Published online 27 October 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D485–D493***
doi: 10.1093/nar/gku1007

ComPPI: a cellular compartment-specific database for protein–protein interaction network analysis

Daniel V. Veres[1,†], Dávid M. Gyurkó[1,†], Benedek Thaler[1,2], Kristóf Z. Szalay[1], Dávid Fazekas[3], Tamás Korcsmáros[3,4,5] and Peter Csermely[1,*]

[1]Department of Medical Chemistry, Semmelweis University, Budapest, Hungary, [2]Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary, [3]Department of Genetics, Eötvös Loránd University, Budapest, Hungary, [4]TGAC, The Genome Analysis Centre, Norwich, UK and [5]Gut Health and Food Safety Programme, Institute of Food Research, Norwich, UK

Received August 13, 2014; Revised September 22, 2014; Accepted October 08, 2014

ABSTRACT

Here we present ComPPI, a cellular compartment-specific database of proteins and their interactions enabling an extensive, compartmentalized protein–protein interaction network analysis (URL: http://ComPPI.LinkGroup.hu). ComPPI enables the user to filter biologically unlikely interactions, where the two interacting proteins have no common subcellular localizations and to predict novel properties, such as compartment-specific biological functions. ComPPI is an integrated database covering four species (*S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*). The compilation of nine protein–protein interaction and eight subcellular localization data sets had four curation steps including a manually built, comprehensive hierarchical structure of >1600 subcellular localizations. ComPPI provides confidence scores for protein subcellular localizations and protein–protein interactions. ComPPI has user-friendly search options for individual proteins giving their subcellular localization, their interactions and the likelihood of their interactions considering the subcellular localization of their interacting partners. Download options of search results, whole-proteomes, organelle-specific interactomes and subcellular localization data are available on its website. Due to its novel features, ComPPI is useful for the analysis of experimental results in biochemistry and molecular biology, as well as for proteome-wide studies in bioinformatics and network science helping cellular biology, medicine and drug design.

INTRODUCTION

Biological processes are separated in the cellular and subcellular space, which helps their precise regulation. Compartmentalization of signalling pathways is a key regulator of several main biochemical processes, such as the nuclear translocation-mediated activation of transcription factors (1). Several proteins are located in more than one subcellular localizations. As an example, IGFBP-2 is a predominantly extracellular protein with a key role in insulin growth factor signalling (2), while its translocation into the nucleus results in vascular endothelial growth factor-mediated angiogenesis (3). Another important example is the HIF-1 Alpha with translocation from the cytosol to the nucleus, where it acts as a transcription factor involved in the maintenance of cellular oxygen homeostasis (4) (Supplementary Figure S1). Their shuttling between these localizations is a key regulatory mechanism, which implicates the importance of improving the systems level analysis of compartmentalized biological processes.

Protein–protein interaction data are one of the most valuable sources for proteome-wide analysis (5), especially to understand human diseases on the systems-level (6) and to help network-related drug design (7). However, protein–protein interaction databases often contain data with low overlap (8), and are designed using different protocols (9), therefore, their integration is needed to improve our comprehensive knowledge (10). Low-throughput data sets often use several different protein naming conventions causing difficulties in data analysis and integration. Manual curation of data yields a large improvement of data quality (11).

Interaction data often contain interactions, where the two interacting proteins have no common subcellular localizations (12). These interactions could be biophysically possible, but biologically unlikely (13). Thus, these interactions cause data bias that leads to deteriorated reliability in interactome-based studies (14), especially those involv-

[*]To whom correspondence should be addressed. Tel: +361 459 1500; Fax: +361 266 3802; Email: csermely.peter@med.semmelweis-univ.hu
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

ing subcellular localization-specific cellular processes (15). Unfortunately, subcellular localization data are incomplete. Despite the need of experimentally verified subcellular localizations for reliable compartmentalization-based interactome filtering (16), only computationally predicted subcellular localization information is available for a large part of the proteome. Moreover, subcellular localization data are redundant, often poorly structured and miss to highlight the reliability of data (17).

Existing analysis tools involving subcellular localizations offer the download of filtered interactomes for a subset of proteins (like MatrixDB (18)). Several databases use only Gene Ontology (GO (19)) cellular component terms as the source of the subcellular localization data (such as HitPredict (20) or Cytoscape BiNGO plugin (21)), while GO still contains data inconsistency despite its highly structured annotations (22). Cytoscape Cerebral plugin (23) generates a view of the interactome separated into layers according to their subcellular localization. In different data sets the subcellular localization structure is not uniform, which makes their comparisons often difficult.

ComPPI-based interactomes introduced here provide a broader coverage (Supplementary Tables S1 and S2), using several curation steps in data integration. ComPPI offers highly structured subcellular localization data supplemented with Localization and Interaction confidence Scores, all presented with user-friendly options. As a key feature ComPPI allows the construction of high-confidence data sets, where potentially biologically unlikely interactions in which the interacting partners are not localized in the same cellular compartment, have been deleted. As our examples will show, this gives novel options of interactome analysis and also suggests potentially new subcellular localizations and localization-based functions.

DESCRIPTION OF THE DATABASE

Overview of ComPPI

Our goal by constructing ComPPI was to provide a reliable subcellular compartment-based protein–protein interaction database for the analysis of biological processes on the subcellular level. A key feature of ComPPI is that it allows the filtering of localization-based biologically unlikely interactions resulting in localization-wise more reliable interaction data. During the integration of 17 databases to build up ComPPI, we used the following four curation steps to improve data quality (Figure 1). (i) Source databases were selected by comparing them to a large number of other potential databases and their data content was manually reviewed. (ii) Subcellular localization data were consistently structured to a hierarchical subcellular localization tree (Supplementary Figure S2) containing more than 1600 individual sublocalizations. (iii) We developed an algorithm to map different protein naming conventions to UniProt accession numbers (24,25). (iv) Finally, a manual follow-up by six independent experts was performed in order to revise the data content searching for data inconsistence and false entries, and to test the functions of the web interface (Supplementary Table S3).

