Hepatocellular carcinoma computational models identify key protein-complexes associated to tumor progression

Maxime Folschette, Vincent Legagneux, Arnaud Poret, Carito Guziolowski, Nathalie Théret

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Hepatocellular carcinoma computational models identify key protein-complexes associated to tumor progression

Maxime Folschette [1,2,3,4], Vincent Legagneux [2], Arnaud Poret [4], Carito Guziolowski [4,5,*] and Nathalie Théret [1,2,*]

[1]Univ Rennes, Inria, CNRS, IRISA, UMR 6074, Rennes, France,
[2]Univ Rennes, Inserm, EHESP, Irset, UMR S1085, Rennes, France,
[3]IFB-CORE, Institut Français de Bioinformatique, UMS CNRS 3601, Évry, France,
[4]LS2N, Laboratoire des Sciences du Numérique de Nantes, UMR 6004, Nantes, France
[5]École centrale de Nantes, Nantes, France,

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Integrating genome-wide gene expression patient profiles with regulatory knowledge is a challenging task because of the inherent heterogeneity, noise and incompleteness of biological data. From the computational side, several solvers for logic programs are able to perform extremely well in decision problems for combinatorial search domains. The challenge then is how to process the biological knowledge in order to feed these solvers to win insights in a biological study. It requires formalizing the biological knowledge to give a precise interpretation of this information; currently, very few pathway databases offer this. The presented work proposes a workflow to generate novel computational predictions related to the state of expression or activity of biological molecules in the context of hepatocellular carcinoma (HCC) progression.

Results: Our working base is a graph of 3,383 nodes and 13,771 edges extracted from the KEGG database, in which we integrate 209 differentially expressed genes between *low* and *high aggressive* HCC across 294 patients. Our computational model predicts the shifts of expression of 146 initially non-observed biological components. Our predictions were validated at 88% using a larger experimental dataset and cross-validation techniques. In particular, we focus on the protein-complexes predictions and show for the first time that NFKB1/BCL-3 complexes are activated in aggressive HCC. In spite of the large dimension of the reconstructed models, our analyses over the computational predictions discover a well constrained region where KEGG regulatory knowledge constrains gene expression of several biomolecules. These regions can offer interesting windows to perturb experimentally such complex systems.

Availability: Data and scripts are freely available at `https://zenodo.org/record/2635752` and `https://github.com/arnaudporet/stream`

Contact: carito.guziolowski@ls2n.fr, nathalie.theret@univ-rennes1.fr

Supplementary information: Supplementary Material, Figures and Tables are available in appendix.

1 Introduction

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer, which counts for more than 800,000 deaths each year. The incidence of HCC is associated with the development of chronic

hepatitis mainly linked to viral infection, alcohol consumption and non-alcoholic fatty liver disease (NAFLD) (Global Burden of Disease Liver Cancer Collaboration *et al.*, 2017). Lifestyles (Saran *et al.*, 2016) and environmental pollution such as particulate matter air pollution (VoPham *et al.*, 2018) also contribute to increase burden in HCC worldwide. HCC is a heterogeneous disease and various genomic alterations associated with the etiologies and the stages of the pathology have been widely documented (Khemlina *et al.*, 2017; Schulze *et al.*, 2016). A pivotal step in the course of HCC progression is the epithelial-mesenchymal transition (EMT) which allows hepatocytes to transdifferenciate into mesenchymal phenotype whereby escaping to host control and acquiring anti-apoptotic and motility features (Giannelli *et al.*, 2016). Upregulation of EMT markers has been associated with tumor aggressiveness and bad prognosis (Kim *et al.*, 2010; Yamada *et al.*, 2014) and associated with inflammatory microenvironment (Yan *et al.*, 2018). However, in vivo monitoring of EMT processes remains difficult, due to the spatio-temporal dynamics of these molecular events and the snap-shot nature of biopsies sampling. Understanding EMT to identify new therapeutic targets require integrative and modeling approaches.

