Published online 11 February 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 1 **1**
doi: 10.1093/nargab/lqaa005

UniProt-Related Documents (UniReD): assisting wet lab biologists in their quest on finding novel counterparts in a protein network

Theodosios Theodosiou[1], Nikolaos Papanikolaou[1], Maria Savvaki[1,2], Giulia Bonetto[1], Stella Maxouri[1,3], Eirini Fakoureli[1], Aristides G. Eliopoulos[4], Nektarios Tavernarakis[1,2], Grigoris D. Amoutzias[5], Georgios A. Pavlopoulos[6], Michalis Aivaliotis[2,7,8], Vasiliki Nikoletopoulou[2], Dimitris Tzamarias[2], Domna Karagogeos[1,2] and Ioannis Iliopoulos[1,*]

[1]University of Crete, School of Medicine, Department of Basic Sciences, Heraklion 71003, Crete, Greece, [2]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, Nikolaou Plastira 100, 70013 Heraklion, Crete, Greece, [3]Medical School of Patras University, Laboratory of General Biology, Asklipiou 1, 26500 Rio Patras, Greece, [4]Department of Biology, Medical School, National and Kapodistrian University of Athens, Mikras Asias 75, 11527 Athens, Greece, [5]Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, Larisa 41500, Greece, [6]Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", 34 Fleming Street, 16672 Vari, Greece, [7]Laboratory of Biological Chemistry, Faculty of Health Sciences, School of Medicine, Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece and [8]Functional Proteomics and Systems Biology (FunPATh), Center for Interdisciplinary Research and Innovation (CIRI-AUTH), Balkan Center, Thessaloniki, 10th km Thessaloniki-Thermi Rd, P.O.Box 8318, GR 57001, Greece

Received August 02, 2019; Revised January 20, 2020; Editorial Decision January 23, 2020; Accepted January 31, 2020

ABSTRACT

The in-depth study of protein–protein interactions (PPIs) is of key importance for understanding how cells operate. Therefore, in the past few years, many experimental as well as computational approaches have been developed for the identification and discovery of such interactions. Here, we present UniReD, a user-friendly, computational prediction tool which analyses biomedical literature in order to extract known protein associations and suggest undocumented ones. As a proof of concept, we demonstrate its usefulness by experimentally validating six predicted interactions and by benchmarking it against public databases of experimentally validated PPIs succeeding a high coverage. We believe that UniReD can become an important and intuitive resource for experimental biologists in their quest for finding novel associations within a protein network and a useful tool to complement experimental approaches (e.g. mass spectrometry) by producing sorted lists of candidate proteins for further experimental validation. UniReD is available at http://bioinformatics.med.uoc.gr/unired/

INTRODUCTION

Protein–protein interactions (PPIs) play an essential role in most complex biological processes and are often classified as direct (when proteins interact physically) or indirect (e.g. when proteins are involved in the same pathway or biological process). Due to their important role in biochemical and cellular processes, PPIs have been for many years subjected to intensive study in order to understand how complex biological networks function and how a complex biological system can work as a unit.

For a large-scale experimental identification of PPIs, several high-throughput techniques have been used. Widely used methods are the yeast-two-hybrid system (1,2), protein arrays (3,4), co-immunoprecipitation and mass spectrometry (5). However, these techniques have certain drawbacks including high cost, low accuracy, high false positive rate and they are often time-consuming (6). In addition, it has been reported that there is a significant discrepancy when comparing data produced by different high-throughput experiments (7). Therefore, the need for computational tools capable of identification, prediction and validation of protein associations as a complement to existing experimental approaches (8) emerges. The latter provides the option for prediction of not only direct/physical PPIs, but also for indirect interactions like for example possible involve-

*To whom correspondence should be addressed. Tel: +30 2810 394539; Fax: +30 2810 394530; Email: iliopj@med.uoc.gr

ment in a pathway (8,9). *In silico* approaches for the prediction of PPIs can be roughly summarized in two categories based on the source of the employed data: (i) those exploiting molecular/genomic data (e.g. protein structures (10), phylogenetic profiling (11), gene fusion detection (12) and genomic neighborhood (13), and (ii) biomedical literature mining methods, taking advantage, for example, of the huge amount of information hidden in today's >30 million abstracts hosted in PubMed (version 10/2019).

The concept that text mining techniques can be used for knowledge extraction and new knowledge discovery came after Swanson's studies, where he directly linked the treatment of Raynaud's disease with fish oil (14,15). During the last thirty years, an increasing amount of publications reporting text mining techniques for exploiting biomedical literature has become evident (16). Among them, many efforts trying to extract PPIs from PubMed abstracts or create a network of genes/proteins/biological terms have appeared (17). For example, PubGene can detect associations between genes using terms from the medical subject heading (MeSH) index and terms from the gene ontology (GO) database (18), CoPub Mapper provides online access to co-occurrence associations between genes and biological terms extracted from PubMed (19). HIPPIE (20) generates reliable and meaningful human PPI networks through PPI network scoring, integration of different types of experimental information and basic graph algorithms for highlighting important proteins. More complex systems, such as iHOP (21) allows to search information through hyperlinks so that literature is clustered according to gene names and text topics, thus leading to possible PPIs. Other approaches extract PPIs from scientific literature through the identification of protein names in text (17), others are able to perform sentence-based semantic analysis (22–24) and others connect proteins to concept profiles following the assumption that proteins which share one or more concept profiles have an increased probability to interact (25). Notably, all PPIs which are experimentally verified are highlighted elsewhere (21).

In addition to the above methods, other tools which integrate data from various sources and predict a PPI, exist. One such web tool is STRING (26), which uses information from biomedical literature (co-occurrences of gene names in PubMed abstracts), high-throughput experiments, conserved co-expression, gene neighborhood, gene fusion events, phylogenetic profiling and curated databases, to show protein associations for more than 1200 species. Similarly, van Haagen *et al*. (27) developed a method similar to STRING, which uses biomedical literature data, mRNA expression patterns and protein domain information. Furthermore, there are several online manually curated databases which contain known and experimentally validated PPIs. Few of them which are also often used for benchmarking are the Reactome (28), BioGRID (29), DIP (30), HitPredict (31), MINT (32), IntAct (33) and BIND (34).

Herein, we introduce a novel computational approach which is able to not only identify known associations between proteins described in the biomedical literature, but also to predict novel interactions which are not yet experimentally documented. The methodology is offered as online service and its functionality is accompanied by both a statistical and an experimental validation of results.

MATERIALS AND METHODS

UniReD (UniProt-Related Documents) is a text mining based computational tool able to extract known protein–protein associations/interactions and predict novel ones from information from the scientific biomedical literature. Protein associations can either be physical interactions (e.g. direct binding between proteins) or indirect (e.g. functional associations). The methodology consists of seven main steps (Figure 1) whereas each step is described below:

Step 1: Retrieval of UniProt reviewed records

In this step, all reviewed records for the *Mus musculus* and *Homo sapiens* organisms are retrieved from UniProt knowledge base (35). Notably, reviewed records are of high quality, manually annotated and non-redundant. Non-reviewed records are omitted due to them being less reliable.

Step 2: Extraction of PubMed IDs from UniProt records

In this step, all UniProt records are parsed and all PubMed IDs are extracted from them. Publications which are tagged as high-throughput by UniProt are filtered out, as non-protein-specific. Additionally, publications appearing in more than 10 (empirical threshold) UniProt records are also excluded from the following steps of the pipeline. The purpose of this exclusion step is to retain publications referring to a specific protein and discard the broader ones, e.g. articles discussing the whole proteome of an organism.

Step 3: Retrieval of related documents

In this step, we take advantage of PubMed's functionality called 'similar documents' or previously known as 'related documents'. With this functionality all relevant documents to the ones collected by the previous step (36) are collected and ranked by relevance. The ranking score is used to assign weights to the connections.

Step 4: Graph construction

Each document is represented as a node and each edge represents the relatedness between them. The result is a weighted graph of documents (based on the PubMed related score).

Step 5: Document clustering

In this step, the MCL clustering algorithm (37–39) is applied on the graph in order to generate clusters of documents. By adjusting a single parameter, called *Inflation value*, clusters of different scales of granularity are generated.

