



Funded by the Seventh Framework
Programme of the European Union



Project full title:

Hepatic and Cardiac Toxicity Systems modelling

Project acronym:

HeCaTos

Collaborative project

HEALTH.2013.1.3.-1:

Modelling toxic response in case studies for predictive human safety assessment

FP7-HEALTH-2013-INNOVATION-1-602156-HeCaTos

Deliverable Report D2.2:

Report on Consensus PathDBTOX update

Work package 2

Due date of deliverable: M36

Actual submission date: October 2016

Start date of project: October, 2013

Duration: 60 months

Maastricht University (UM)

Project co-funded by the European Commission within the 7th Framework Programme (2013-2018)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributions to deliverable - Internal review procedure

Deliverable produced by:	Date:
Ralf Herwig - Partner MPIMG	10.10.2016
Deliverable internally reviewed by:	Date:
Jos Kleinjans - Partner UM	October 2016

Contents

Publishable Summary.....	3
Objectives	3
Introduction	3
Results.....	4
Analysis features of ConsensusPathDB.....	4
Annotation sets.....	5
Identifier mapping	6
Pathway annotation - specificity and redundancy	6
Interaction retrieval for single biomolecules.....	6
Over-representation analysis.....	7
Enrichment analysis	8
Network module analysis.....	9
Summary	10
Analysis features of ToxDB	10
Difficulties	11
References	11
Annex	11

PUBLISHABLE SUMMARY

ConsensusPathDB consists of a comprehensive collection of human (as well as mouse and yeast) molecular interaction data integrated from 32 different public repositories and a web interface featuring a set of computational methods and visualization tools exploring these data. MPIMG has updated ConsensusPathDB with respect to the functional and network-based characterization of biomolecules (genes, proteins, metabolites) that are submitted to the system either as a priority list or together with associated experimental data, such as RNA-seq or proteome data. The tool reports interaction network modules, biochemical pathways and functional information that are significantly enriched by the user's input applying computational methods for statistical over-representation, enrichment and graph analysis. The resulting network associations will be used by WP11 and WP12 in order to interpret high-throughput data mechanistically, to characterize and prioritize biomarkers and to integrate different omics levels.

OBJECTIVES

The goal of this deliverable is to adapt the ConsensusPathDB functionality, as the central pathway resource of HeCaToS, to the needs of the consortium. MPIMG will use, extend and modify the integrated interaction graph of the ConsensusPathDB in order to identify functional network modules for liver and heart toxicity from heterogeneous data (WP11 and WP12). These networks will be evaluated for integration in physiological models in WP3 and WP4. A toxicity-focused instance of the ConsensusPathDB will be developed based on these results (MPIMG).

INTRODUCTION

In ConsensusPathDB, we have agglomerated the contents of 32 major public repositories on human molecular interactions of heterogeneous types as well as biochemical pathways (Table 1) resulting in one of the largest interactome collections available. In addition, the database integrates the contents of 15 mouse and 14 yeast interaction repositories.

In addition to gene ontology (GO) and pathway annotations, ConsensusPathDB explores systematically the protein-protein interaction (PPI) network, since PPIs are key drivers of biological function. Proteins do not exert their function in isolation but rather through interactions with other proteins. However, still only a minor fraction of the estimated ~650,000 human protein interactions have been experimentally measured. Moreover, these interactions, along with other interaction information on pathways, biochemical reactions, drug-target interactions, etc., are scattered across more than 500 different databases worldwide what necessitates the integration of as many as possible interaction resources in meta-databases such as ConsensusPathDB. The benefit of such interaction integration is on the one hand a better coverage of the interactome improving guidance in the functional interpretation of *omics* data and on the other hand a partially redundant, and thus reassuring, content that allows judging interaction confidence.