ComPPI database includes comprehensive and integrated data of four species (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*) cataloguing 125 757 proteins, their 791 059 interactions and 195 815 major subcellular localizations in its current, 1.1 version. The proteome-wide data set contains localizations for five main subcellular organelles (nucleus, mitochondrion, cytosol, secretory-pathway, membrane) and the extracellular compartment. Importantly, 60% of the ComPPI entries have high resolution cellular localization data assigning them to one or several of >1600 GO cellular

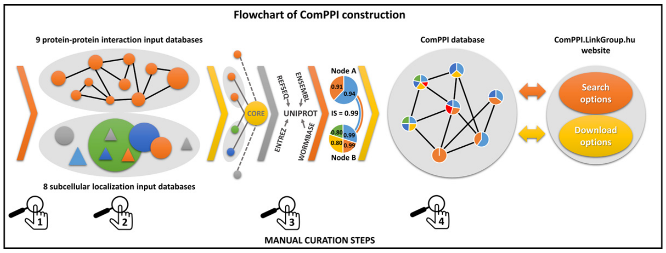


Figure 1. Flowchart of ComPPI construction highlighting the four curation steps. Constructing the ComPPI database we first checked the data content of 24 possible input databases for false entries, data inconsistence and compatible data structure in order to minimize the bias in ComPPI coming from the input sources **(1)**. As a consequence we selected nine protein–protein interaction (BioGRID (29), CCSB (30), DiP (31), DroID (26), HPRD (27), IntAct (32), MatrixDB (18), MINT (33) and MIPS (28)) and eight subcellular localization databases (eSLDB (37), GO (19), Human Proteinpedia (34), LOCATE (38), MatrixDB (18), OrganelleDB (39), PA-GOSUB (36) and The Human Protein Atlas (35)) in order to integrate them into the ComPPI data set. The subcellular localization structure was manually annotated creating a hierarchic, non-redundant subcellular localization tree using >1600 GO cellular component terms (19) for the standardization of the different data resolution and naming conventions **(2)**. All input databases were connected to the ComPPI core database with newly built interfaces in order to improve data consistency, to allow easy extensibility with new databases and to incorporate automatic database updates. As part of the curation steps the filtering efficiency of our newly built interfaces were tested on 200 random proteins for every input databases, and the interfaces were accepted only when all the requested false-entries and data content errors were filtered, in order to establish a more reliable content (Supplementary Table S3). During data integration, different protein naming conventions were mapped to the most reliable protein name. In this process we used publicly available mapping tables (UniProt (24) and HPRD (27)). For 30% of protein names we applied manually built mapping tables with the help of online ID cross-reference services (PICR (25) and Synergizer (http://llama.mshri.on.ca/synergizer/translate/)) **(3)**. After data integration Localization and Interaction Scores were calculated (for detailed description see Figure 2). As an illustration we show the example of Figure 2 with two interacting proteins (nodes A and B corresponding to HSP 90-alpha A2 and Survivin, respectively) with shared cytosolic and nuclear localizations (light blue and orange). Node B has an additional membrane (yellow) subcellular localization and an extracellular localization (green). Numbers in the circles of nodes A and B refer to their Localization Scores. The Interaction Score of nodes A and B is 0.99 (see Figure 2 for details). The integrated ComPPI data set was manually revised by six independent experts **(4)**. During the revision two of the six experts tested our database on 200 random proteins each to ensure high-quality control requirements, and searched for exact matches between the entries in the input sources and the ComPPI data set. All the experts searched for false entries, data inconsistency, protein name mapping errors in the downloadable data and tested the operation of the online services as well. After the revision we updated our source databases, their interfaces, the subcellular localization tree and the algorithm generating the downloadable data, in order to acquire all the changes proposed during the tests. As the final result, the webpage http://ComPPI.LinkGroup.hu is available for search and download options in order to extract the biological information in a user-friendly way.

component terms (19) associating these proteins with distinct subcellular compartments.

Design and implementation

Both protein–protein interaction and subcellular localization data are incorporated to ComPPI automatically using their own interface to bridge the difference in data structure (Supplementary Table S3). New interfaces can be added without limitations. The incoming data are merged to form a consistent internal data pool using a comprehensive protein name mapping algorithm, in order to deal with the redundancy in the input data sets (http://comppi.linkgroup.hu/help/naming_conventions). The website and the downloadable contents are generated from this integrated internal data pool. All curated parts are stored in separate, yet interconnected containers to maintain persistency between ComPPI releases.

The website follows the hierarchical model-view-controller design pattern to ensure the separation of the data layer from the business logic and the user interface. Each functional unit is implemented as a module to further support easy maintenance and extensibility. Protein search algorithms have been extensively optimized, and the served content is cached to ensure quick response times even on low-end infrastructure. Due to these features ComPPI can be easily run on a general laptop or desktop computer.

The downloadable data sets are pre-generated and validated automatically and manually in every release to fulfill our high quality control requirements (Figure 1). The Python script that generates these data sets also contains basic tools for data retrieval and manipulation in a network-oriented manner, which enables the user to perform bioinformatics analysis on the interactome using the open source code and also gives space for further improvement.

End-user documentation is available at the website as tutorials, detailed descriptions and location-specific tooltips. All components of ComPPI and the underlying software stack are open source. The source code is available in a revision controlled repository at http://bificomp2.sote.hu:22422/comppi/summary.

Third-party tools and technologies were selected with open accessibility and scientific reproducibility in mind including the Ubuntu Linux 14.04 operating system (http://ubuntu.com/), the nginx HTTP server (http://nginx.org/), the MySQL 5 Community Edition database server (http://www.mysql.com/), the git version control system (http://git-scm.com/), the PHP 5 scripting language (https://php.net/), the Symfony 2 PHP framework (http://symfony.com/), the jQuery JavaScript framework (http://jquery.com/), the D3.js JavaScript library for network visualization (http://d3js.org/) and the Python3 scripting language (https://python.org/).

Database content and access

Input databases. The low overlap of protein–protein interaction and subcellular localization databases (11) prompted us to integrate several source databases in order to improve data coverage and quality (Supplementary Figure S3 and Supplementary Table S2). In this process we used publicly downloadable license-free data sources, preferably containing proteome-wide data sets. Protein–protein interaction data were selected to contain only physical interactions with experimental evidence coming from high-throughput, as well as low-throughput techniques. We incorporated the widely used species-specific (DroID (26), HPRD (27), MatrixDB (18) and MIPS (28)) and general (BioGRID (29), CCSB (30), DiP (31), IntAct (32) and MINT (33)) protein–protein interaction databases having high data quality, update frequency and freely downloadable latest releases for academic research.

Subcellular localization data can be obtained from experimental evidence or using predictions. Several source databases contained only experimentally verified subcellular localization entries (such as Human Proteinpedia (34) and the Human Protein Atlas (HPA) (35)). Other source data had only computationally predicted information (such as PA-GOSUB (36)). Several data sources had integrated data structure (such as eSLDB (37), GO (19), LOCATE (38), MatrixDB (18), OrganelleDB (39)) containing data of both experimental and predicted origin. During the selection of the proteome-wide predicted subcellular localization databases with downloadable content we focused on the use of prediction algorithms with combined methods using robust machine learning tools validated on highly reliable training sets.

ComPPI data set. The availability of the data sources differs between various species. As an example ComPPI contains eight protein–protein interaction and eight subcellular localization databases for human proteins (Supplementary Figure S3). Database integration was based on protein ID mapping to the most reliable naming convention available, primarily to UniProt Swiss-Prot accession numbers (11). The 4 curation steps (Figure 1) allow the users to access interaction and localization data at a single resource having a higher coverage and reliability than the incorporated databases.