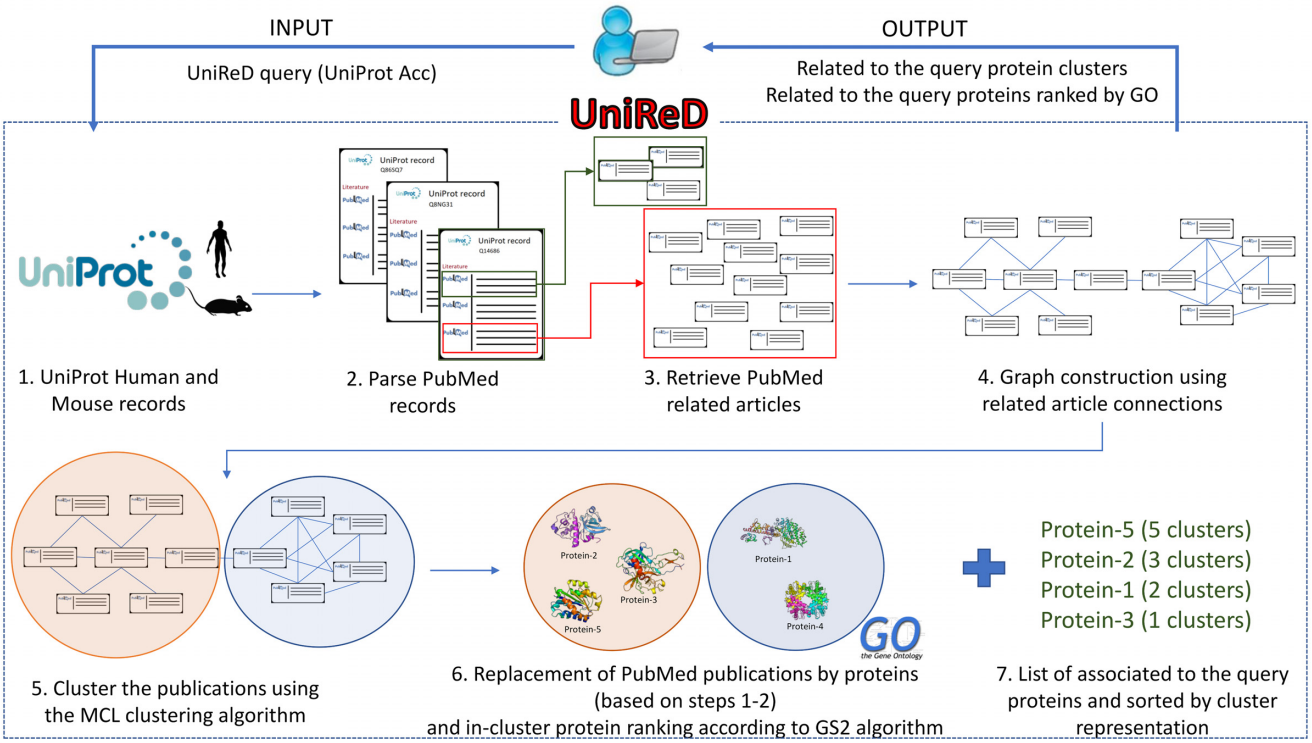

Figure 1. Main steps of the UniReD methodology.

Step 6: Protein clustering

This step involves the creation of protein clusters (UniProt ACs) from the document clusters (PubMed IDs) created in step 5. Each document (PubMed ID) which is related to at least one protein (UniProt AC) based on information from UniProt is replaced by these proteins. The outcome will be clusters of proteins (UniProt ACs) based on the clustering in step 5. We posit that proteins within the same cluster are associated (directly or indirectly) whereas a protein may appear in more than one cluster. Proteins within a cluster are ranked by using the GS2 scoring measure (40). The GS2 measure is a GO-based measure and estimates the similarity between a set of genes/proteins by averaging the contribution of each gene's/protein's GO terms and their ancestor terms with respect to the GO vocabulary graph. Thus, clusters with proteins more relevant to each other have higher GS2 measure scores.

Step 7: List of associated proteins

A non-redundant list of proteins which were found in clusters associated to the query protein is generated. The proteins are sorted by their cluster representation, thus the more clusters a protein is found to belong to, the higher its placement in the list.

Input/Output

The whole pipeline runs in the background every 6 months and all results are updated and pre-calculated. Users are able to query for a protein (UniProt AC) and get back: (i) clusters containing the query of interest and other associated proteins and (ii) a ranked list of proteins from all clusters associated to the query.

Protocols used for experimental validation

As a proof of principle, we experimentally validated six UniReD PPI predictions in three case studies. Further bellow, we provide information about the protocols used for the experiments. The selection of the protein pairs that were chosen for experimental validation (direct or indirect association) was based on manual inspection of UniReD results for the protein of interest. After that, depending on the availability of resources (i.e. antibodies, plasmids), the appropriate experiments were conducted.

Case study 1 (Contactin-2)

Brain cortex from adult (2mo) and embryonic (E15.5) C57BL6/SV129 mice were carefully dissected and homogenized in ice-cold glucopyranoside lysis buffer (85 mM Tris, pH 7.5, 30 mM NaCl, 1 mM ethylenediaminetetraacetic acid, 120 mM glucose, 1% Triton X-100, 60 mM octyl Q-D glucopyranoside (Sigma-Aldrich)), protease inhibitor mixture diluted 1:1000 (Sigma-Aldrich) and phosphatase inhibitors (ThermoFisher Scientific), followed by a brief sonication on ice. Alternatively, lysates from co-transfected HEK293T cells were used. The following expression plasmids were transfected in 10 cm culture plates: human CNTN2 in pEGFP-C1, mouse Sema6A-c-myc epitope in pCX, mouse Nfasc140FLAG in pCMVTag4 and rat Nfasc155FLAG in pCMVTag4 (from Diane Sherman).

Table 1. Sequences of primers for the yeast Gcn5p experiment

ZWF1-S1	GAAAGAGTAAATCCAATAGAATAGAAAACCACATAAGGCAAGcgtacgctgcaggtcgac
ZWF1-R3	ATTTCAGTGACTTAGCCGATAAATGAATGTGCTTGCATTTTTCtcgatgaattcgagctcg
GCN5 (HATΔ) fw	CATCTTTCCATGGCTGTCATTAGGAAGCCATTGACTGTCGTAGGtttttgattccggtttctttg
GCN5 (HATΔ) rev	TTTAATATATCCCATCCATATACTTTTATCCAACGTGATTTCTTTagattcccgggtaataactg

Fifteen micrograms of total DNA were used for each transfection. For co-immunoprecipitation (Co-IP) studies, protein lysates were precleared with protein G Sepharose beads (GE Healthcare LifeSciences) for 1 h at 4°C. Co-IP was performed by incubating the lysate with the antibody overnight at 4°C. The next day, 40 μl of protein G Sepharose beads were added and incubated for at least 1 h at 4°C. The following antibodies were used: rabbit polyclonal antibody against CNTN2 (TG2, (41)) and mouse monoclonal antibody for CNTN2 (1C12, Developmental Studies Hybridoma Bank, (42)). Immunoprecipitates were analyzed on a sodium dodecyl sulphate-polyacrylamide gel of appropriate acrylamide percentage and transferred to a 0.45 μM Protran nitrocellulose transfer membrane (GE Healthcare LifeSciences), over 1 h using a wet transfer unit (Bio-Rad Laboratories). After blocking (5% powdered BSA and 0.1% Tween-20 in phosphate-buffered saline (PBS)) for 1 h, the membrane was incubated overnight at 4°C with the primary antibodies. After washing three times for 15 min in 0.1% Tween-20 in PBS, samples were incubated for 1 h at room temperature with horseradish peroxidase–coupled secondary antibodies, and proteins were visualized by enhanced chemiluminescence (Luminata Classico HRP Substrate, Millipore). The following antibodies were used: mouse monoclonal against Reelin E4 (1:1000, Developmental Studies Hybridoma Bank), mouse monoclonal against MAP1B (1:1000, Santa Cruz Biotechnology), rabbit polyclonal antibody against CNTN2 (1:4000, (41)),rabbit polyclonal against c-Myc (A-14) (1:1000, Santa Cruz Biotechnology, sc-789), Pan-Neurofascin antibody and horseradish peroxidase-coupled secondary antibodies (1:6000; GE Healthcare).

Case study 2 (Necdin)

Lysates from E18 mouse brains were obtained by incubation and sonication of the tissue in RIPA buffer supplemented with mini-complete protease inhibitors (Roche). Immunoprecipitation of Necdin was performed following standard procedures, using the NC243 antibody raised against Necdin amino acids 83–325 (kind gift of Dr Yoshikawa). Immunoprecipitated proteins were separated by western blot and QCR-1 was identified on the blot using a specific antibody (Abcam, ab110252).