ConsensusPathDB is well-adopted by the research community, mostly due to its ease of use, comprehensive data basis, and a set of features missing in many other tools, such as pathway-centric metabolite set analysis and gene set analysis with protein complexes and network modules. The ConsensusPathDB web interface is regularly used for over-representation analysis to characterize

diverse sets of molecules, gene set enrichment analysis or to derive gene regulatory network modules in various contexts. Furthermore, ConsensusPathDB is used as a data basis by other tools, for example for enrichment analysis by Chipster using web service connections or by Cytoscape using a Java plugin for assessing interaction confidence of PPIs. In addition to these analyses, which explore the content of ConsensusPathDB for characterization of gene lists, the tool can be used as a resource for generating molecular interaction gene sets, which themselves can be used as predictive signatures (cf. Deliverable Report D12.1). For example, it has been shown that network modules and pathways can be derived as predictive patterns in cancer diagnostics as well as tumor progression monitoring and by that enable extrapolating biomarker analysis from single molecules to entire pathways.

RESULTS

Analysis features of ConsensusPathDB

A detailed protocol of the analysis features of ConsensusPathDB has been published as a Nature Protocol during the third reporting period (Herwig et al., 2016).

ConsensusPathDB integrates comprehensive interaction networks (**Table 1**) and contains pre-defined annotation sets that hold functional information such as pathways, GO categories, protein complexes and PPI network neighborhoods that were derived from the integrated resources.

Depending on the user's input, ConsensusPathDB allows the following analyses (**Fig. 1**):

- **Analysis path 1:** the interaction neighborhood of a single molecule can be tracked and a corresponding network can be generated; this can be done, for example, to infer network-level information (i.e. interaction partners) for biomarkers of interest.
- **Analysis path 2:** uploading a list of molecules (genes, proteins and metabolites) allows either performing over-representation analysis with pre-defined annotation sets from ConsensusPathDB or computing network associations between the molecules of interest through mining the integrated interaction network.
- **Analysis path 3:** inserting a weighted list of molecules allows computing enrichment analysis with the annotation sets of ConsensusPathDB using experimental data; this path employs a more unbiased analysis compared to analysis path 2 because it is not dependent on a pre-defined priority list of molecules.

Content type		Human	Mouse	Yeast
Integrated databases		32	15	14
Unique physical entities		158,523	31,679	17,672
Unique interactions		458,570	34,064	272,094
	gene regulations	17,098	2,196	316
	protein interactions	261,085	23,488	123,842
	genetic interactions	443	194	145,151
	biochemical reactions	21,070	8,186	2,785
	drug-target interactions	158,874	0	0
Pathway gene sets		4,593	2,173	1,101

Table 1: Content of ConsensusPathDB

Annotation sets

ConsensusPathDB offers four types of pre-defined annotation sets: neighborhood-based entity sets (NESTs), protein complexes, pathways and gene ontology terms (GOs). NESTs: These sets are derived from the integrated interaction network, including four types of biological interactions: protein-protein, biochemical, gene regulatory and genetic interactions. A NEST is defined as a central protein and its network neighbors. The size of the network neighborhood is determined by its radius. The user can choose between radius equal to one and two respectively. Radius equal to one adds only the direct neighbors to the center protein while radius equal to two adds, in addition, all direct neighbors of the direct neighbors. We recommend rather using radius equal to one, otherwise the neighborhoods grow too large and lose specificity.

There are as many NESTs as proteins in the integrated network.

1. **Protein complexes:** These sets are derived from specific databases that hold information on protein complexes. Most annotated protein complex sets are rather small (2-3 members);
2. **Pathways:** These sets comprise metabolic, signaling and gene regulatory pathways annotated by 12 source databases for human (4 for mouse and yeast each). Pathways range from very large biological processes covering for example the complete metabolism with >1000 members, to very specific concepts that describe detailed processes;
3. **GO terms:** ConsensusPathDB offers four levels of GO categories ranging from very general terms (level = 2) with >1000 members, to more specific terms (level = 5). In the analysis, the user can restrict the categories to specific level(s) or to the specific GO tree branches covering '*biological process*', '*molecular function*' and '*cellular compartment*'.

