



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 D12.3:

Report on genotype phenotype correlation

Work package 12

Due date of deliverable: M36

Actual submission date: M45

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	October 2016
Ralf Herwig - Partner MPIMG	June 2017
Deliverable internally reviewed by:	Date:
Jos Kleinjans - Partner UM	July 2017

Contents

Publishable Summary.....	3
Objectives	4
Introduction	4
Results.....	5
Developing a pipeline for predicting network modules from omics data	5
BioNet	6
HotNet2.....	6
Differential equation model.....	9
Difficulties	10
References	10
Annex 1	11

PUBLISHABLE SUMMARY

The goal of this deliverable is to establish computational methods for predicting network and pathway components from experimental molecular genotype data that can be correlated with the toxicity phenotype under study. The modules represent “minimal” numbers of network components that trigger the disease phenotype. They will be derived from either protein-protein interaction networks or components of existing pathways.

Partners have set up a computational pipeline which is adapted from components built in WP2 (Deliverable Report 2.3). The computational prediction of compound effects from molecular data is an important task in hazard and risk assessment and pivotal for judging the safety of any drug, chemical or cosmetic compound. In particular, the identification of such compound effects at the level of molecular interaction networks is helpful for the construction of an adverse outcome pathway (AOP) that emerged as a guiding concept for toxicity prediction, because of the inherent mechanistic information of such networks. In fact, adding molecular interactions as an additional layer of information and observing expression changes in highly interacting genes might allow identifying the key molecular initiating events of compound toxicity.

The approach is composed of different tool:

- i)* for primary data analysis, i.e. the biostatistical quantification of the gene expression changes;
- ii)* for functional annotation and prioritization, such as pathway enrichment analysis and literature mining, as well as
- iii)* for the construction of appropriate interaction networks and the identification of predictive modules from these molecular networks.

The approach has been described in a recent publication (Hardt et al., Methods Mol Biol, in press).

OBJECTIVES

General objectives of WP12 are to:

- Collect benchmark data sets for liver and heart toxicity AOPs;
- Perform model analysis and comparison with respect to relevant use cases;
- Optimize modelling parameters.

Objective of Deliverable 12.3 is to develop tools for genotype-phenotype correlation from molecular data and to apply these tools in the course of the case studies carried out by WP12.

INTRODUCTION

The explanatory power of a single event (gene/protein expression fold change, differentially methylated region, single mutation, etc.) for toxicity is rather low, meaning that a multitude of different events can lead to the same phenotypic outcome. Thus, toxicity progression might be highly individual with respect to the patient/cell line under study. As a consequence the contribution of a single factor cannot be properly used for explaining genotype-phenotype relationships. A common explanation of the fact that different patterns of molecular changes evolve to the same phenotype is that these changes can dysregulate the same pathways or network environment (Muller et al., 2008). Thus, recent research, with cancer being the pioneering field, has focused on biological networks as a model of the biological system rather than single biomarkers and the analysis of genotype in the context of these networks.

The power of network-based interpretation of mutations has been shown efficiently in cancer (Creixell et al., 2015) but is missing yet for toxicity prediction.

Essentially two types of analyses have been performed:

- i) *Supervised* pathway analysis, and
- ii) *Unsupervised* network analysis.

The first approach utilizes our knowledge on biological pathways and interprets variants on the respective gene/protein/metabolite sets defined by these pathways. Methods are then either over-representation analysis or enrichment analysis based on different statistical models. Such methods are implemented in many web-accessible tools such as GSEA (Subramanian et al., 2005), DAVID (Huang et al., 2009) or the ConsensusPathDB (Herwig et al., 2016) that has been developed by Partner MPIMG. This approach identifies significantly enriched pathways that agglomerate a large number of the observed mutations and is supervised in the sense that only pre-defined, well-annotated pathways are taken under consideration. The second approach utilizes interaction networks (e.g. protein-protein interactions) and maps the observed experimental data onto these networks. Then, methods from mathematical graph theory are applied in order to identify subnetworks, called network modules that are enriched for the effected genes (Creixell et al., 2015). The second approach is unsupervised because no grouping of genes/proteins/metabolites into pathways is done prior to data analysis. This allows, for example, also quantifying cross-talk between different pathways or novel network associations.