The ComPPI database contains three types of predefined data sets: (i) the compartmentalized interactome catalogues of those protein–protein interactions, where the interacting proteins have at least one common subcellular localization, (ii) the integrated protein–protein interaction data set which can be customized by the four species included and (iii) the subcellular localization data set, which is one of the biggest existing subcellular localization resource with a comprehensive structure for interactome analysis. All downloadable ComPPI resources are license free and publicly available for academic and industrial research.

Search and download features

Search features. The internally hyper-linked web application of ComPPI enables even those users, who have no bioinformatics expertise, to search for the interactions of individual proteins. Search options (http://comppi.linkgroup.hu/protein_search) are available for protein names with autocomplete function giving their subcellular localization, their interactions and the likelihood of their interactions considering the subcellular localization of the interacting partners. Using the Advanced Settings of the Search page the user is able to filter the list of the possible query proteins for species, subcellular localizations and/or localization probability. These settings can be set for the interactors of the query protein too and are adjustable with the Custom Settings on the Results page, which allows the filtering of the interactors for subcellular localizations, localization probability and interaction score. The properties of the query protein and its interactors are available for download. After filtration only those interactions are exported that fulfill the custom filtering requirements set by the user. Network visualization of the whole or filtered first-neighbour interactome of the query protein is also available, where the width of the edges corresponds to the Interaction Score of the given interaction. These options together provide a user-friendly web interface for data mining for both non-experts and computational biologists. A Direct Search option is also available via URL, which gives the opportunity to interconnect the ComPPI database with other resources, or to generate multiple searches for data mining.

Download options. All ComPPI data are available for download at the website. Predefined data sets can be customized by the user to contain only data for a requested species or localization: (i) Compartmentalized interactomes have interactions, where the two interacting protein-nodes have at least one common subcellular localization. These interactomes can be filtered to species besides subcellular localizations. (ii) Integrated protein–protein interaction data sets contain all the interactions, and can be customized to the four species included. (iii) Integrated subcellular localization data sets contain proteins together with their localization data. The user can select species and localizations to customize these data sets. (iv) The current and previous releases of the full database can also be downloaded. A detailed help and a tutorial for the Search and Download functions are both available.

Output. ComPPI output data provide lists of interactions, interaction scores of the interacting proteins and localizations with localization scores. Moreover, the user receives the PubMed IDs and references of the source databases for both the interactions and subcellular localizations, and the additional information (if available) of the data type. The user-defined interactomes as results of the Basic or Advanced Search options and the predefined data sets on the Downloads page are available for download in plain text format to ensure convenient data handling. The complete current and previous releases of the database are downloadable in SQL format to provide full access to all the data in ComPPI.

Localization and interaction scores

Subcellular localization structure. Subcellular localization data are coming from different source databases, containing localizations having experimental evidence (in the followings: experimental), coming from unknown sources (unknown) or predictions (predicted; Figure 2). Experimental data usually have high resolution, where the exact localization of the protein is often defined, such as the nuclear pore complex for Nup107 (40). Predicted localizations have usually low resolution. As an example nuclear localization can be predicted from the existence of a nuclear localization signal in the amino acid sequence (41) without any experimental evidence.

Because of the incongruity in the resolution of the localization data and the different naming conventions between the source databases, we standardized the subcellular localization data using GO cellular component terms (19). In order to solve the problem of the unequivocally mapped GO terms (Figure 3) we created a manually built, non-redundant, hierarchical localization tree (Supplementary Figure S2). With the help of this we clustered the >1600 GO cellular component terms to six major compartments (cytosol, nucleus, mitochondrion, secretory-pathway, membrane, extracellular) (Supplementary Table S4). This new structure allows ComPPI to store all localization entries from different sources and to assign the proteins efficiently to six major compartments (Figure 3 and Supplementary Figure S4).

Localization and interaction scores. The ComPPI Localization Score is a novel measure to score the probability of a localization for a given protein. The Localization Score depends on the subcellular localization evidence type (experimental, unknown, predicted) and the number of sources (Figure 2). The Interaction Score characterizes the probability of the subcellular localization of a protein–protein interaction, and is based on the consensus of the compartment-specific Localization Scores of the interacting proteins. With the help of the scoring algorithm ComPPI provides a novel localization probability describing how likely it is that the protein exists in the given subcellular compartment, and gives the opportunity to build high-confidence interactomes based on the distribution of the interaction scores (Supplementary Figure S5).

Localization Scores are calculated using probabilistic disjunction (marked with operator V) among the different localization evidence types and the number of ComPPI localization data entries of the respective evidence type (Equation (1), see top panel of Figure 2 for details)

$$\varphi_{\mathrm{LocX}} = V_{\mathrm{res}} p_{\mathrm{LocX}} \tag{1}$$

where φ_{LocX} and p_{LocX} are the Localization Score and the localization evidence type (experimental, unknown or predicted) for protein X and localization Loc, respectively, while res is the number of available ComPPI localization data entries for protein X.

As the first step of Interaction Score calculation, compartment-specific Interaction Scores are obtained by multiplying the Localization Scores of the two interactors for each of the six major compartments. Finally, the Interaction Score is calculated as the probabilistic disjunction (marked with operator V) of the Compartment-specific Interaction Scores of all major localizations available for the interacting pair from the maximal number of six major localizations (Equation (2), see bottom panel of Figure 2 for details)

$$\varphi_{\mathrm{Int}} = V_{i=1}^{6} \varphi_{\mathrm{LocA}} * \varphi_{\mathrm{LocB}} \tag{2}$$

where φ_{Int} is the Interaction Score, while φ_{LocA} and φ_{LocB} are the Compartment-specific Localization Scores of interacting proteins A and B, respectively.