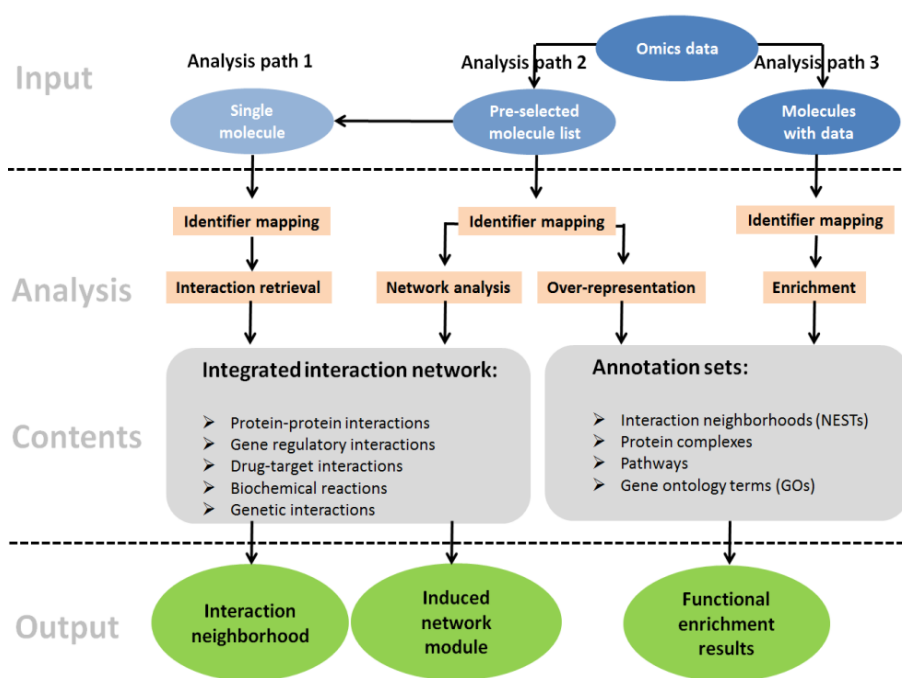


Figure 1: ConsensusPathDB analysis features. Three paths of analyses are possible depending on the user's input. The content of the ConsensusPathDB (i.e. the integrated interaction graph and the pre-defined annotation sets) can be explored with single molecules (analysis path 1), with priority lists of molecules (genes, proteins and metabolites; analysis path 2) or with associated experimental data (analysis path 3). The web server functionality includes over-representation analysis, enrichment analysis and network module analysis. The outputs are the generated tables and graphs that can be downloaded for further inspection.

Identifier mapping

A recurrent problem when integrating data from different resources, or when analyzing high-throughput data by comparison with existing databases, is the non-uniformity of gene/protein/metabolite identifiers used. In ConsensusPathDB, we have created comprehensive identifier maps by parsing the contents of 11 major genomic, proteomic and metabolite databases such as Ensembl, Uniprot and PubChem. These maps were used to match genes, proteins and metabolites from the 32 sources of interaction and pathway information currently integrated in ConsensusPathDB. Furthermore, they are used to map identifiers from the user input to these physical entities, and hence allow great flexibility with respect to what identifier namespace is used by the user.

Pathway annotation - specificity and redundancy

The pathway concept is essential for modern biology and usually describes a certain cellular process, for example '*apoptosis*', in which the involved proteins or metabolites exert specific functions and are interconnected by molecular interactions.

ConsensusPathDB agglomerates 4,593 human pathway concepts (mouse: 2,173, yeast: 1,101) originating from 12 different resources. These pathway concepts are partially redundant, on the one hand because they describe sub-pathways of a given pathway that are annotated by the same database. For example, the pathway '*apoptosis*' might cover the sub-pathways '*extrinsic apoptosis*' and '*intrinsic apoptosis*' among others with corresponding subsets of proteins. On the other hand, most generic pathways are annotated by several databases leading, for example, to more than one annotation set referring to '*apoptosis*'.

It is worth noting that pathway concepts from different resources might in fact involve different sets of molecules even when describing similar molecular processes. As a consequence, this could lead to differences in functional enrichment analyses (analysis paths 2 and 3) because the different annotation sets might have deviating overlaps with the gene list submitted by the user.