In this Deliverable Report D12.3 we propose a computational pipeline for inferring omics data at the level of molecular networks in order to compute network modules that are predictive for drug toxicity. We surveyed and implemented several approaches from literature and tested the pipeline on public and project data.

RESULTS

Developing a pipeline for predicting network modules from omics data

An approach to better explain how multiple genomic aberrations integrate into toxic phenotypes is given by network approaches (Yi et al., 2017). Genotype-phenotype correlation network analysis was originally developed for associating mutations with phenotypic (clinical or assay-based) functional measurements. In this project we extrapolate these methods to other omics data since not only mutations can exert their effects over the network but also gene/protein expression changes etc. Thus, the original concept of "genotype" in our context is extrapolated to all omics measurements. Methodologically, it is about the way to weight the nodes in the underlying interaction graph.

In HeCaToS Partner MPIMG uses such network approaches as an integration concept. There are different ways how to incorporate cross-omics. "Genotype-phenotype" approaches employ three components:

- i. The underlying network (for example protein-protein interactions);
- ii. The node weighting (for example gene scores (cf. Deliverable Report D12.1));
- iii. The network propagation method.

Cross-omics data could be introduced through points i or ii.

In a first attempt Partner MPIMG has inferred the protein-protein network and incorporated cross-omics data through the node weighting process. This can be done either by running the algorithm multiple times (transcriptome, proteome, methylome, etc.) with node weights adapted from the respective fold-changes and then identify network overlaps of the computed modules or by running the algorithm once and use a combined node score. These approaches are conceptually integrated into the pipeline described in Deliverable Report D2.3.

Another idea to introduce cross-omics data through *i* would be to use a network similar to "weighted gene co-expression networks". Here, one would for example use all data (all treatments, all time points, all omics platforms) and draw a link between two genes when they are highly correlating. This network would then represent a microtissue co-expression network. Then, for an individual experiment the nodes would be weighted according to ii and network modules would be computed according to iii. This would lead to modules that deliver aberrant parts of the co-expression network due to the specific treatment.

The pipeline was built in cooperation with partners in WP2 in the previous reporting periods (cf. D2.3). In the 3rd reporting period this pipeline was substantially extended with new functionality. The joint work of Partner MPIMG and Partner MD resulted in a recent publication in *Methods Mol. Biol. on Computational Cell Biology* (Hardt et al., in press; Annex 1).

New functionality includes the incorporation of pipelines for sequencing data (RNA-seq, Methyl-seq) and Proteomics, new node weighting functions as well as new network propagation approaches.

Genotype-phenotype network prediction methods

BioNet

BioNet is available as an R-package that can easily be integrated in computational workflows (Beisser et al., 2010).

Node weighting procedure is done by assigning each node (gene/protein) a P-value according to experimental measurements (for example fold-change significance when comparing toxic vs control states) but could alternatively also be done with any kind of experimentally-informed value.

The distribution of P-values is transformed into a mixture of two Beta-functions in order to separate signal ($B(a,1)$ distribution) from noise ($B(1,1)$ uniform distribution).

Such a function has the form:

$$f(x|a, \lambda) = \lambda + (1 - \lambda)ax^{a-1} \quad \text{with } 0 < x \leq 1, 0 < a < 1.$$

The function uses the initial P-values to derive estimates for a and λ .

In the next step, given a certain false-discovery rate, $t(FDR)$, this fit is used to derive for each node x a node score, $S(x)$, by:

$$S(x) = (a - 1) \left(\log(x) - \log(t(FDR)) \right).$$

Once the nodes of the PPI network have been weighted according to these values the BioNet approach applies linear integer programming in order to derive a high-scoring sub-network by the following steps:

1. In the first step all positive connected nodes are aggregated into higher level nodes called meta-nodes;
2. Edges are scored based on the connected node's scores;
3. From these edge scores a minimum spanning tree (MST) is calculated;
4. All paths between positive meta-nodes are calculated based on the MST to obtain the negative nodes between the positives;
5. Upon these negative nodes again an MST is calculated from which the path with the highest score, which is based on node scores of negative nodes and the positive meta-nodes they connect, gives the resulting approximated module.