Case study 3 (Yeast Gcn5p)

Known interactors of Gcn5p include, among others, Kar2p (43,44), Sod2p (45), Tsa1p (43,46–47), Ssc1p (43), Gsh1p (46), Rpo21p (43) and Cdc28p (48). The parental wild-type yeast strain (FT5, S288c) that was used in this study has been previously described in detail (49). All yeast strains derivatives were cultured in rich YPD media and yeast transformations were performed by standard methods, as described in (50). Yeast transformants were selected by growth in appropriate minimal or antibiotic(s) containing media, then purified and tested for growth defects as described in the text. The zwf1Δ mutant strain was constructed following the standard strategy by using the pair of primers ZWF1-S1 και ZWF1-R3 (Table 1) and the plasmid pYM6 as a template. The resulting polymerase chain reaction (PCR) product was used to transform the wild-type strain as referred above. Disruption of GCN5 was accomplished using the forward GCN5 (HATΔ) and reverse GCN5 (HATΔ) primers along with the plasmid YEp24-GCN5, as previously described (51). In order to confirm the corresponding genes' disruptions, PCR analysis was performed with appropriate internal primers. The zwf1Δ disruption was also confirmed by the Methionine auxotrophy phenotype known to be displayed by the zwf1Δ mutation. The zwf1Δ, gcn5Δ double mutant was constructed by sequential genes' disruption in the FT5 wild-type strain, as described above.

RESULTS

UniReD's functional landscape

UniReD methodology was applied on *M. musculus* and *H. sapiens*. The query used to retrieve the reviewed UniProt records for *M, musculus* for example was: (*reviewed:yes AND organism:'M. musculus (Mouse) [10090]'),* where 10 090 refers to the NCBI taxonomy ID. Similarly, the respective query used to retrieve the reviewed UniProt records for *H. sapiens* was: (*reviewed:yes AND organism:'H. sapiens (Human) [9606]'*). While in UniReD's GUI we allow clustering using five different MCL inflation values (2.0, 2.2, 2.5, 2.7 and 3.0—higher inflation values entail more but tighter clusters), the results reported in this article for analysis and benchmarking have been generated using the inflation value of 2.5.

Briefly, the total number of generated document clusters for mouse were 66 780, containing 13 655 unique proteins. While 27 390 of these clusters were discarded as singletons (clusters with only one member), UniReD reported 13 487 unique proteins appearing in 39 390 clusters.

Similarly, 97 944 clusters containing 16 369 proteins were generated for human. From them, 47 193 clusters were discarded as singletons and 50 751 composed of 16 137 unique proteins were used for further analysis.

The web interface

UniReD's functionality is offered through an online application (Figure 2). The backend routines were written in Perl whereas its front end in HTML/PHP/JavaScript. For navigation and performance reasons, all results are precomputed thus making them immediately available to the

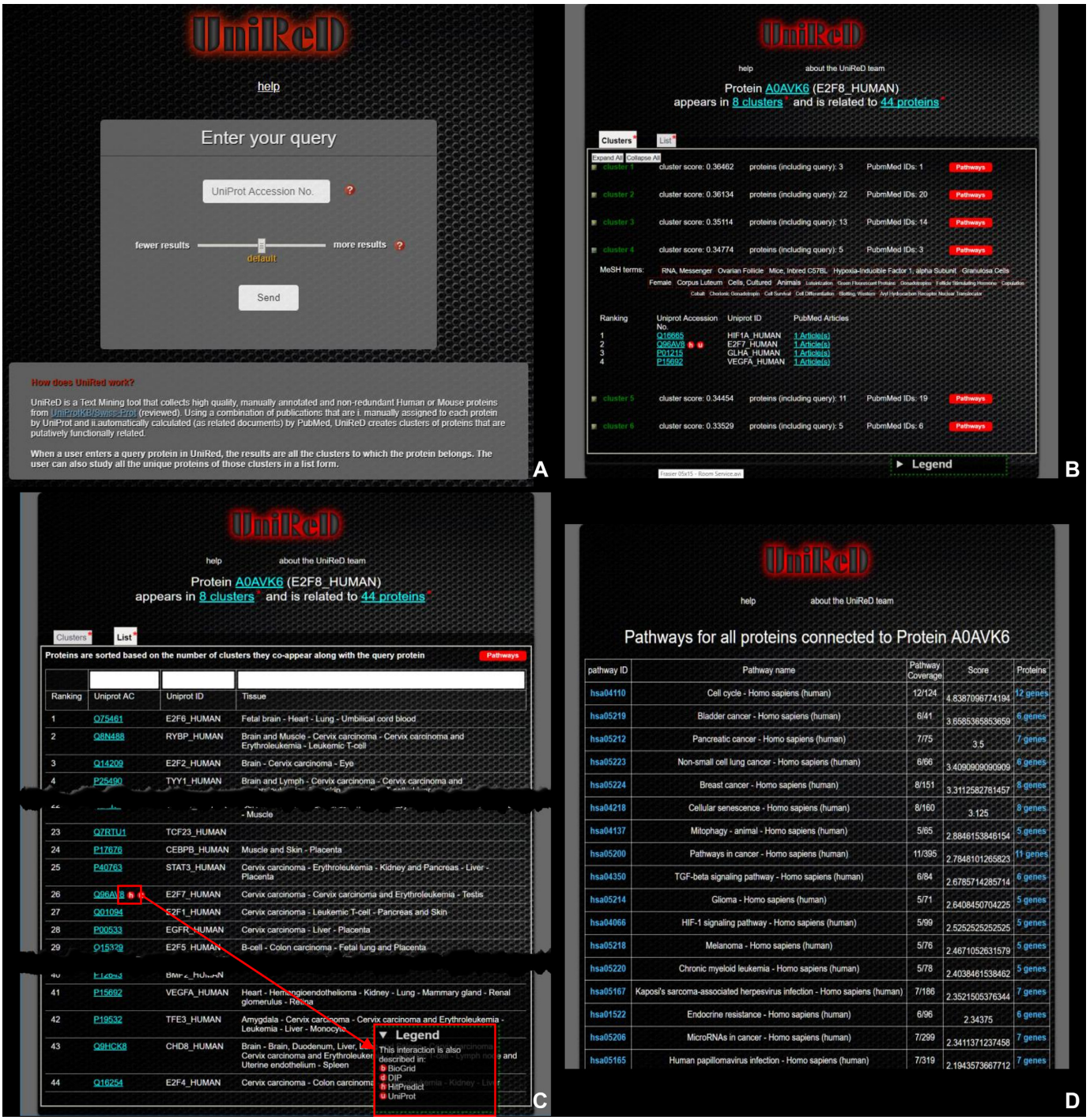

Figure 2. UniReD application. (**A**) UniReD's landing page. The user may enter a mouse or human UniProt accession number and choose the granularity of the results (slidebar -MCL inflation value). (**B**) Cluster View: Each cluster contains the query protein and other associated proteins found in literature. Interactions which are found in external databases (BioGRID, DIP, HitPredict, UniProt) are marked. A protein may appear in more than one cluster. Cluster are ranked according GO similarity. (**C**) List View: All proteins predicted to interact with the query protein, along with information whether the PPIs are described in an external PPI database. Proteins are sorted according to the number of clusters they appear in. (**D**) KEGG pathways related to query protein.

user. Through UniReD's interface, a user can query the system for any protein of interest by using a UniProt accession number. For the time being, only *M. Musculus* and *H. sapiens* accession numbers are acceptable, but more organisms will be supported soon. Using an interactive slider, the user can in addition adjust MCL's inflation value and subsequently the number of clusters (the higher the inflation value, the more and tighter the clusters). Every time a query is performed, all clusters containing the UniProt Accession Number are pulled from our precalculated results and presented to the user in various ways. At the top of the web page, a summary displaying the UniProt Accession Number & the UniProt ID of the UniProt entity is presented. In addition, the number of clusters in which the query protein appears along with the number of predicted protein interactors are shown. Detailed information can be downloaded whereas the results can be displayed in two modes, namely *Cluster View* and *List View*.

In a typical Cluster view, all clusters are displayed and sorted by a cluster score. Next to each cluster, the user can view the cluster score, the number of proteins in the cluster and the number of unique PubMed IDs where this information came from. The clusters are initially collapsed but users can expand one or all of them. Upon expansion, a detailed report for each cluster is generated and accompanied by a list of MeSH terms. These MeSH terms derive from the PubMed publications that relate to each cluster. The results are normalized and the MeSH terms that appear more often are displayed in a word cloud format. The size is analogous to their occurrence frequency. The proteins in each cluster are ranked by their GO similarity using the GS2 algorithm (40). In addition, icons are used to highlight whether a protein in a cluster was found to interact with the query protein in curated PPI databases such as BioGRID (29), DIP (30), HitPredict (31) and UniProt (52). At last, a KEGG representation analysis per cluster is available where users can see whether the proteins in a cluster belong to a KEGG pathway (53). Links to KEGG pathways with highlighted proteins are provided.