For example, comparison of gene sets for the '*apoptosis*' pathway in the three widely used databases KEGG, Reactome and WikiPathways reveals that 81% of the annotated proteins are specific to a single database compared to the number of proteins that are shared by at least two of the three databases (**Fig. 2a**). The reason for this is that pathway boundaries are not clearly defined and that expert opinion on the extent of cross-talk with other pathways is highly variable. Additionally, pathway annotations are commonly focused on specific substructures or specific cellular contexts (e.g. tissues, diseases, organisms) what might result in variations of the assembled gene lists.

Consequently, in ConsensusPathDB such overlapping pathway concepts are not merged to generalized pathways; instead, the redundancy is kept and the annotated pathway set is always disclosed together with its source database.

Interaction retrieval for single biomolecules

ConsensusPathDB holds 158,523 unique physical entities (mouse: 31,679, yeast: 17,672) and offers the possibility to retrieve interaction information for these entities. The concept of an interaction in ConsensusPathDB is very general so that proteins can have connections not only to other proteins, but also to drugs, complexes or metabolites. By selecting specific interactions, the user can generate fairly complex interaction networks.

The source database of each interaction is tracked by a color-code, on the one hand providing the user with the information where the interaction is originally coming from and, on the other hand, displaying the redundancy which might serve as an indicator for assessing confidence of the particular interaction. Figure 2b shows the distribution of the different interaction types and their origin. Most interactions are present for protein-protein and drug-target interaction types and predominantly specific for a single or low number of databases.

Another level of confidence assessment is available for binary PPIs. Because a lot of PPI resources are integrated in ConsensusPathDB, control of false positive interaction is of utmost importance. Therefore, binary PPIs have a quality score (range [0,1]) which is displayed with a color code. This score was computed as a meta-score integrating different methods for interaction confidence assessment including graph-based topological criteria, literature evidence and pathway co-occurrence and semantic similarity using our IntScore web tool.

Over-representation analysis

This feature comprises the interrogation of the annotation sets (pathways etc.) with lists of genes, proteins or metabolites. This analysis requires prior data analysis by the user outside of ConsensusPathDB, for example by applying a statistical test to the genome data and pre-selecting the most significant molecules as is typically done for RNA-seq or microarray data. For computing the significance of the over-representation of the annotation sets with respect to user-input molecules, ConsensusPathDB applies Fisher's exact test evaluating a 2x2 contingency table and uses a false discovery rate (FDR) correction for multiple testing. This is a widely used technique also applied by many other tools. After computation, annotation sets are ordered according to significance and can be downloaded in table format. Additionally, specific sets and their overlaps can be visualized. Importantly, in contrast to many other tools, not only gene and protein lists can be submitted to ConsensusPathDB but also lists of metabolites; furthermore, the resulting functional categories can be visualized as overlap graphs, described below.