The result of the BioNet algorithm is a sub-network that agglomerates highly significant nodes. The sizes of the modules are dependent on the algorithmic parameter, i.e. the false-discovery rate $t(FDR)$.

Results on identification of network modules with respect to cardiac and hepatic compounds have been documented in the accompanying paper (Annex 1).

HotNet2

Hotnet2 has been developed originally for propagating the effect of mutations through protein-protein networks but can also be used to predict network modules from gene/protein expression data. HotNet2 uses a biophysical model of heat diffusion assuming that each affected node (for example a significantly differentially expressed gene) sends heat to its neighbouring nodes whereby the heat diffuses upon network distance. Thus, neighbouring nodes collect more heat from the affected node than more distant nodes. This procedure is done for all affected nodes. At the end of the heat diffusion procedure the heat is collected for each individual node and network modules are computed which consists of edges between nodes with a sufficient heat. HotNet2 comes with a standalone software with own visualization features (Leiserson et al., 2015).

Partner MPIMG has adapted this software and tested it on patient proteome data. For each node heat was introduced according to the correlation of the protein with the LVEF (left ventricular ejection fraction) clinical parameter as the phenotypic output (cf. D11.1). Network propagation results in 12 network modules with sizes ≥ 5 .

Network modules computed with HotNet.

Module	Size	Nodes
1	37	ACPM_HUMAN, ACS2L_HUMAN, AL1B1_HUMAN, DECR_HUMAN, DHRS4_HUMAN, ECH1_HUMAN, ECHB_HUMAN, ETFA_HUMAN, FUMH_HUMAN, HYES_HUMAN, IF2M_HUMAN, IPYR2_HUMAN, ITA7_HUMAN, LYRM4_HUMAN, LYRM7_HUMAN, MIC60_HUMAN, NAR3_HUMAN, NDUAA_HUMAN, NDUB9_HUMAN, NDUS1_HUMAN, NDUS8_HUMAN, NPS3A_HUMAN, OPA1_HUMAN, PRDX3_HUMAN, PSMD9_HUMAN, PYGB_HUMAN, RL10_HUMAN, RL15_HUMAN, RL21_HUMAN, RL26_HUMAN, RL3_HUMAN, SODM_HUMAN, SPG7_HUMAN, SYDC_HUMAN, VAPA_HUMAN, XPP1_HUMAN, ZADH2_HUMAN
2	18	ANXA2_HUMAN, APOH_HUMAN, BCAT2_HUMAN, CATB_HUMAN, CO6A1_HUMAN, LEG1_HUMAN, LEG3_HUMAN, LG3BP_HUMAN, MYOC_HUMAN, NID2_HUMAN, PEDF_HUMAN, PGBM_HUMAN, PTGDS_HUMAN, RAB1A_HUMAN, RB11B_HUMAN, S10AA_HUMAN, SBP1_HUMAN, TMM65_HUMAN
	15	DTNA_HUMAN, KCC2D_HUMAN, MAP4_HUMAN, MARK4_HUMAN, MY18A_HUMAN, MYH10_HUMAN, NDUA8_HUMAN, PRS10_HUMAN, RYR2_HUMAN, S10A4_HUMAN, SNTA1_HUMAN, STIP1_HUMAN, SYNEM_HUMAN, UB2V2_HUMAN, UFM1_HUMAN
3	9	KTN1_HUMAN, LDB3_HUMAN, MYOZ2_HUMAN, PSMD7_HUMAN, RHG01_HUMAN, RHOA_HUMAN, RL18_HUMAN, RS21_HUMAN, UFD1_HUMAN
6	6	ACPH_HUMAN, BAP31_HUMAN, DNM1L_HUMAN, FIS1_HUMAN, HS74L_HUMAN, MFF_HUMAN
5	6	ATP5H_HUMAN, MATR3_HUMAN, NDUAC_HUMAN, STML2_HUMAN, SYIM_HUMAN, SYSM_HUMAN
6	6	1433S_HUMAN, IF4B_HUMAN, RL29_HUMAN, RL4_HUMAN, RL8_HUMAN, SRBS1_HUMAN
7	6	MIC13_HUMAN, MPCP_HUMAN, MTX2_HUMAN, NDUB8_HUMAN, NDUV3_HUMAN, ODPX_HUMAN
8	5	GELS_HUMAN, MYO5C_HUMAN, NLTP_HUMAN, TMOD1_HUMAN, TYB10_HUMAN
9	5	AAKG2_HUMAN, EIF3A_HUMAN, FUBP2_HUMAN, NP1L4_HUMAN, TBB6_HUMAN
10	5	ANKR1_HUMAN, CASQ2_HUMAN, GDIR2_HUMAN, HEBP2_HUMAN, LPPRC_HUMAN
11	5	GDE_HUMAN, LGUL_HUMAN, NUDC_HUMAN, VPS25_HUMAN, VPS35_HUMAN
12	5	AAKG2_HUMAN, EIF3A_HUMAN, FUBP2_HUMAN, NP1L4_HUMAN, TBB6_HUMAN