In a typical List view all unique proteins found to be associated with the query protein of interest are displayed. The list is non-redundant and proteins are sorted by their cluster representation. This way, if a protein was found in many clusters, it will be sorted higher compared to a protein which was found in fewer clusters. To reduce complexity, by using text filtering, users can sort or filter the list based on various features. One can for example filter by expression in a specific tissue using substrings (e.g *uter* for *uterus*, *uteral* etc.). At last, while all known interactions are marked by an icon (similarly to the Cluster View), tissue expression derived from the SwissProt database (local updated copy) is also displayed for each Uniprot entity.

Experimental validation

As a proof of principle, we experimentally validated six PPI predictions, as suggested by UniReD.

Prediction 1-4. The first case focuses on Contactin-2/TAG-1 (CNTN2, UniProt AC:Q61330), a cell adhesion molecule of the immunoglobulin superfamily (IgSF) that is known to exert its functions through homophilic and/or heterophilic interactions (54,55). UniReD revealed a number of known protein interactors for the target molecule CNTN2, such as L1CAM (56), NRCAM (57), CNTNAP2/CASPR2 (54), potassium voltage-gated channel subfamily A member 1 KCNA1 (58), and protein CD24 (55). Apart from the known molecules that interact with CNTN2, several other putative interactors were identified. Four of them were validated with co-immunoprecipitation experiments (Figure 3). More specifically, interactions were detected between CNTN2 and the cell adhesion molecules Sema6A and Neurofascin isoforms 140 and 155 in HEK293 co-transfected cells (Figure 3A-F). Furthermore, CNTN2 was found to interact with the microtubule-associated protein MAP1B both in embryonic and adult mouse brain tissue (Figure 3F-L). MAP1B promotes similar processes as CNTN2 such as axonal growth, development, branching and regeneration, playing also an important role in axon guidance and neuronal migration (59). At last, an interaction was verified between CNTN2 and Reelin, a large extracellular glycoprotein secreted by several neurons, particularly, in the embryonic cortex, by Cajal–Retzius cells (60). This result points toward a possible crosstalk between Reelin and CNTN2, which may contribute to neuronal migration.

Prediction 5. The second case concerns Necdin (UniProt AC:P25233), a protein belonging to the Melanoma-associated Antigen Gene Family (MAGE) of proteins. Necdin, is encoded by an imprinted gene mapped to human chromosome 15q11-q13, a genetic locus that is invariably deleted in patients with Prader-Willi and Angelman Syndromes (61). As the biological role of Necdin is still not entirely clear, UniReD could provide useful insights as to pathways that may be deregulated in the absence of Necdin, ultimately contributing to the phenotypic manifestations of these diseases. Interestingly, UniReD identified pathways that were already known to intersect with Necdin function. One such example is the insulin pathway. UniReD identified several members of this pathway, including IGF-1, IGF-II, IGF-1R, IR, IRS-1 as potential interactors (62,63). In addition, recent work indicates that Necdin promotes mitochondrial biogenesis in neurons by stabilizing PGC1a (64). Interestingly, UniReD suggested several mitochondrial proteins as potential Necdin interactors. Of interest was a cluster of mitochondrial proteins, that included members of the electron transport chain, such as CoxVIII, CoxVIa1, QCR-1 and UCP2, as well as proteins involved in mitochondrial metabolism such as FABP-1, sterol 26 hydroxylase, Glo2, endophilin-B1 and p19ARF isoform 4. We verified by co-immunoprecipitation that Necdin interacts physically with QCR-1 in lysates of embryonic murine brain, as shown in Figure 4. This finding invites the speculation that Necdin may be directly involved in regulating cell metabolism, a possibility that could account for the metabolic defects associated with Necdin deficiency. However, this hypothesis needs further exploration.

Prediction 6. The last case involves the yeast Gcn5p (UniProt AC: Q03330). Gcn5p is the catalytic histone acetyl transferase subunit of three chromatin modifying

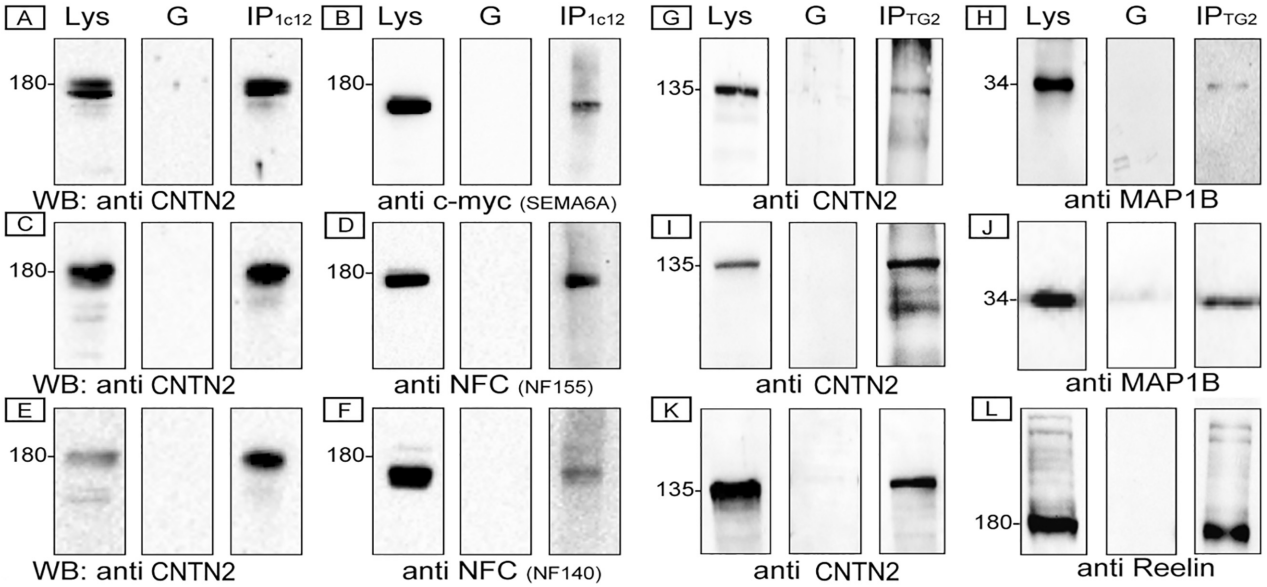

Figure 3. Co-immunoprecipitation analysis of HEK293 co-transfected cells and mouse embryonic and adult brain tissue. (**A**–**F**) Direct interactions of CNTN2 with Sema6A (A and B), Neurofascin155 (NF155) (C and D) and Neurofascin140 (NF140) (E and F) in HEK293 co-transfected cells. Immunoprecipitation was performed with the monoclonal anti CNTN2 antibody 1c12. Western blot analysis of the lysates (Lys), G-beads used for the preclearance step (**G**) and immunoprecipitates (IP1c12) revealed the direct interaction of GFP-tagged CNTN2 with Sema6A-c-myc (B), NF155 (D) and NF140 (F). (G–**L**) Interaction of CNTN2 with MAP1B in mouse adult (G and H) and embryonic tissue (I and J), and Reelin (K and L) in embryonic tissue. Immunoprecipitation was performed with the rabbit polyclonal antibody against CNTN2, TG2. Western blot analysis of the lysates (Lys), G-beads used for the preclearance step (G) and immunoprecipitates (IPTG2) revealed the interaction of CNTN2 with MAP1B (H and J) and Reelin (L).