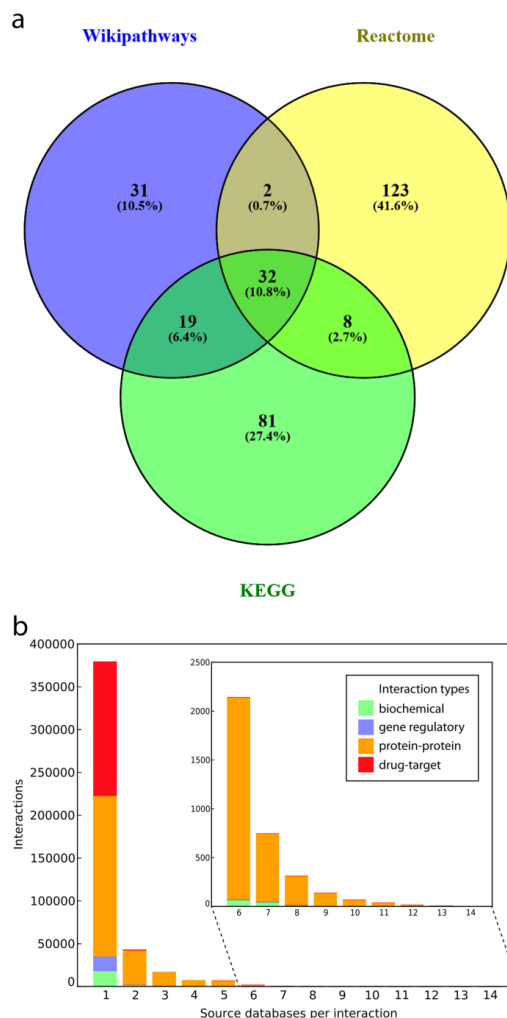


Figure 2. (a) Annotation specificity. VENN diagram showing the proteins annotated for the apoptosis signaling pathway in three different databases (Wikipathways, WP254; Reactome, R-HSA-109581; KEGG, hsa:04210). In total, 255 different proteins are annotated for apoptosis signaling, 84 in Wikipathways, 161 in Reactome (proteins with gene symbols) and 85 in KEGG. 49 of these proteins are common to all or at least two databases (19%) while the vast majority of proteins (206; 81%) are specific for a single database. **(b)** Histogram of the number of contributing databases per interaction (genetic interactions have been omitted in this figure since their total number, $n=443$, is comparatively too small to be visible).

Enrichment analysis

This feature comprises the enrichment analysis using all experimental data in an unbiased way rather than prioritized molecule lists. It is worth noting that this analysis path is complementary to the above: With over-representation analysis only the most significant changes are typically evaluated whereas with enrichment analysis also subtle but congruent changes of many molecules in parallel are appreciated. Therefore, in practice a pathway might emerge as significantly enriched although none of the genes in that pathway were in the priority list submitted for over-representation analysis. Furthermore, completely novel information can be retrieved, for example involving interaction neighborhoods (NESTs) of proteins that were not even captured by the omics platform under consideration.

ConsensusPathDB assumes that the user submits case-control data, for example disease vs. normal or treated vs. untreated conditions, and that for each molecule both values are given. Alternatively, a single log-fold change value for each molecule can be submitted.

For computing enrichment P-values for the annotation sets ConsensusPathDB applies Wilcoxon's matched-pairs signed rank test. This test is very robust against experimental outliers and is based on ranks rather than real values. For each annotation set i with n molecules m_{i1}, \dots, m_{in} and experimental measurements x_{i1}, \dots, x_{in} and y_{i1}, \dots, y_{in} the ranks of the differences, $x_{ij} - y_{ij}$, of the two experimental conditions are computed; next, all ranks with positive differences R^+ and negative differences R^- are summed. The test statistic is the minimum of both rank sums: $R = \min\{R^+, R^-\}$. Expectation, $E(R)$, and variance, $Var(R)$, of R can be derived and a final Z-score measures the observed deviation:

$$Z = \frac{R - E(R)}{\sqrt{Var(R)}},$$

$$\text{with } E(R) = \frac{n(n+1)}{4} \text{ and } Var(R) = \frac{n(n+1)(2n+1)}{24} \text{ respectively}^{45}.$$

If both conditions have similar values for m_{i1}, \dots, m_{in} , then both rank sums are equal and the resulting test statistic is not significant, while otherwise the Z-score gets significant.

As in the previous section, annotation sets are ordered according to significance and can be downloaded in table format. Additionally, specific sets and their overlaps can be visualized also by overlaying color-coded experimental (e.g. gene expression) data.

Network module analysis

This feature is relevant for both gene or protein lists. The idea behind this kind of analysis is to start with a priority list of genes/proteins (*seed nodes*) and to compute from the underlying interaction graph a subgraph that connects the seed nodes together through functional and physical links. In ConsensusPathDB we implemented a previously published induced network approach. The output subnetwork produced by this algorithm may optionally include nodes, which are not originally part of the user's list but have significantly many connections to seed nodes. These nodes are called *intermediate* nodes. Intermediate nodes are potential specific regulators and specific participants in pathways, protein complexes and modules involving the input seed list. For example, if a list of down-regulated genes as measured through RNA-seq is uploaded, ConsensusPathDB may output a common transcription factor regulating those genes that might be mutated or otherwise dysregulated (but is not on the transcriptional level and hence not included in the seed list).