As expected, the modules agglomerate proteins that are correlating with the LVEF clinical parameter, here e.g. for the first module of size 37:

No.	Protein	LVEF-expression correlation
1	ACPM_HUMAN	0,266569
2	ACS2L_HUMAN	0,06922
3	AL1B1_HUMAN	-0,27246
4	DECR_HUMAN	0,381444
5	DHRS4_HUMAN	0,368189
6	ECH1_HUMAN	0,524301
7	ECHB_HUMAN	0,182622
8	ETFA_HUMAN	0,173785
9	FUMH_HUMAN	0,244477
10	HYES_HUMAN	-0,0854
11	IF2M_HUMAN	0,384389
12	IPYR2_HUMAN	0,431517
13	ITA7_HUMAN	-0,27688
14	LYRM4_HUMAN	0,10162
15	LYRM7_HUMAN	-0,26362
16	MIC60_HUMAN	-0,23859
17	NAR3_HUMAN	0,394699
18	NDUAA_HUMAN	0,298969
19	NDUB9_HUMAN	0,460973
20	NDUS1_HUMAN	0,60383
21	NDUS8_HUMAN	0,586157
22	NPS3A_HUMAN	-0,26362
23	OPA1_HUMAN	-0,44477
24	PRDX3_HUMAN	0,147276
25	PSMD9_HUMAN	0,521355
26	PYGB_HUMAN	-0,27982
27	RL10_HUMAN	-0,49926
28	RL15_HUMAN	-0,47717
29	RL21_HUMAN	-0,51105
30	RL26_HUMAN	-0,14728
31	RL3_HUMAN	-0,3703
32	SODM_HUMAN	0,321061
33	SPG7_HUMAN	-0,53745
34	SYDC_HUMAN	0,483064
35	VAPA_HUMAN	-0,40943
36	XPP1_HUMAN	-0,38439
37	ZADH2_HUMAN	0,132548

Furthermore, enrichment analysis shows that the proteins enrich biological functions related to cardiotoxicity:

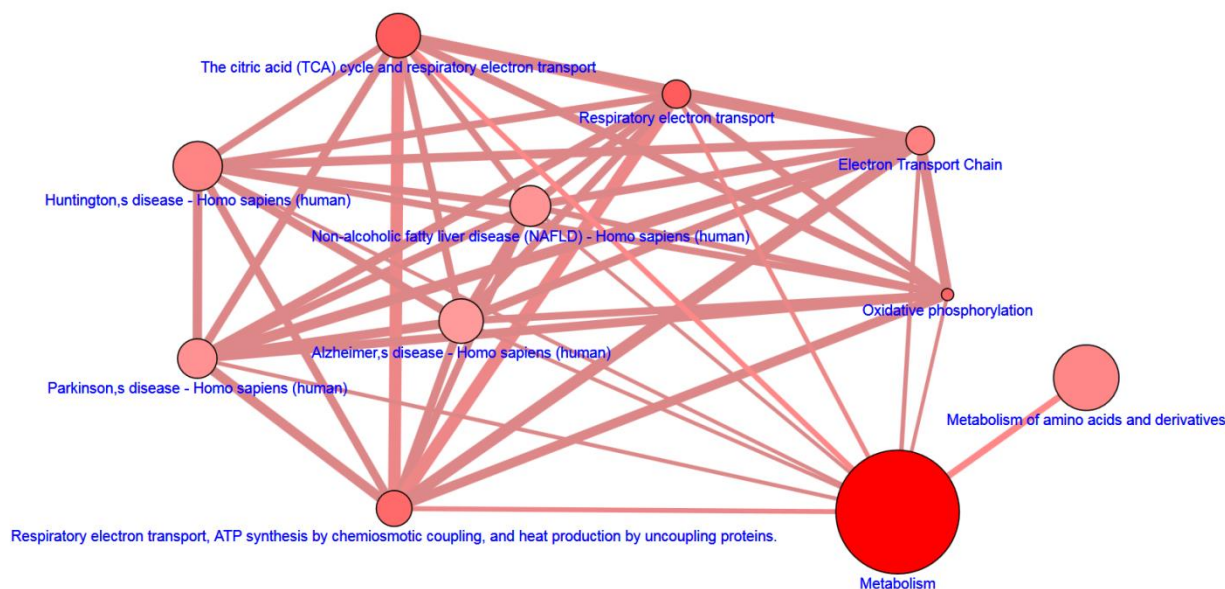


Fig.1: Pathways enriched by the most significant module of connected proteins computed by HotNet2. Pathways have been taken from KEGG, Reactome and WikiPathways.

Both approaches Hotnet2 and BioNet will be used in the course of the use cases that are conducted in the next reporting periods.

Differential equation model

In addition to the two published methods, Partner MPIMG has worked in the 3rd reporting period on an alternative network propagation model which is based on an epidemiological approach and models the effect of network perturbations by an SI (susceptible-infected) ODE model. Currently, the method is tested and compared to the alternative methods.

One motivation to this development was the fact that existing methods are heavily dependent on node degree as a central concept of the propagation methods. This leads to overweight of so-called hubs, i.e. well-studied or ubiquitous proteins with many interactions. For example, TP53 and EGFR are heavily studied leading to more than 1000 interaction partners in the ConsensusPathDB network. This leads to bias in the computation of the network modules such as star-like structured modules that are less informative. To avoid this problem, Partner MPIMG has adapted a model from epidemiology, called SI. Here, the “infected” nodes correspond to nodes that are experimentally altered (e.g. mutated proteins, differentially expressed proteins). These infected nodes can influence other nodes (the “susceptible” nodes) and turn them into an infected state. Thus, the resulting model describes a potential network propagation that is initiated by the effected proteins and that might identify network modules that are affected by this protein. The ODE has the following form:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI$$

Here, S is the population of susceptible nodes (i.e. not yet infected), I is the population of infected nodes and β is the rate at which the starting node affects its network neighbours. This parameter can be learned from the experimental results. The procedure is repeatedly done for all affected nodes and a post-processing step will then collect the most affected subnetwork.

Currently, the method is under development and work will be ongoing in the next reporting period (D12.4).

DIFFICULTIES

The Deliverable D12.3 was delayed by 9 months due to delays in data generation (WP6 and WP7) and to the test-case on anthracycline cardiac toxicity (4th reporting period); this use-case is still ongoing.

REFERENCES

- Yi S., Lin S., Li Y., Zhao W., Mills GB, Sahni N. Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nature Reviews Genetics* 18:395-410 (2017).
- Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, et al. Regulatory networks define phenotypic classes of human stem cell lines. *Nature* 455:401-405 (2008).
- Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JI, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 47:106-114 (2015).
- Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. *Nat Methods*, 12:615-621.
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102, 15545-15550 (2005).
- Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocol* 4, 44-56 (2009).
- Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat Protocol* 11:1889-1907 (2016).
- Hardt C, Bauer C, Schuchhardt J, Herwig R. Computational network analysis for drug toxicity prediction. *Methods Mol Biol*, in press.
- Beisser D, Klau GW, Dandekar T, Müller T, Dittrich MT. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, 26:1129-1130 (2010).

ANNEX 1

A scientific publication describing the pipeline implemented during preparation of D2.3 and D12.3 has been accepted for publication in Methods Molecular Biology.