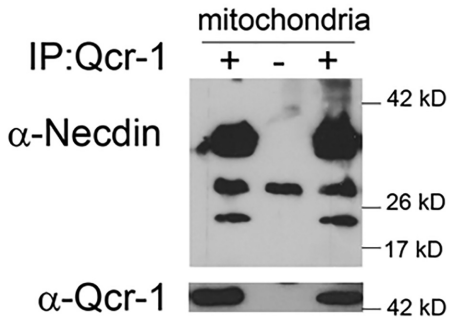

Figure 4. Whole brain lysates from E18 mouse embryos were fractionated to isolate mitochondria. The mitochondrial fraction was then used for co-immunoprecipitation experiments with antibodies against Qcr-1, a complex III mitochondrial protein and Necdin. Necdin physically interacts with Qcr-1, as shown.

complexes (ADA, SAGA and SLIK/SALSA) that target-specific lysine residues of nucleosomal histones (H3 and H2B), regulating transcription of numerous genes (4,65–68). In order to find interactors for yeast Gcn5p, we used the mouse homolog Kat2a (UniProt AC: Q9JHD2) as a query. UniReD revealed several protein interactors for the target molecule in mouse. Those include proteins whose yeast homologs are known to interact with Gcn5p, as well as several putative novel interactors. UniReD also revealed several putative interactors for Gcn5p. Apart from the known molecules that interact with CNTN2, several other putative interactors were identified. For some of them we identified the corresponding homolog proteins in yeast: (i) the 40S ribosomal proteins Rps15p and Rps16ap, as well as the alpha subunit of the Translation Elongation Factor Tef1, (ii) the cytosceleton structural unit actin Act1p, (iii) Kin28p, a subunit of the General Transcription Factor TFIIH, which is an essential serine/threonine-protein kinase that targets RNA polII C-terminal domain and regulates the transcription initiation process (69) and at last (iv) two physiologically related cytoplasmic enzymes, Sod1p and Zwf1p (70). Sod1p is a Cu-Zn superoxide dismutase that scavenges harmful superoxide anions and is required for cell protection from oxidative stress. Zwf1p encodes Glucose-6-phosphate dehydrogenase, the first and rate-limiting enzyme of the pentose phosphate pathway that reduces NADP+ to NADPH, which is also a critical metabolite for cell survival upon oxidative stress. We tested whether Gcn5p interacts functionally with the cytoplasmic Zwf1p (G6PD) protein. To this end, we generated yeast mutants lacking either Zwf1p (zwf1Δ) or Gcn5p (gcn5Δ), as well as double mutants lacking both proteins (zwf1Δ, gcn5Δ). The zwf1Δ mutant strain does not display any growth defect, while the gcn5Δ strain displays a severe slow growth phenotype. Interestingly, the double mutant strain zwf1Δ, gcn5Δ grows normally. We concluded that the zwf1Δ mutation suppresses the growth defect caused by the gcn5Δ mutation, indicating a functional interaction between the two proteins.

Additional evidence. At last, we evaluated UniReD's performance against a standard mass spectrometry-based affinity purification (AP-MS) of mouse PPIs analysis presented by Chatzinikolaou *et al.* (71). In this study, XPF (UniProt AC: Q9QZD4, also known as ERCC4) was used as affinity bait in order to 'fish' putative direct/indirect In-

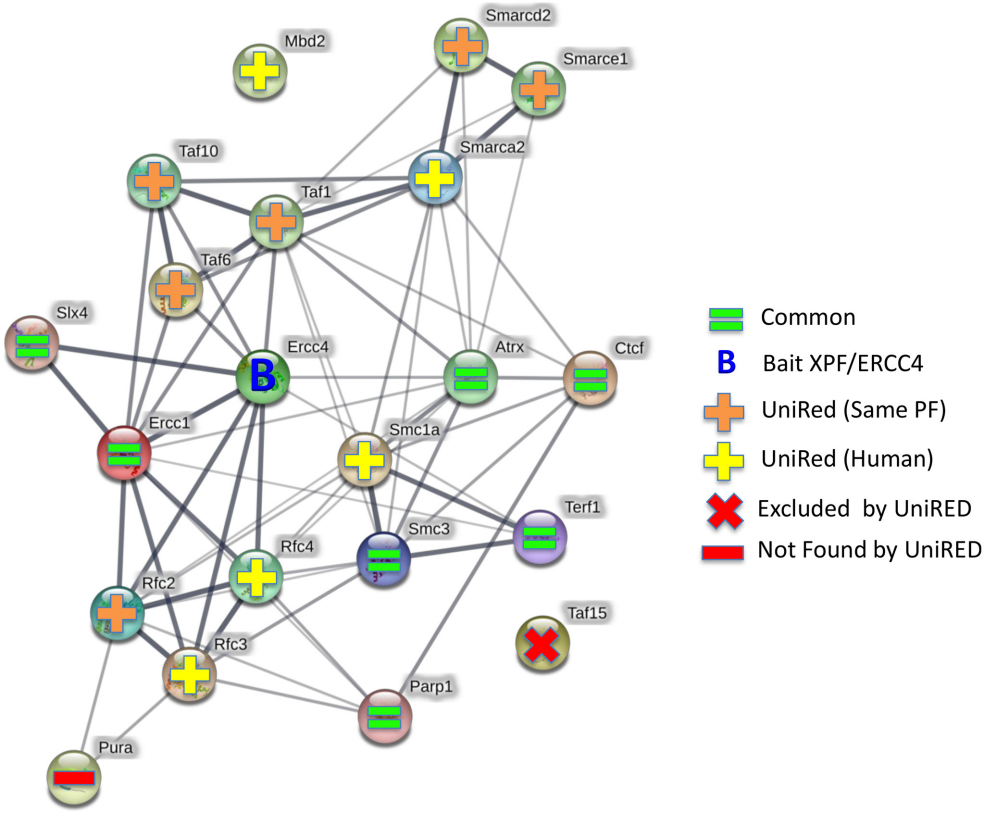

Figure 5. Protein interaction network of ERCC4 (XPF-Q9QZD4) generated by String. This experimental study identified 306 putative PPIs that were reduced to 20 proteins after manual validation. UniReD was able to identify 18 out of 19 (94.5%) proteins selected by the researchers for experimental verification as interacting partners of XPF. One (Taf15-Q8BQ46) was excluded from the UniReD results, as UniReD requires that all proteins are reviewed and excludes large-scale experiment publications. The symbols in the inset box explain how proteins were found in UniReD. PF means protein family and means that UniReD detected proteins of the same family with the identified interacting partner. Human indicates that UniReD detected Human homologs of the interacting partner.

Table 2. UniReD evaluation using different databases

Database (DB)	Organism	Coverage
HitPredict Small Scale	Human	60.06%
HitPredict Small Scale	Mouse	68.06%
BioGRID Small Scale	Human	55.34%
BioGRID Small Scale	Mouse	66.95%
UniProt High Confidence (>0.60)	Human	53.21%
UniProt High Confidence (>0.60)	Mouse	96.83%
DIP Small Scale	Human	77.59%
DIP Small Scale	Mouse	76.62%
Lit-BM	Human	73.48%
PrePPI 201008 dataset	Human	81.57%
Pickle (results with >1 publication)	Human	73.27%

Two datasets were used: i) the Lit-BM dataset, which is a highly curated human interactome network and ii) the PrePPI dataset (the Human High Confidence set—interactions supported by at least two publications prior to August 2010 - https://honiglab.c2b2.columbia.edu/PrePPI/ref/data/human.db.hc.201008.intm). Last column represents the same evaluation against the STRING database, using its medium and high confidence text mining evidence.

teractors. ERCC1 (UniProt AC: P07903) is a known interactor of XPF, and both are essential components of the NER system. Normally, ERCC1 joins XPF endonuclease to form heterodimeric endonuclease (XPF-ERCC1), which excises the 5′ end of DNA to the damaged site. XPF-ERCC1 complex also participates in the homologous recombination and repair of inter-strand crosslinks. This experimental study identified 306 putative PPIs. Furthermore, manual validation of the putative PPIs resulted in four XPF-bound protein complexes consisting of 20 proteins, only four of them being previously confirmed as direct interactors. We queried UniReD with mouse XPF using UniProt data predating the publication of the study. UniReD predictions exhibit a minor overlap with the automatically derived experimental data (∼12%), while a higher overlap with the manually derived short-listed putative PPIs is evident (∼39%). Four of these proteins were previously confirmed as direct interactors (TERF1-P70371, ERCC1-P07903, PARP1-P11103, SLX4-Q6P1D7) and three of them are predicted as putative PPIs (CTCF-Q61164, SMC3-Q9CW03, ATRX-Q61687). A manual inspection of the remaining 11 proteins, showed that 4 of them appear in the UniReD results when human XPF is used as query, and for six proteins UniReD was able to identify one or more proteins belonging in the same protein family. Summing up, UniReD was able to identify 94.5% of the proteins selected by the researchers for experimental verification as interactors of XPF (Figure 5). This again reveals the usefulness of our method when combined with high-throughput experimental results and can be

Table 3. Topological analysis and comparison between UniReD and STRING networks (human and mouse)

	MOUSE							
	Inflation	Nodes	Edges	Centralization	Average #neighbors	Density	Heterogeneity	Clustering coefficient
UniReD	2.0	13 608	7 208 519	0.59	1059	0.078	1.21	0.742
UniReD	2.2	13 565	5 238 233	0.53	772	0.057	1.33	0.724
UniReD	2.5	13 487	3 335 666	0.44	494	0.037	1.48	0.693
UniReD	2.8	13 413	2 ,397 365	0.38	357	0.027	1.59	0.667
UniReD	3.0	13 357	1 982 808	0.36	296	0.022	1.67	0.653
BioGRID		7332	26 060	0.23	6	0.001	5.32	0.130
STRING		21 291	5 972 403	0.27	561	0.026	0.92	0.245
	HUMAN							
	Inflation	Nodes	Edges	Centralization	Average #neighbors	Density	Heterogeneity	Clustering coefficient
UniReD	2.0	16 318	6 711 598	0.56	822	0.050	1.21	0.669
UniReD	2.2	16 253	4 452 816	0.52	547	0.034	1.33	0.643
UniReD	2.5	16 137	2 593 089	0.45	321	0.026	1.52	0.612
UniReD	2.8	16 031	1 759 127	0.39	219	0.014	1.66	0.594
UniReD	3.0	15 952	1 423 064	0.35	178	0.011	1.73	0.584
BioGRID		17 793	475 919	0.16	39	0.002	2.22	0.121
STRING		19 354	5 879 727	0.36	607	0.031	0.87	0.205

a practical guide for researchers in order to identify a list of proteins for further analysis.