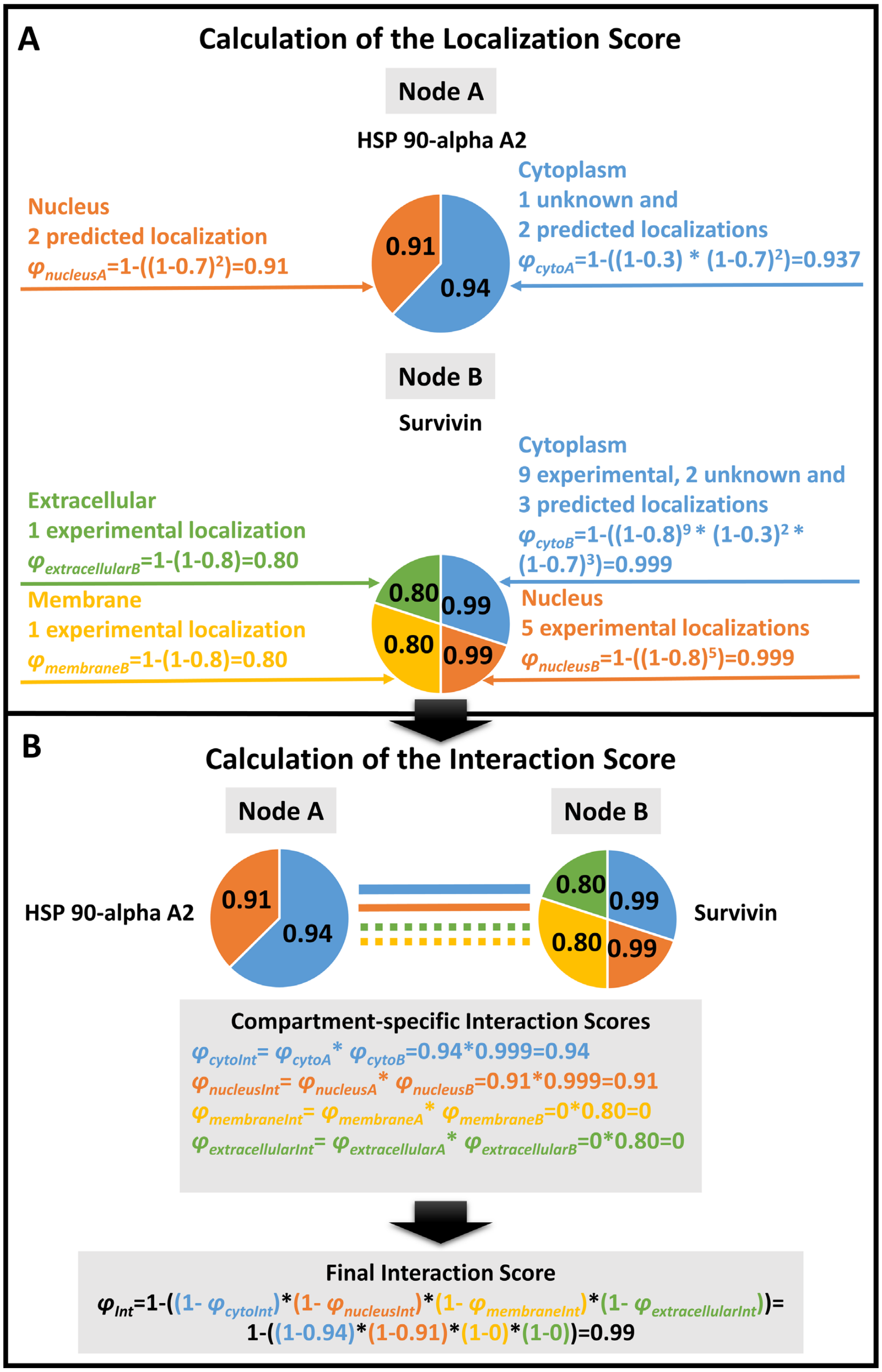

A
Calculation of the Localization Score
Node A
HSP 90-alpha A2
Nucleus
2 predicted localization
$\varphi_{nucleusA}$=1-((1-0.7)^2)=0.91
Cytoplasm
1 unknown and
2 predicted localizations
φ_{cytoA}=1-((1-0.3) * (1-0.7)^2)=0.937
0.91
0.94
Node B
Survivin
Extracellular
1 experimental localization
$\varphi_{extracellularB}$=1-(1-0.8)=0.80
Membrane
1 experimental localization
$\varphi_{membraneB}$=1-(1-0.8)=0.80
Cytoplasm
9 experimental, 2 unknown and
3 predicted localizations
φ_{cytoB}=1-((1-0.8)^9 * (1-0.3)^2 *
(1-0.7)^3)=0.999
Nucleus
5 experimental localizations
$\varphi_{nucleusB}$=1-((1-0.8)^5)=0.999
0.80
0.99
0.80
0.99
B
Calculation of the Interaction Score
Node A
Node B
HSP 90-alpha A2
0.91
0.94
0.80
0.99
0.80
0.99
Survivin
Compartment-specific Interaction Scores
$\varphi_{cytoInt}$= φ_{cytoA}* φ_{cytoB}=0.94*0.999=0.94
$\varphi_{nucleusInt}$= $\varphi_{nucleusA}$* $\varphi_{nucleusB}$=0.91*0.999=0.91
$\varphi_{membraneInt}$= $\varphi_{membraneA}$* $\varphi_{membraneB}$=0*0.80=0
$\varphi_{extracellularInt}$= $\varphi_{extracellularA}$* $\varphi_{extracellularB}$=0*0.80=0
Final Interaction Score
φ_{Int}=1-((1- $\varphi_{cytoInt}$)*(1- $\varphi_{nucleusInt}$)*(1- $\varphi_{membraneInt}$)*(1- $\varphi_{extracellularInt}$))=
1-((1-0.94)*(1-0.91)*(1-0)*(1-0))=0.99

Figure 2. Calculation of the subcellular localization-based ComPPI scores. We illustrate the Localization Score calculation steps on the examples of Heat Shock Protein (HSP) 90-apha A2 and Survivin. HSP 90-alpha A2 has two major subcellular localizations, while Survivin has four ($\varphi_{\mathrm{nucleusA}}$, φ_{cytoA} and $\varphi_{\mathrm{extracellularB}}$, $\varphi_{\mathrm{membraneB}}$, $\varphi_{\mathrm{nucleusB}}$, φ_{cytoB}, respectively). Localizations were manually categorized into major localizations before the calculation (see the text in section 'Subcellular Localization Structure' for details). **(A)** A Localization Score (such as φ_{cytoA}) is calculated for every available major subcellular localization for both HSP 90-alpha A2 and Survivin based on the available localization evidence types and the number of the respective localization data entries (corresponding to p_{LocX} and V_{rec} of Equation (1)). The Localization Score calculation uses the optimized localization evidence type weights of 0.8, 0.7 and 0.3 for experimental, predicted or unknown localization evidence types, respectively. (For details of the weight optimization procedure see section 'Score Optimization' of the main text and Supplementary Figure S6.) The Localization Score (i.e. the likelihood for the respective protein to belong to a major compartment) is represented by the probabilistic disjunction among the different localization evidence types and the number of ComPPI localization data entries of the respective evidence type (Equation (1)). **(B)** Calculation of the Interaction Score (φ_{Int}) is based on the Localization Scores of the interacting proteins. First, Compartment-specific Interaction Scores (such as $\varphi_{\mathrm{cytoInt}}$) are calculated as pair-wise products of the relevant Localization Scores of the two interacting proteins (HSP 90-alpha A2 and Survivin). The final Interaction Score (φ_{Int}) is calculated as the probabilistic disjunction of the Compartment-specific Interaction Scores of all major localizations available for the interacting pair of proteins (in the example four major localizations for HSP 90-alpha A2 and Survivin) from the maximal number of six major localizations (Equation (2)).

