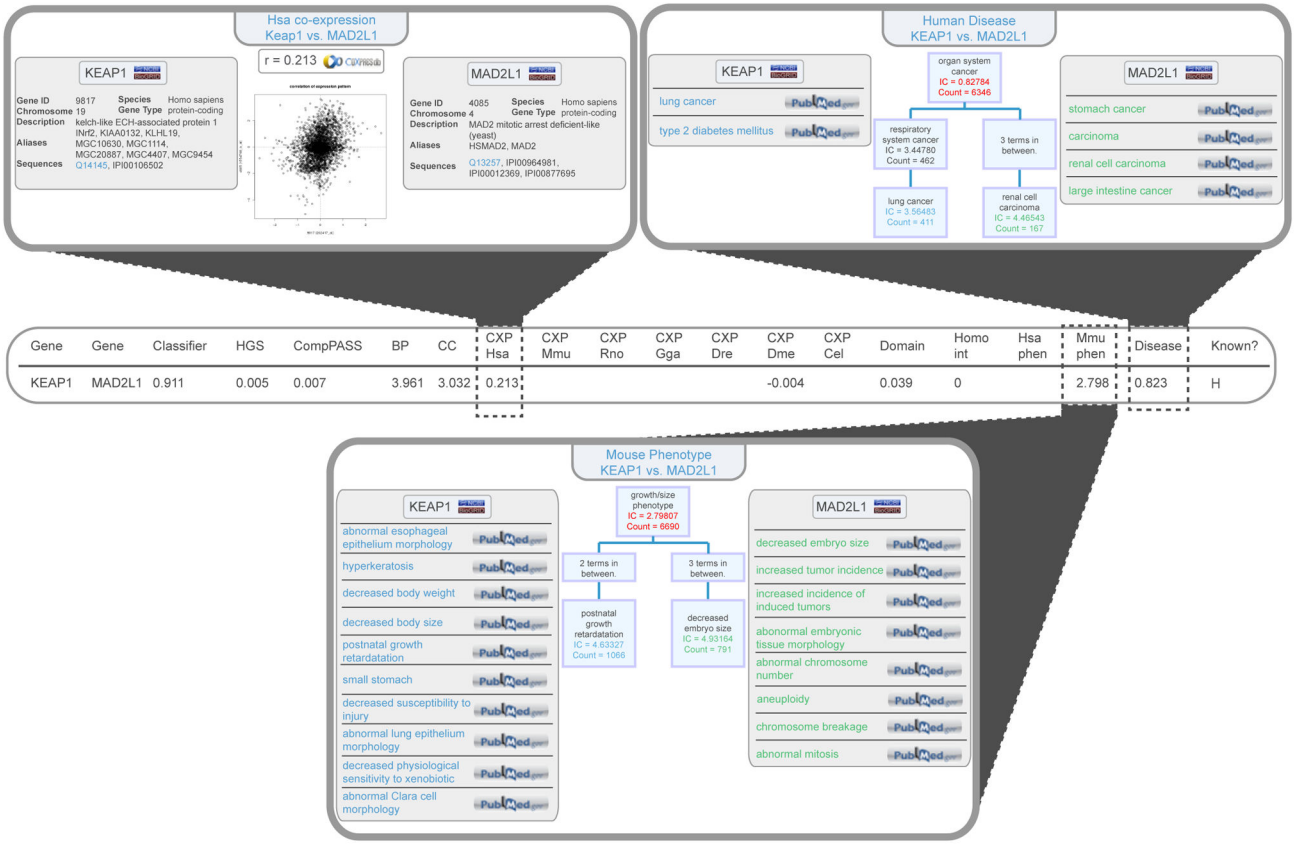

Figure 5. Screenshots of Spotlite visualization for KEAP1-MAD2L1 data

Column headers on the main results screen are the following: Spotlite score (Classifier), HGSCore (HGS), CompPASS, gene ontologies for biological process (BP) and cellular component (CC), gene co-expression for seven species (CXP), domain-domain binding score (Domain), number of homologous interactions (Homo int), shared phenotypes (phen), shared human diseases (Disease) and whether the proteins have previously been shown to interact (Known?; H=high throughput, L=low throughput). Transparency is provided through a series of user-triggered pop-up windows which details the information used to generate the Spotlite feature score.

Table I

Public dataset statistics

Dataset	AP Method	Experiments	Baits	Distinct Interactions	Total Interactions	Mean Clustering Coefficient[a]
Complexome	Antibody	3268	1082	254452	388414	0.2080
DUB	HA	201	101	35735	58068	0.0757
AIN	HA	127	64	19720	31328	0.2264
TIP49	FLAG	35[b]	27	5436	6218	0.2307

[a]Computed using a protein-protein interaction network comprised of only bait nodes, and edges between them derived from BioGRID using experiments testing direct interactions- reconstituted complex, co-crystal structture, protein-peptide, FRET, and two-hybrid.

Table II

Feature importances for Random Forest classifier

Feature	Type[a]	Database Coverage[b]	Training Coverage[c]	Gini Importance[d]
HGSCore	Direct	11.79%	100.00%	1.88
CompPASS	Direct	11.79%	100.00%	1.54
Human co-expression	Indirect	69.29%	92.92%	1.37
Chicken co-expression	Indirect	11.89%	22.33%	1.25
Fish co-expression	Indirect	4.51%	17.11%	1.24
Fly co-expression	Indirect	2.20%	9.85%	1.18
Mouse co-expression	Indirect	29.31%	47.77%	1.13
Worm co-expression	Indirect	1.36%	5.48%	1.09
Rat co-expression	Indirect	13.63%	26.50%	1.08
Disease ontology	Functional	3.98%	11.69%	1.02
Mouse phenotype ontology	Functional	8.53%	20.75%	0.95
Human phenotype ontology	Functional	0.80%	1.91%	0.93
Biological process GO	Functional	48.66%	84.33%	0.81
Homologous interactions	Sequence	85.86%	99.53%	0.77
Cellular localization GO	Functional	61.69%	86.02%	0.71
Domain-domain binding affinity	Sequence	70.32%	88.33%	0.05

[a] Classification of the type of evidence a feature represents with respect to co-complexed proteins.

[b] Percentage of all potentially co-complexed pairs of genes within the Spotlite database containing values for a feature. HGSCore and CompPASS coverages represent the percentage of bait-prey interactions tested, including preys with 0 spectra. Ontology coverages computed by taking the percentage of gene pairs in which both genes have ≥ 1 annotation. Homologous interactions coverage - both genes must have a known homolog in the same species. Domain-domain binding affinity coverage - both genes must contain a known domain.

[c] Coverages calculated identically to b - restricted to the training dataset.

[d] Feature importance measure computed by Random Forests.