In order to judge the significance of the intermediate node a Z-score value is computed using a binomial proportions test as follows:

$$Z = \frac{\left(\frac{a}{c} - \frac{b}{d}\right)}{\sqrt{\frac{\frac{b}{d}\left(1 - \frac{b}{d}\right)}{d}}}.$$

Here, a equals the number of links from the intermediate node being examined to nodes from the input seed list, b equals the number of total links for the intermediate node in the consolidated background reference network, c is the number of total links in the output subnetwork, and d is the number of total links in the consolidated background reference network. The threshold for the Z-score can be set

interactively by the user. It is clear that for large lists and/or low Z-scores the outputted network can be fairly large so that we recommend either not to use more than 100 genes/proteins as seed nodes or to restrict the analysis to specific interaction types in the parameter setting.

Summary

To summarize, ConsensusPathDB analysis tools aim to enable network level interpretation and functional characterization of user-specified lists of molecules (genes, proteins, metabolites) and associated high-throughput data. ConsensusPathDB helps users working with such data to infer heterogeneous interaction networks for genes, proteins, metabolites, drugs and other biomolecules to:

- Compute over-represented pathways, PPI networks, protein complexes and GO annotations from a priority list of genes, proteins or metabolites;
- Compute enriched pathways, PPI networks, protein complexes and GO annotations from genome-wide data such as RNA-seq, ChIP-seq, or array technology;
- Generate network modules that are over-represented by genes or proteins and thereby exploring heterogeneous interactions such as PPI, drug-target, gene regulatory and genetic interactions.

Analysis features of ToxDB

A detailed protocol of the analysis features of ToxDB has been published in the Database journal during the third reporting period (Hardt et al., 2016) and was reported already in Milestone Report MS27 in the second reporting period.

ToxDB is a repository containing benchmark data sets and stores pathway-level information on drug action derived from omics data along with chemical and toxicity information and information on experimental design. Toxicity pathway data is derived from the combination of pre-defined pathway concepts, as given by the ConsensusPathDB, and gene expression and potentially other omics data provided by two large public studies, TG-GATES and DrugMatrix.

ToxDB allows the user accessing drug-treatment data available for currently 437 different drugs measured in human hepatocytes as well as rat *in vitro* and *in vivo* data with respect to several target organs (liver, heart, kidney). Benchmark data includes 5,205 human pathways agglomerated from 12 different pathway repositories, 11 different cell types and a total of 7,464 drug treatments. The methodological basis for such an approach is a statistical method that was previously developed by us and that allows scoring entire pathways rather than single genes with any kind of omics data. In particular, scoring methods that take into account time and dosage factors have been tested and compared.

The user can access the pathway-level information in ToxDB for a given compound and derive most responding pathways for a specific experimental condition. Switching from pathway to gene level important markers in these pathways can be identified. Additionally, the web server allows user interaction by computing statistical tests for sets of experimental conditions, for example comparing differential pathway responses for specific compounds or sets of compounds *in vitro* against *in vivo*, human against rat or with different target organs. Furthermore, for a given pathway results for all compounds can be retrieved what allows tracking the response of this particular pathway across all compounds. In order to overlay toxicity information for the chemicals under study and in order to inform on chemical information we have linked the system to published toxicity classification, chemical resources. ToxDB can be accessed by the consortium via the compounds section of the central HeCaToS project repository (WP9).

DIFFICULTIES

None.

REFERENCES

- Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature Protocols*. 2016 Oct;11(10):1889-907.
- Hebels DG, Rasche A, Herwig R, VanWesten GJ, Jennen DG, Kleinjans JC. A Systems Biology Approach for Identifying Hepatotoxicant Groups Based on Similarity in Mechanisms of Action and Chemical Structure. *Methods Mol Biol*. 2016;1425:339-59.
- Hardt C, Beber ME, Rasche A, Kamburov A, Hebels DG, Kleinjans JC, Herwig R. ToxDB: pathway-level interpretation of drug-treatment data. *Database* (Oxford). 2016 Apr 13;2016. pii: baw052. doi: 10.1093/database/baw052.

ANNEX

Pdf files of the corresponding papers for ToxDB and ConsensusPathDB have been submitted as supplementary material to this deliverable report and accessible at www.hecatos.eu website.