Benchmarking against known protein–protein interaction (PPI) DBs

In order to validate UniReD, we benchmarked its coverage against several established PPI databases, such as HitPredict (31), BioGRID (29), UniProt (52), Lit-BM (72), PrePPI (73) and PICKLE (74). All UniReD results were produced using UniProt version 2017 and PubMed's related articles feature from the same period (March 2017). Document clusters were generated after applying MCL inflation value 2.5. For the analysis, we only looked at associations/interactions in which both ends (interactors) are proposed by UniReD. The proposed PPIs were generated based on the assumption that all proteins in a UniReD cluster may be functionally related. In addition, to validate UniReD results against known well-documented PPIs, we only considered the interactions which have high accuracy/confidence from only PPIs from small scale experiments, the interactions with high confidence score when available or the interactions from PPIs which were found in at least two different references in the biomedical literature. Coverage results are shown in Table 2.

Comparison with STRING

Looking at simple topological features such as the number of nodes and edges as well as other features like the centralization, the average number of neighbors, the density, the heterogeneity and the clustering coefficient (Table 3), we observe that UniReD's networks (human and mouse) are very comparable to STRING networks. For an inflation value of 2.5, for example, UniReD generates a protein association network consisting of 13 487 nodes and 3 335 666 edges for mouse and a network consisting of 16 137 nodes and 2 593 089 edges for human. Similarly, STRING (version 10/2019—all evidence channels, 0.4 score), generates a network consisting of 21 291 nodes and 5 972 403 edges for mouse and a network consisting of 19 354 nodes and 5 879 727 edges for human. While networks have similar topologies in terms of density and average neighbor connectivity, UniReD's networks are more modular.

As a next step we compared UniReD's and STRING's (26) coverage against widely-used databases such as HitPredict (31), DIP (30), BioGRID (29), UniProt (52), Lit-BM (72), PrePPI (73) and PICKLE (74). For a fairer comparison, we isolated STRING's human and mouse networks for three different evidence scores (0.9 highest, 0.7 high and 0.4 medium) and filtered by text mining evidence channel. The coverage results are reported in Tables 4 and 5 respectively and show that UniReD achieves a better coverage than STRING when it is compared against the known PPI databases.

In order to demonstrate that the coverage (human and mouse) differences between UniReD and STRING (high score:0.7 and medium score 0.4) are statistically significant, we used *t*-test since the data follow a normal distribution (Shapiro-Wilk $W = 0.91869$, $p = 0.07141$) and the variances between the two datasets do not have a statistically significant difference ($F(10,10) = 1.2089$, $P = 0.77$). Similarly, when comparing UniReD and STRING (medium score: 0.4), we show that the data follow a normal distribution (Shapiro–Wilk $W = 0.95496$, $P = 0.3946$) too and that the variances of the two datasets do not have a statistically significant difference ($F(10,10) = 0.80085$, $P = 0.7322$). The Shapiro–Wilk test shows whether a distribution deviates from a normal distribution. In both cases, the results indicate that there is a statistically significant difference between the coverage of UniReD and high confidence STRING ($t(20) = 9.2169$, $P < 0.001$) as well as between the coverage of UniReD and medium confidence STRING ($t(20) = 4.4965$, $P < 0.001$), thus highlighting the importance of UniReD's existence.

Retrospective analysis

We performed a retrospective analysis (as conducted elsewhere (25)), where we used articles published until 2010 in order to predict non-documented PPIs and crosschecked how many of the UniReD predictions were indeed pub-

Table 4. Comparison of the STRING database PPI predictions based on text mining evidence against the set of **human** PPI databases UniReD was compared for evaluation purposes

	Confidence		
STRING DB	Medium (0.4)	High (0.7)	Highest (0.9)
biogrid_human	31.34% (18264/58259)	11.99% (6991/58259)	2.74% (1600/58259)
preppi_human	63.38% (4118/6497)	36.44% (2368/6497)	11.95% (777/6497)
lit_bm_human	46.7% (4626/9904)	22.49% (2228/9904)	7.4% (733/9904)
hitpredict_human	32.83% (20149/61362)	12.87% (7898/61362)	3.07% (1886/61362)
dip_human	58.49% (2583/4416)	33.67% (1487/4416)	11.91% (526/4416)
uniprot_human	31.93% (2080/6513)	18.68% (1217/6513)	7.04% (459/6513)
pickle_human	47.46% (10007/21083)	23.62% (4980/21083)	6.59% (1390/21083)

Medium, high and highest confidence imply a score higher than 0.4, 0.7 and 0.9, respectively. The percentage corresponds to the coverage, the numbers inside the parenthesis are the number of PPIs common in both DBs and the total PPIs of each DB.

Table 5. Comparison of the STRING database PPI predictions based on text mining evidence against the set of mouse PPI databases UniReD was compared for evaluation purposes

	Confidence	
STRING DB	Medium (0.4)	High (0.7)
uniprot_mouse	71.81% (158/220)	48.63% (107/220)
biogrid_mouse	35.66% (2363/6625)	15.87% (1052/6625)
dip_mouse	48.7% (581/1193)	25.56% (305/1193)
hitpredict_mouse	34.35% (3635/10580)	15.19% (1608/10580)

Medium, high and highest confidence imply a score higher than 0.4, 0.7 and 0.9, respectively. The percentage corresponds to the coverage, the numbers inside the parenthesis are the number of PPIs common in both DBs and the total PPIs of each DB.

lished from 2011–2015; we were able to predict 57.1% of the published (and experimentally verified) PPIs, a result that reveals the high potential of UniReD. At last, in order to compare our method against a well-established and very popular PPI prediction tool, we compared STRING (26) against the eleven databases used for evaluating UniReD. As shown in Table 2, UniReD achieved competitive coverage scores when compared to the scores achieved by STRING.

Random clustering

The information produced by UniReD clusters was further validated by randomizing the assignment of UniProt proteins to the clusters (random clustering). Using the bootstrap method (75), we created 1000 random clusterings (samples) for *H. sapiens*, where the number of clusters and their size remain the same as in the UniReD results. The UniProt proteins are randomly assigned to each cluster. Then the coverage is calculated for each of the random clusterings to known PPI databases shown in Table 2. Using the results for each of the 1000 bootstrap samples, a distribution graph based on kernel density estimation was plotted. If the coverage calculated for the UniReD clustering is not overlapping with the respective coverage calculated for the random clusterings, UniReD results can be considered informative (i.e. not random). The results are depicted in Supplementary Figure S1. It is obvious that the computed coverages do not overlap with the respective coverage calculated for the random clusterings.

Validation against KEGG

To ascertain KEGG's coverage percentage, all *H. sapiens* and *M. musculus KEGG* pathways were collected. UniReD results were mapped against the KEGG pathways. Only the KEGG ids (i.e. proteins in the KEGG pathways) that could be translated to UniProt ACs were used following the same reasoning applied when comparing to PPI databases. In the case of *H. sapiens*, the coverage reaches a percentage of 98.02% and in the case of *M. musculus* the coverage is 96.94%. Similarly, to the method followed previously, 1000 random clusterings were created. Supplementary Figure S2 presents the distribution graph of the coverage based on the kernel density estimation. The maximum coverage is 58%, which is lower than the one computed when the PPIs are not random.