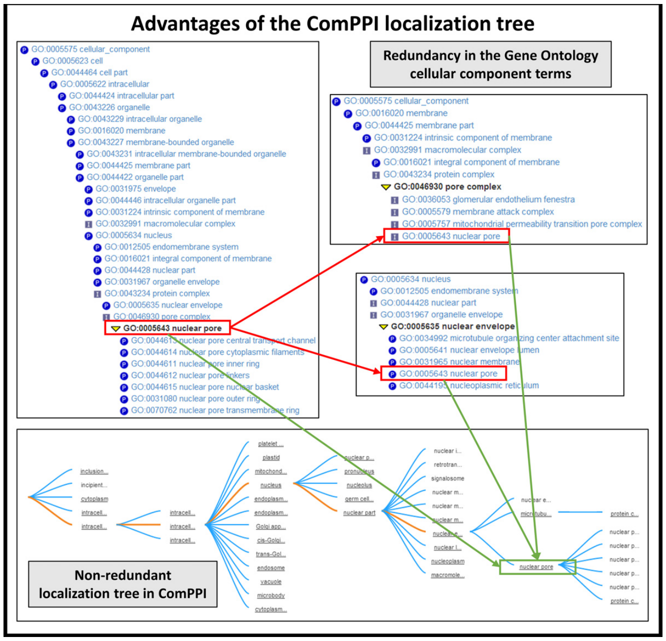

Advantages of the ComPPI localization tree
GO:0005575 cellular_component
GO:0005623 cell
GO:0044464 cell part
GO:0005622 intracellular
GO:0044424 intracellular part
GO:0043226 organelle
GO:0043229 intracellular organelle
GO:0016020 membrane
GO:0043227 membrane-bounded organelle
GO:0043231 intracellular membrane-bounded organelle
GO:0044425 membrane part
GO:0044422 organelle part
GO:0031975 envelope
GO:0044446 intracellular organelle part
GO:0031224 intrinsic component of membrane
GO:0032991 macromolecular complex
GO:0005634 nucleus
GO:0012505 endomembrane system
GO:0016021 integral component of membrane
GO:0044428 nuclear part
GO:0031967 organelle envelope
GO:0043234 protein complex
GO:0005635 nuclear envelope
GO:0046930 pore complex
GO:0005643 nuclear pore
GO:0044613 nuclear pore central transport channel
GO:0044614 nuclear pore cytoplasmic filaments
GO:0044611 nuclear pore inner ring
GO:0044612 nuclear pore linkers
GO:0044615 nuclear pore nuclear basket
GO:0031080 nuclear pore outer ring
GO:0070762 nuclear pore transmembrane ring
Redundancy in the Gene Ontology cellular component terms
GO:0005575 cellular_component
GO:0016020 membrane
GO:0044425 membrane part
GO:0031224 intrinsic component of membrane
GO:0032991 macromolecular complex
GO:0016021 integral component of membrane
GO:0043234 protein complex
GO:0046930 pore complex
GO:0036053 glomerular endothelium fenestra
GO:0005579 membrane attack complex
GO:0005757 mitochondrial permeability transition pore complex
GO:0005643 nuclear pore
GO:0005634 nucleus
GO:0012505 endomembrane system
GO:0044428 nuclear part
GO:0031967 organelle envelope
GO:0005635 nuclear envelope
GO:0034992 microtubule organizing center attachment site
GO:0005641 nuclear envelope lumen
GO:0031965 nuclear membrane
GO:0005643 nuclear pore
GO:0044195 nucleoplasmic reticulum
inclusion...
incipient...
cytoplasm
intracell...
intracell...
intracell...
intracell...
intracell...
platelet ...
plastid
mitochond...
nucleus
endoplasm...
endoplasm...
Golgi app...
cis-Golgi...
trans-Gol...
endosome
vacuole
microbody
cytoplasm...
nuclear p...
pronucleus
nucleolus
germ cell...
nuclear part
nuclear i...
retrotran...
signalosome
nuclear m...
nuclear m...
nuclear m...
nuclear e...
nuclear l...
nucleoplasm
macromole...
nuclear e...
microtubu...
protein c...
nuclear pore
nuclear p...
nuclear p...
nuclear p...
nuclear p...
nuclear p...
protein c...
Non-redundant localization tree in ComPPI

Figure 3. Advantages of ComPPI subcellular localization structure. The subcellular localization structure of ComPPI is based on a manually curated, non-redundant subcellular localization tree extracted from GO data (19) containing more than 1600 GO cellular component terms (Supplementary Figure S2). On Figure 3 an example of the redundancy in the GO cellular component tree structure is shown, where the 'nuclear pore' cellular component can be found under several branches in the tree, such as in the 'nucleus' -> 'nuclear envelope' -> 'nuclear pore' or the 'membrane' -> 'membrane part' -> 'intrinsic component of the membrane' -> 'integral component of the membrane' -> 'pore complex' pathways (highlighted in red). Because of the need of the mapping of high-resolution subcellular localization data into major cellular components (Supplementary Table S4) a localization tree with a non-redundant structure was built. In our example, it can be seen that with the help of this structure the 'nuclear pore' derives unequivocally from the 'nuclear envelope' term (highlighted in green).

Score optimization. As mentioned before the ComPPI localization evidence type can be experimental, unknown or predicted. ComPPI characterizes each of these localization evidence types by a parameter called the evidence type weight to achieve a unified scoring system applicable to the diverse data sources. To obtain these evidence type weights we performed their data-driven optimization. Based on the fact that experimentally validated entries are the most reliable, while localization entries coming from unknown or predicted origin are less reliable, we set the following order of evidence type weights: experimental > predicted AND experimental > unknown as the two requirements of the optimization process. We chose the HPA database (35) containing only experimentally verified subcellular localizations in order to build a positive control data set, where the interactors have at least one common localization according to HPA. Our goal was to find a specific ratio of the experimental, unknown and predicted evidence type weights that maximizes the number of high confidence interactions in the positive control data set (HPA) and simultaneously maximizes the number of low confidence interactions in the ComPPI data set not containing HPA data. These ensure that the quality of data marked as high confidence will have a good match to the quality of experimentally verified data.

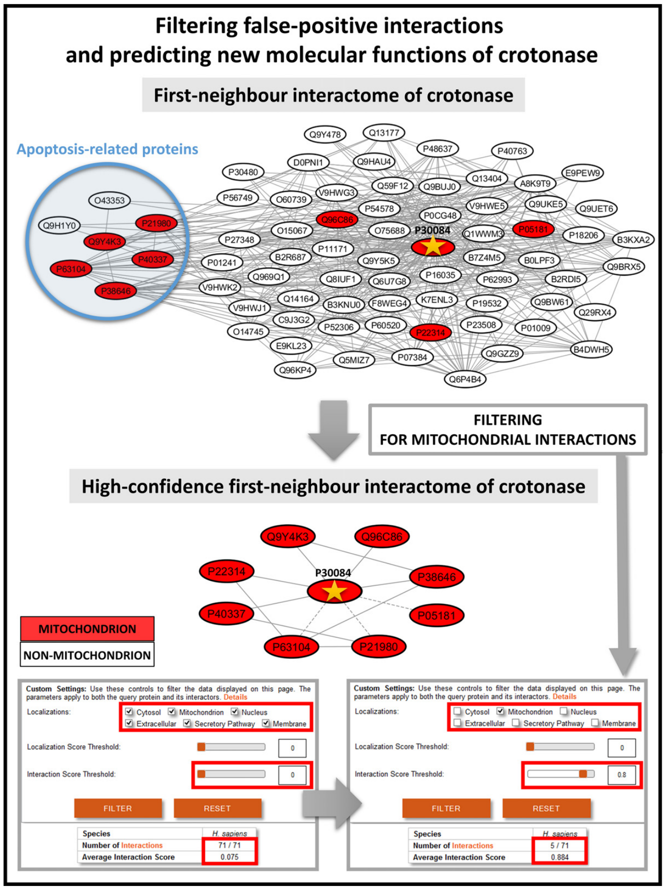
All combinations of the experimental, unknown and predicted evidence type weights were set up from 0 to 1 with 0.1 increments. The kernel density of the interactions were calculated with all these settings (with a bandwidth of 0.01), which gave us the ratio of interactions belonging to a given confidence level compared to the distribution of all the interactions. Finally, the 285 possible kernel density solutions were tested to find the parameter combination that maximizes the number of both the low and high confidence interactions as described above. This resulted in 0.8, 0.7 and 0.3 as the relative evidence type weights for experimental, predicted and unknown data types, respectively (Supplementary Figure S6). Note that this optimization is driven by the reliability of the subcellular localization data, and was not tested using gold standard protein–protein interaction data sets, therefore the Interaction Score reflects the reliability of the interaction in a subcellular localization-dependent but not in an interactome-dependent manner.

Application examples

Merging of subcellular localization and interactome data provides several application opportunities: (i) the filtration of localization-based biologically unlikely interactions—where the two interacting proteins have no common localization and (ii) the prediction of possible new localizations and localization-based biological functions (15). Both are important features of ComPPI as illustrated by an example in this section.

ComPPI-based interaction filtering. First, we made a systematic search for an example, which highlights the importance of the removal of localization-based biologically unlikely interactions looking for key hubs and bridges, where interaction structure changed the most after the filtering step. Here we calculated the degree distribution of the whole human interactome and the high-confidence interactome (containing 23 265/19 386 proteins and their 385 481/260 829 interactions, respectively) where from the latter biologically unlikely interactions with no common subcellular localizations have already been removed. We also calculated the distribution of the betweenness centrality in the two data sets. After these procedures we manually reviewed the first 20 proteins from the UniProt Swiss-Prot subset (15 258 proteins out of 19 386) with the highest differences in degree and centrality measures (Supplementary Table S5). Enoyl-CoA hydratase (crotonase) had the largest absolute change of degree among the top 20 proteins, thus we selected crotonase as our illustrative example (Figure 4). Crotonase catalyses the second step in the beta-oxidation pathway of fatty acid metabolism (42), and is a key member of the crotonase protein superfamily (43). Beta-oxidation takes place primarily in the mitochondrion (44). Crotonase has only a mitochondrial ComPPI localization with experimental evidence, which is in agreement with its cellular function.

Crotonase has 71 interacting partners in the integrated data set, of which only 8 is present in the mitochondrion, and only 5 have an interaction score equal or higher than 0.8. After the manual review of crotonase neighbours, it turned out that only one of the 8 mitochondrial interactors (mitochondrial Hsp70, (45)) has experimental evidence for



Custom Settings: Use these controls to filter the data displayed on this page. The parameters apply to both the query protein and its interactors. Details

Localizations: ☑ Cytosol ☑ Mitochondrion ☑ Nucleus ☑ Extracellular ☑ Secretory Pathway ☑ Membrane

Localization Score Threshold: 0

Interaction Score Threshold: 0

Species	*H. sapiens*
Number of Interactions	71 / 71
Average Interaction Score	0.075

Custom Settings: Use these controls to filter the data displayed on this page. The parameters apply to both the query protein and its interactors. Details

Localizations: □ Cytosol ☑ Mitochondrion □ Nucleus □ Extracellular □ Secretory Pathway □ Membrane

Localization Score Threshold: 0

Interaction Score Threshold: 0.8

Species	*H. sapiens*
Number of Interactions	5 / 71
Average Interaction Score	0.884

Figure 4. Advantages of the ComPPI data set to filter biologically unlikely interactions and to predict compartment-specific, new properties and functions. The figure shows the interactions of crotonase (enoyl-CoA hydratase, UniProt ID: P30084), involved in fatty acid catabolism having a mitochondrial localization, and its first neighbours supported with experimental evidence before and after filtering to mitochondrial localization. Interactions with an Interaction Score below 0.80 are shown with dashed lines. On one hand, out of the original 71 neighbours of crotonase only 8 remain as mitochondrial interacting partners with a significantly higher average Interaction Score than the whole first-neighbour network, which highlights the importance of compartment-specific filtering in the detection of high-confidence interactors in a subcellular localization-dependent manner. On the other hand, the blue circle of the upper left side of the figure shows those cytosolic crotonase interacting partners, which are involved in apoptosis, a recently discovered function of crotonase (45–47). Thus, the very same example also reveals a potential new function of crotonase, which partially involves its unexpected cytosolic localization, which was recently verified experimentally (46).

mitochondrial localization. Mitochondrial localization of the other 7 interactors is not based on strong evidence, while 63 out of 71 interactors have no known mitochondrial localization at all. Figure 4 shows the interactome of crotonase and its 71 first neighbours containing 428 edges. In the mitochondrial interaction subset only 13 edges remained, while the high-confidence part contains only 10 interactions (Figure 4). Second neighbours of crotonase contain 82% of the interactome, and their network contains 14 803 nodes and 319 305 edges. The filtered mitochondrial network of the second neighbours is much smaller, having only 2107 nodes and 8381 interactions.