Comparison with Negatome 2.0

It is not trivial to assess UniReD's false positive prediction rate because many PPIs have not been discovered and described yet. In addition, UniReD is an exploratory tool and does not only capture direct interactions but also indirect ones (associations). In an effort to assess UniReD's false positive rates, we compared its performance against Negatome 2.0 (76). Negatome is a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. In order to be as thorough as possible, we used the largest, most extensive dataset provided by Negatome (combined dataset, 6532 protein pairs). Out of the 1526 human protein pairs in Negatome, 756 were also detected by UniReD. The respective numbers in mice are 239 out of 411. In order to account for indirect interactions and more generic associations, we employed STRING and filtered out the UniReD pairs which were found in Negatome but not in STRING (with at least medium confidence). Here, we make the assumption that interactions in the intersection between STRING and Negatome are indirect interactions. For human, out of the 756 Negatome pairs found by UniReD, 618 were also detected by STRING. These pairs might be involved in indirect interactions and should not be considered false positives, thus leaving only 138 pairs to be considered as false positives. Similarly, for mouse, out of the 239 Negatome pairs found by UniReD, 187 are also detected by STRING, thus leaving only 52 pairs to be considered as possible false positives. The false positive rate was based using

Table 6. Comparison with Negatome 2.0

Organism	Negatome pairs (combined)	UniReD pairs	UniReD pairs found in Negatome	UniReD pairs in Negatome but not in STRING	False positive ratio
Human	1526	2593090	756	138	15.1%
Mouse	411	3335667	239	52	23.2%

FPR = (UN-UNS)/(N-UNS)
UN = Common pairs between UniReD & Negatome
UNS = Common pairs between UniReD & Negatome & STRING
N = Negatome pairs
This analysis resulted in a false positive rate of 15.1% for humans and 23.2% for mice.

the following formula: *(UN-UNS)/(N-UNS)* (where *UN* = Common pairs between UniReD & Negatome, *UNS* = Common pairs between UniReD & Negatome & STRING and *N* = Negatome pairs). This analysis resulted in a false positive rate of 15.1% for human and 23.2% for mouse. Results are reported in Table 6.

DISCUSSION

UniReD is a tool that can assist wet lab scientists in their quest to discover proteins associated with a protein of interest and a tool for accurately extracting known interactions and pathways. It mainly parses UniProt records for literature links and enriches this dataset with related articles from PubMed. Then upon abstract clustering, it substitutes the articles with original UniProt records and reports many associated-to-a-query proteins in the form of ranked clusters or as a ranked list. UniReD's performance has been validated both computationally and experimentally and benchmark tests show a high coverage rate and an impressive predictive percentage, indicators for its effectiveness and accuracy. Experimental validations which have been conducted by different wet lab researchers, revealed five new direct interactions (co-immunoprecipitation) and a genetic one. For broader use, UniReD comes with a user-friendly GUI where users can simply start by entering a UniProt AC of the protein of interest as input and get back a list of functionally related proteins and pathways. At last, and most importantly, UniRed can be safely used as a tool for exploratory analyses, as well as for prioritizing hits obtained from complementary high throughput studies and can aid researchers to perform more targeted experiments. This methodology may be able to unravel potentially functional counterparts of the protein of interest or unreported pathways where a protein is involved in.

DATA AVAILABILITY

UniReD is available at: http://bioinformatics.med.uoc.gr/unired/

Both the curated PPI databases as well as information in SwissProt and PubMed are scheduled for update every 6 months.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors would also like to thank Dr Christos Ouzounis, Dr Vassilis Promponas, Dr Nikos Darzentas and Dr Pantelis Hatzis for the fruitful discussions and comments. We also thank several experimental biologists from the University of Crete, Massachusetts Institute of Technology and Icahn School of Medicine at Mount Sinai (NY) for testing UniReD and providing us with valuable feedback.

FUNDING

Financed by the State Scholarship Foundation (IKY), funded by the Action "Scholarships for post-graduate studies" (Operational Program "Education and Lifelong learning", 2014–2020) and project "BIOIMAGING-GR" (MIS5002755), which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation"(NSRF 2014-2020). GAP was supported by the Operational Program "Competitiveness, Entrepreneurship and Innovation", NSRF 2014-2020, Action code: MIS 5002562, co-financed by Greece and the European Union (European regional Development Fund).
Conflict of interest statement. None declared.

REFERENCES

1. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 4569–4574.
2. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions inSaccharomyces cerevisiae. *Nature*, **403**, 623–627.
3. Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A., Bertone,P., Lan,N., Jansen,R., Bidlingmaier,S., Houfek,T. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.
4. Lee,T.I. and Young,R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, **34**, 77–137.
5. Free,R.B., Hazelwood,L.A. and Sibley,D.R. (2009) Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. *Curr. Protoc. Neurosci.*, **Chapter 5**, Unit 5.28.
6. Pitre,S., Alamgir,M., Green,J.R., Dumontier,M., Dehne,F. and Golshani,A. (2008) Computational methods for predicting protein-protein interactions. *Adv. Biochem. Eng. Biotechnol.*, **110**, 247–267.
7. Aloy,P. and Russell,R.B. (2002) The third dimension for protein interactions and complexes. *Trends Biochem. Sci.*, **27**, 633–638.
8. Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

9. Qi,Y., Bar-Joseph,Z. and Klein-Seetharaman,J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins Struct. Funct. Bioinform.*, **63**, 490–500.
10. Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 5896–5901.
11. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 4285–4288.
12. Promponas,V.J., Ouzounis,C.A. and Iliopoulos,I. (2014) Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey. *Brief. Bioinform.*, **15**, 443–454.
13. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
14. Swanson,D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, **30**, 7–18.
15. Swanson,D.R. (1990) Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.*, **78**, 29–37.
16. Krallinger,M., Valencia,A. and Hirschman,L. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9(Suppl 2)**, S8.
17. Papanikolaou,N., Pavlopoulos,G.A., Theodosiou,T. and Iliopoulos,I. (2015) Protein-protein interaction predictions using text mining methods. *Methods*, **74**, 47–53.
18. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.
19. Alako,B.T.F., Veldhoven,A., van Baal,S., Jelier,R., Verhoeven,S., Rullmann,T., Polman,J. and Jenster,G. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6**, 51.
20. Alanis-Lobato,G., Andrade-Navarro,M.A. and Schaefer,M.H. (2017) HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.*, **45**, D408–D414.
21. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
22. Ono,T., Hishigaki,H., Tanigami,A. and Takagi,T. (2001) Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.
23. Novichkova,S., Egorov,S. and Daraselia,N. (2003) MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, **19**, 1699–1706.
24. Rebholz-Schuhmann,D., Jimeno-Yepes,A., Arregui,M. and Kirsch,H. (2010) Measuring prediction capacity of individual verbs for the identification of protein interactions. *J. Biomed. Inform.*, **43**, 200–207.
25. van Haagen,H.H.H.B.M., 't Hoen,P.A.C., Botelho Bovo,A., de Morrée,A., van Mulligen,E.M., Chichester,C., Kors,J.A., den Dunnen,J.T., van Ommen,G.-J.B., van der Maarel,S.M. *et al.* (2009) Novel Protein-Protein interactions inferred from literature context. *PLoS One*, **4**, e7894.
26. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
27. van Haagen,H.H.H.B.M., 't Hoen,P.A.C., de Morrée,A., van Roon-Mom,W.M.C., Peters,D.J.M., Roos,M., Mons,B., van Ommen,G.-J. and Schuemie,M.J. (2011) In silico discovery and experimental validation of new protein-protein interactions. *Proteomics*, **11**, 843–853.
28. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
29. Chatr-aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
30. Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: The database of interacting Proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.
31. López,Y., Nakai,K. and Patil,A. (2015) HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species. *Database*, **2015**, doi:10.1093/database/bav117.
32. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
33. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
34. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
35. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
36. Lin,J. and Wilbur,W.J. (2007) PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, **8**, 423.
37. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
38. Azad,A., Pavlopoulos,G.A., Ouzounis,C.A., Kyrpides,N.C. and Buluç,A. (2018) HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Res.*, **46**, e33.
39. Theodosiou,T., Darzentas,N., Angelis,L. and Ouzounis,C.A. (2008) PuReD-MCL: a graph-based PubMed document clustering methodology. *Bioinformatics*, **24**, 1935–1941.
40. Ruths,T., Ruths,D. and Nakhleh,L. (2009) GS2: an efficiently computable measure of GO-based similarity of gene sets. *Bioinformatics*, **25**, 1178–1184.
41. Traka,M., Dupree,J.L., Popko,B. and Karagogeos,D. (2002) The neuronal adhesion protein TAG-1 is expressed by Schwann cells and oligodendrocytes and is localized to the juxtaparanodal region of myelinated fibers. *J. Neurosci.*, **22**, 3016–3024.
42. Savvaki,M., Tivodar,S., Theodorakis,K., Stylianopoulou,F., Stamatakis,A., Karagogeos,D., Zoupi,L. and Kyriacou,K. (2010) The expression of TAG-1 in glial cells is sufficient for the formation of the juxtaparanodal complex and the phenotypic rescue of Tag-1 homozygous mutants in the CNS. *J. Neurosci.*, **30**, 13943–13954.
43. Lee,K.K., Sardiu,M.E., Swanson,S.K., Gilmore,J.M., Torok,M., Grant,P.A., Florens,L., Workman,J.L. and Washburn,M.P. (2014) Combinatorial depletion analysis to assemble the network architecture of the SAGA and ADA chromatin remodeling complexes. *Mol. Syst. Biol.*, **7**, 503–503.
44. Vembar,S.S., Jonikas,M.C., Hendershot,L.M., Weissman,J.S. and Brodsky,J.L. (2010) J domain co-chaperone specificity defines the role of BiP during protein translocation. *J. Biol. Chem.*, **285**, 22484–22494.
45. Graumann,J., Dunipace,L.A., Seol,J.H., McDonald,W.H., Yates,J.R., Wold,B.J. and Deshaies,R.J. (2004) Applicability of tandem affinity purification MudPIT to pathway proteomics in yeast. *Mol. Cell. Proteomics*, **3**, 226–237.
46. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L.Y., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
47. Pan,X., Ye,P., Yuan,D.S., Wang,X., Bader,J.S. and Boeke,J.D. (2006) A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell*, **124**, 1069–1081.
48. Chymkowitch,P., Eldholm,V., Lorenz,S., Zimmermann,C., Lindvall,J.M., Bjørås,M., Meza-Zepeda,L.A. and Enserink,J.M. (2012) Cdc28 kinase activity regulates the basal transcription machinery at a subset of genes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 10450–10455.
49. Tzamarias,D. and Struhl,K. (1994) Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. *Nature*, **369**, 758–761.
50. Sheff,M.A. and Thorn,K.S. (2004) Optimized cassettes for fluorescent protein tagging inSaccharomyces cerevisiae. *Yeast*, **21**, 661–670.