ComPPI-based prediction of new or non-conventional functions. Importantly, 52 out of the 71 interactors, and more specifically, 7 out of the 8 mitochondrial interacting partners of crotonase have cytosolic localization with a localization probability over 0.95. This indicates that crotonase may have a cytosolic localization as well. Indeed, crotonase was shown to be overexpressed and localized in the cytosol in hepatocarcinoma cells, where it contributes to lymphatic metastatis (46). GO (19) biological process term enrichment analysis of the mitochondrial crotonase interacting partners using BiNGO (21) revealed that besides the known function of the crotonase in ‘catabolic process’ the ‘negative regulation of apoptosis’ and related terms were also significantly enriched (Supplementary Table S6). In agreement with this, previous studies showed that crotonase is overexpressed in several cancer types (47), and the knockdown of crotonase decreased cell viability and enhanced cisplatin-induced apoptosis in hepatocellular carcinoma (48). The anti-apoptotic effect of crotonase also exists in breast cancer, where its down-regulation potentiates PP2-induced apoptosis (49).

These findings may implicate that the high ratio of ‘biologically unlikely’ interactions may also be a result of a transient and dynamic cytosolic subcellular localization of crotonase, where the enzyme may be involved in currently not widely crotonase-associated biological processes, such as the inhibition of apoptosis. Importantly, these compartment-specific crotonase functions may be applied as potential therapeutic targets in the treatment of hepatocellular carcinoma or breast cancer.

In summary, the crotonase example shows the utility of ComPPI both (i) to filter low-confidence interactions concentrating on high-confidence subcellular localizations and (ii) to predict unknown biological functions in previously unknown or non-conventional subcellular localizations. Another example of ComPPI-based prediction of potential, novel functions besides crotonase, is Monopolar Spindle 1 protein (MPS1) having a centromere-associated cytosolic localization (50). We identified a number of relatively undiscovered MPS1 functions related to the ComPPI analysis of nuclear MPS1 interactome as detailed in Supplementary Figure S7 and Supplementary Table S6.

CONCLUSIONS AND FUTURE DIRECTIONS

In summary, ComPPI provides a unique data set for the analysis of protein–protein interaction networks at the subcellular level. The assembly of the integrated ComPPI database with manual curation protocols (Figure 1) provides an improvement of both coverage and data quality. ComPPI subcellular localization data have a novel structure in order to incorporate localizations from different data sources (Figure 3 and Supplementary Figure S4), and to reveal compartment-specific biological functions based on the analysis of the interactomes extended with high-resolution localization data in a hierarchical structure. With the use of the optimized Localization and Interaction Scores (Figure 2) high-confidence interactomes could be created for further investigation in the field of compartment-specific biological processes (15).

Comparison of integrated protein–protein interaction data and the compartmentalized interactome allow the filtering of biologically unlikely interactions, where the interacting partners have no common subcellular localization. Our examples (Figure 4 and Supplementary Figure S7) illustrate that besides filtering, ComPPI has a strong predictive power to find new localizations of the proteins based on the underlying network or to suggest new compartment-specific biological functions. The comprehensive data set for four species gives the opportunity to analyse evolutionary aspects of the compartmentalization, such as the prediction of subcellular localization ortologes ('localogs').

The web interface of ComPPI (http://ComPPI.LinkGroup.hu) provides user-friendly search and download options. Besides the basic Search feature to explore and download the interactions of individual proteins, Advanced Settings could be applied to both query proteins and their interactors. Interactome-wide studies could be applied using the downloadable compartment-specific interactomes or the integrated protein–protein interaction data set, while the integrated subcellular localization data set is also available on the webpage for further analyses.

ComPPI is available at http://ComPPI.LinkGroup.hu, and has an open source code, which allows further improvement and the construction of 'ComPPI-based databases'. ComPPI is a community-annotated resource, which will be continuously enriched by a user community of experts helped by a public issue-tracking system and by feedbacks from the core-team, and will be updated and upgraded annually for minimum 5 years.

We plan to resolve current ComPPI limitations, such as the relatively low amount (29% of total) of experimental subcellular localization entries with the incorporation of newly available experimental data. Future plans include the development of improved gold standard-based Localization and network neighbourhood-based Interaction Scores, as well as further advanced download and search options, such as advanced localization-based network visualization and extended number of output formats.

In summary, the ComPPI-based interactomes introduced here provide a broader coverage, offer highly structured subcellular localization data, as well as offer Localization and Interaction confidence Scores, all in a user-friendly manner. Importantly, ComPPI enables the user to filter biologically unlikely interactions, where the two interacting proteins have no common subcellular localizations, and to predict novel subcellular localization as well as localization-based properties, such as compartment-specific biological functions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

Authors acknowledge the members of the LINK-Group (http://LinkGroup.hu) and Cellular Network Biology Group (http://NetBiol.elte.hu) for their advice, as well as Judit Gyurkó for the design of the ComPPI website.

FUNDING

Hungarian Science Foundation [OTKA-K83314]; Fellowship in computational biology at The Genome Analysis Centre (Norwich, UK) in partnership with the Institute of Food Research (University of Norwich, UK); Biotechnological and Biosciences Research Council, UK [to T.K.]; János Bolyai Scholarship of the Hungarian Academy of Sciences [to T.K.]. Funding for open access charge: Hungarian Science Foundation [OTKA-K83314].
Conflict of interest statement. None declared.

REFERENCES

1. Hao,N. and O'Shea,E.K. (2012) Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nat. Struct. Mol. Biol.*, **19**, 31–39.
2. Firth,S.M. and Baxter,R.C. (2002) Cellular actions of the insulin-like growth factor binding proteins. *Endocr. Rev.*, **23**, 824–854.
3. Azar,W.J., Zivkovic,S., Werther,G.A. and Russo,V.C. (2014) IGFBP-2 nuclear translocation is mediated by a functional NLS sequence and is essential for its pro-tumorigenic actions in cancer cells. *Oncogene*, **33**, 578–588.
4. Semenza,G.L. (2009) Regulation of oxygen homeostasis by hypoxia-inducible factor 1. *Physiology (Bethesda)*, **24**, 97–106.
5. Koh,G.C.K.W., Porras,P., Aranda,B., Hermjakob,H. and Orchard,S.E. (2012) Analyzing protein-protein interaction networks. *J. Proteome Res.*, **11**, 2014–2031.
6. Vidal,M., Cusick,M.E. and Barabási,A.-L. (2011) Interactome networks and human disease. *Cell*, **144**, 986–998.
7. Bulusu,K.C., Tym,J.E., Coker,E.A., Schierz,A.C. and Al-Lazikani,B. (2014) canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.*, **42**, D1040–D1047.
8. De Las Rivas,J. and Fontanillo,C. (2010) Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, **6**, e1000807.
9. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
10. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.
11. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
12. Wiwatwattana,N. and Kumar,A. (2005) Organelle DB: a cross-species database of protein localization and function. *Nucleic Acids Res.*, **33**, D598–D604.
13. Levy,E.D., Landry,C.R. and Michnick,S.W. (2009) How perfect can protein interactomes be? *Sci. Signal.*, **2**, pe11.
14. Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol.*, **327**, 919–923.
15. Lee,K., Sung,M.-K., Kim,J., Kim,K., Byun,J., Paik,H., Kim,B., Huh,W.-K. and Ideker,T. (2014) Proteome-wide remodeling of protein location and function by stress. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E3157–E3166.
16. Nastou,K.C., Tsaousis,G.N., Kremizas,K.E., Litou,Z.I. and Hamodrakas,S.J. (2014) The human plasma membrane peripherome: visualization and analysis of interactions. *Biomed. Res. Int.*, **2014**, 397145.
17. Sprenger,J., Fink,J.L. and Teasdale,R.D. (2006) Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformat.*, **7**, Suppl. 5:S3.
18. Chautard,E., Ballut,L., Thierry-Mieg,N. and Ricard-Blum,S. (2009) MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics*, **25**, 690–691.
19. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