51. Topalidou,I., Papamichos-Chronakis,M., Thireos,G. and Tzamarias,D. (2004) Spt3 and Mot1 cooperate in nucleosome remodeling independently of TBP recruitment. *EMBO J.*, **23**, 1943–1948.
52. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
53. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
54. Traka,M., Goutebroze,L., Denisenko,N., Bessa,M., Nifli,A., Havaki,S., Iwakura,Y., Fukamauchi,F., Watanabe,K., Soliven,B. *et al.* (2003) Association of TAG-1 with Caspr2 is essential for the molecular organization of juxtaparanodal regions of myelinated fibers. *J. Cell Biol.*, **162**, 1161–1172.
55. Lieberoth,A., Splittstoesser,F., Katagihallimath,N., Jakovcevski,I., Loers,G., Ranscht,B., Karagogeos,D., Schachner,M. and Kleene,R. (2009) Lewis(x) and alpha2,3-sialyl glycans and their receptors TAG-1, Contactin, and L1 mediate CD24-dependent neurite outgrowth. *J. Neurosci.*, **29**, 6677–6690.
56. Kuhn,T.B., Stoeckli,E.T., Condrau,M.A., Rathjen,F.G. and Sonderegger,P. (1991) Neurite outgrowth on immobilized axonin-1 is mediated by a heterophilic interaction with L1(G4). *J. Cell Biol.*, **115**, 1113–1126.
57. Pavlou,O., Theodorakis,K., Falk,J., Kutsche,M., Schachner,M., Faivre-Sarrailh,C. and Karagogeos,D. (2002) Analysis of interactions of the adhesion molecule TAG-1 and its domains with other immunoglobulin superfamily members. *Mol. Cell. Neurosci.*, **20**, 367–381.
58. Poliak,S. and Peles,E. (2003) The local differentiation of myelinated axons at nodes of Ranvier. *Nat. Rev. Neurosci.*, **4**, 968–980.
59. Sato-Yoshitake,R., Shiomura,Y., Miyasaka,H. and Hirokawa,N. (1989) Microtubule-associated protein 1B: molecular structure, localization, and phosphorylation-dependent expression in developing neurons. *Neuron*, **3**, 229–238.
60. Ogawa,M., Miyata,T., Nakajima,K., Yagyu,K., Seike,M., Ikenaka,K., Yamamoto,H. and Mikoshiba,K. (1995) The reeler gene-associated antigen on Cajal-Retzius neurons is a crucial molecule for laminar organization of cortical neurons. *Neuron*, **14**, 899–912.
61. Jay,P., Rougeulle,C., Massacrier,A., Moncla,A., Mattei,M.G., Malzac,P., Roëckel,N., Taviaux,S., Lefranc,J.L.B., Cau,P. *et al.* (1997) The human NECDIN gene, NDN, is maternally imprinted and located in the Prader-Willi syndrome chromosomal region. *Nat. Genet.*, **17**, 357–361.
62. Cypess,A.M., Zhang,H., Schulz,T.J., Huang,T.L., Espinoza,D.O., Kristiansen,K., Unterman,T.G. and Tseng,Y.-H. (2011) Insulin/IGF-I regulation of necdin and brown adipocyte differentiation via CREB- and FoxO1-associated pathways. *Endocrinology*, **152**, 3680–3689.
63. Tseng,Y.-H., Butte,A.J., Kokkotou,E., Yechoor,V.K., Taniguchi,C.M., Kriauciunas,K.M., Cypess,A.M., Niinobe,M., Yoshikawa,K., Patti,M.E. *et al.* (2005) Prediction of preadipocyte differentiation by gene expression reveals role of insulin receptor substrates and necdin. *Nat. Cell Biol.*, **7**, 601–611.
64. Hasegawa,K., Yasuda,T., Shiraishi,C., Fujiwara,K., Przedborski,S., Mochizuki,H. and Yoshikawa,K. (2016) Promotion of mitochondrial biogenesis by necdin protects neurons against mitochondrial insults. *Nat. Commun.*, **7**, 10943.
65. Georgakopoulos,T. and Thireos,G. (1992) Two distinct yeast transcriptional activators require the function of the GCN5 protein to promote normal levels of transcription. *EMBO J.*, **11**, 4145–4152.
66. Brownell,J.E., Zhou,J., Ranalli,T., Kobayashi,R., Edmondson,D.G., Roth,S.Y. and Allis,C.D. (1996) Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, **84**, 843–851.
67. Grant,P.A., Duggan,L., Côté,J., Roberts,S.M., Brownell,J.E., Candau,R., Ohba,R., Owen-Hughes,T., Allis,C.D., Winston,F. *et al.* (1997) Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: Characterization of an ada complex and the saga (spt/ada) complex. *Genes Dev.*, **11**, 1640–1650.
68. Pray-Grant,M.G., Schieltz,D., McMahon,S.J., Wood,J.M., Kennedy,E.L., Cook,R.G., Workman,J.L., Yates,J.R. and Grant,P.A. (2002) The novel SLIK histone acetyltransferase complex functions in the yeast retrograde response pathway. *Mol. Cell. Biol.*, **22**, 8774–8786.
69. Rodriguez,C.R., Cho,E.-J., Keogh,M.-C., Moore,C.L., Greenleaf,A.L. and Buratowski,S. (2000) Kin28, the TFIIH-Associated Carboxy-Terminal domain kinase, facilitates the recruitment of mRNA processing machinery to RNA polymerase II. *Mol. Cell. Biol.*, **20**, 104–112.
70. Slekar,K.H., Kosman,D.J. and Culotta,V.C. (1996) The yeast copper/zinc superoxide dismutase and the pentose phosphate pathway play overlapping roles in oxidative stress protection. *J. Biol. Chem.*, **271**, 28831–28836.
71. Chatzinikolaou,G., Apostolou,Z., Aid-Pavlidis,T., Ioannidou,A., Karakasilioti,I., Papadopoulos,G.L., Aivaliotis,M., Tsekrekou,M., Strouboulis,J., Kosteas,T. *et al.* (2017) ERCC1–XPF cooperates with CTCF and cohesin to facilitate the developmental silencing of imprinted genes. *Nat. Cell Biol.*, **19**, 421–432.
72. Rolland,T., Taşan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A Proteome-Scale map of the human interactome network. *Cell*, **159**, 1212–1226.
73. Zhang,Q.C., Petrey,D., Garzón,J.I., Deng,L. and Honig,B. (2013) PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.*, **41**, D828–D833.
74. Gioutlakis,A., Klapa,M.I. and Moschonas,N.K. (2017) PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology. *PLoS One*, **12**, e0186039.
75. Efron,B. and Tibshirani,R.J. (1994) *An Introduction to the Bootstrap*. Chapman & Hall.
76. Blohm,P., Frishman,G., Smialowski,P., Goebels,F., Wachinger,B., Ruepp,A. and Frishman,D. (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.*, **42**, D396–D400.