20. Patil,A., Nakai,K. and Nakamura,H. (2011) HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res.*, **39**, D744–D749.
21. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
22. Faria,D., Schlicker,A., Pesquita,C., Bastos,H., Ferreira,A.E.N., Albrecht,M. and Falcão,A.O. (2012) Mining GO annotations for improving annotation consistency. *PLoS ONE*, **7**, e40519.
23. Barsky,A., Gardy,J.L., Hancock,R.E.W. and Munzner,T. (2007) Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, **23**, 1040–1042.
24. Huang,H., McGarvey,P.B., Suzek,B.E., Mazumder,R., Zhang,J., Chen,Y. and Wu,C.H. (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics*, **27**, 1190–1191.
25. Wein,S.P., Côté,R.G., Dumousseau,M., Reisinger,F., Hermjakob,H. and Vizcaíno,J.A. (2012) Improvements in the Protein Identifier Cross-Reference service. *Nucleic Acids Res.*, **40**, W276–W280.
26. Murali,T., Pacifico,S., Yu,J., Guest,S., Roberts,G.G. and Finley,R.L. (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. *Nucleic Acids Res.*, **39**, D736–D743.
27. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* , (2009) Human Protein Reference Database - 2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
28. Pagel,P., Kovac,S., Oesterheld,M., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stümpflen,V., Mewes,H.-W. *et al.* (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.
29. Chatr-Aryamontri,A., Breitkreutz,B.-J., Heinicke,S., Boucher,L., Winter,A., Stark,C., Nixon,J., Ramage,L., Kolas,N., O'Donnell,L. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.
30. Yu,H., Tardivo,L., Tam,S., Weiner,E., Gebreab,F., Fan,C., Svrzikapa,N., Hirozane-Kishikawa,T., Rietman,E., Yang,X. *et al.* (2011) Next-generation sequencing to generate interactome datasets. *Nat. Methods*, **8**, 478–480.
31. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
32. 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.
33. 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.
34. Kandasamy,K., Keerthikumar,S., Goel,R., Mathivanan,S., Patankar,N., Shafreen,B., Renuse,S., Pawar,H., Ramachandra,Y.L., Acharya,P.K. *et al.* (2009) Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res.*, **37**, D773–D781.
35. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
36. Lu,P., Szafron,D., Greiner,R., Wishart,D.S., Fyshe,A., Pearcy,B., Poulin,B., Eisner,R., Ngo,D. and Lamb,N. (2005) PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Res.*, **33**, D147–D153.
37. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2007) eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Res.*, **35**, D208–D212.
38. Sprenger,J., Lynn Fink,J., Karunaratne,S., Hanson,K., Hamilton,N.A. and Teasdale,R.D. (2008) LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res.*, **36**, D230–D233.
39. Wiwatwattana,N., Landau,C.M., Cope,G.J., Harp,G.A. and Kumar,A. (2007) Organelle DB: an updated resource of eukaryotic protein localization and function. *Nucleic Acids Res.*, **35**, D810–D814.
40. Hoelz,A., Debler,E.W. and Blobel,G. (2011) The structure of the nuclear pore complex. *Annu. Rev. Biochem.*, **80**, 613–643.
41. Brameier,M., Krings,A. and MacCallum,R.M. (2007) NucPred–predicting nuclear localization of proteins. *Bioinformatics*, **23**, 1159–1160.
42. Waterson,R.M. and Hill,R.L. (1972) Enoyl Coenzyme A Hydratase (Crotonase): catalytic properties of Crotonase and its possible regulatory role in fatty acid oxidation. *J. Biol. Chem.*, **247**, 5258–5265.
43. Hamed,R.B., Batchelar,E.T., Clifton,I.J. and Schofield,C.J. (2008) Mechanisms and structures of crotonase superfamily enzymes–how nature controls enolate and oxyanion reactivity. *Cell. Mol. Life Sci.*, **65**, 2507–2527.
44. Turteltaub,W. and Murphy,A. (1987) Subcellular localization and capacity of ß-oxidation and aldehyde dehydrogenase in porcine liver. *Arch. Biochem. Biophys.*, **255**, 120–126.
45. Bhattacharyya,T., Karnezis,A.N., Murphy,S.P., Hoang,T., Freeman,B.C., Phillips,B. and Morimoto,R.I. (1995) Cloning and subcellular localization of human mitochondrial hsp70. *J. Biol. Chem.*, **270**, 1705–1710.
46. Zhang,J., Sun,M., Li,R., Liu,S., Mao,J., Huang,Y., Wang,B., Hou,L., Ibrahim,M.M. and Tang,J. (2013) Ech1 is a potent suppressor of lymphatic metastasis in hepatocarcinoma. *Biomed. Pharmacother.*, **67**, 557–560.
47. Yeh,C.-S., Wang,J.-Y., Cheng,T.-L., Juan,C.-H., Wu,C.-H. and Lin,S.-R. (2006) Fatty acid metabolism pathway play an important role in carcinogenesis of human colorectal cancers by microarray-bioinformatics analysis. *Cancer Lett.*, **233**, 297–308.
48. Chen,Z., Dai,Y.-C., Chen,Z.-X., Xie,J.-P., Zeng,W., Lin,Y.-Y. and Tan,Q.-H. (2013) Knockdown of ECHS1 protein expression inhibits hepatocellular carcinoma cell proliferation via suppression of Akt activity. *Crit. Rev. Eukaryot. Gene Exp.*, **23**, 275–282.
49. Liu,X., Feng,R. and Du,L. (2010) The role of enoyl-CoA hydratase short chain 1 and peroxiredoxin 3 in PP2-induced apoptosis in human breast cancer MCF-7 cells. *FEBS Lett.*, **584**, 3185–3192.
50. Fisk,H.A., Mattison,C.P. and Winey,M. (2003) Human Mps1 protein kinase is required for centrosome duplication and normal mitotic progression. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 14875–14880.