To build computational models and integrate experimental data on molecular events, pathway databases can be used. However, despite the fact that numerous publicly available pathway databases currently exist, compiling hundreds of signaling pathways for various species, very few formal representations linked with automatic inference processes have been proposed so far (Neaves *et al.*, 2018). The main difficulty appears to be the transfer from the biological representation of a pathway towards a logic knowledge base. Currently, pathway repositories, such as Reactome (Fabregat *et al.*, 2018), Pathway Commons (Cerami *et al.*, 2010), KEGG (Kanehisa *et al.*, 2017), or OmniPath (Turei *et al.*, 2016) propose their own tools to build graphs. Some of these tools are the Cytoscape (Shannon *et al.*, 2003) plugin CyPath2, PCViz for Pathways Commons; pypath for OmniPath; and ReactomeFIViz (Wu and Stein, 2012) for Reactome. However, the resultant graphs are difficult to be transferred into mechanistic models because the notion of causality is often misinterpreted. This misinterpretation is due to the lack of a formal causal representation of biochemical reactions such as protein-complexes assemblies. For instance tools such as CyPath2, PCViz, ReactomeFIViz, and pypath assume that there is a relation of causality between the protein-complex members (protein-complex members are the cause and consequence of each other); while in our modeling choice, protein-complexes may be triggering other reactions, and their presence is a consequence of the presence of their members. Knowing that signaling cascades are represented by multiple complexes assemblies, this misinterpretation impacts importantly the construction of a mechanistic model when using pathway databases. On the other hand, such tools are very useful to compute topological scores, perform statistical analyses, and to integrate gene expression measurements using enrichment analyses (Mi *et al.*, 2017). They remain, however, limited to extract logical consequences of the representation of the biological mechanisms.

The *sign-consistency* framework proposes a way to automatically confront the logic of large-scale interaction networks and genome-wide experimental measurements, provided that a signed oriented network is given and that the experimental measurements are discretized in 3 expression levels (up-regulated, down-regulated and no-change). This framework, introduced in Veber *et al.* (2004), has being applied to model middle- and large-scale regulatory and signaling networks. The two most recent implementations of it are by the means of integer linear programming (Melas *et al.*, 2013) and logic programming. The latter, implemented in a tool named *Iggy* (Thiele *et al.*, 2015), presents some key aspects: (i) it provides a global analysis applying a local rule which relates a node with its direct predecessors, (ii) it handles a network composed of thousands of components, (iii) it allows the integration of hundreds of measurements, (iv) it performs minimal corrections to restore the logic consistency, and (v) once the consistency is restored, it allows to infer the behaviour (up, down, no-change) of components in the network that were not experimentally measured. In this work we apply this sign-consistency framework to model the HCC progression.

Our case study is composed of two input data which were publicly available. First, gene expression data from patients with HCC were extracted from ICGC database (Hudson et al, 2010). Based on the EMT signature from MSigDB (Subramanian *et al.*, 2005), HCC samples were clustered into either agressive HCCs (high EMT gene expression) or non-agressive HCCs (low EMT gene expression). Second, the up-stream events of the regulatory events of these genes were obtained by querying automatically KEGG to build a causal model from this database. We used Iggy to study what are the regulatory events that explain the differential expression between low and high aggressiveness from the KEGG interaction knowledge (network of 3,383 nodes and 13,771 edges). We discovered that 146 nodes were predicted, of them 33 refer to gene expression, 110 were protein activities, and 3 were protein-complexes activities. 88% of the predictions were in agreement with the ICGC gene expression measurements. Importantly, we predicted the activation of NFKB1/BCL3 and NFKB2/RELB complexes, two critical regulators of NFKB signalling pathway implicated in tumorigenesis. Finally, we proposed a method to discover sensitive network regions that explains HCC progression. This means network components which were highly constrained by multiple experimental data points that could be interesting to target in order to obtain significant changes in the system behavior. We provide a list of 27 nodes discovered by this approach, including TP53.

2 Material & Methods

2.1 Differential Analysis

We set up a pilot study aiming at comparing gene expression in aggressive, versus less aggressive HCC. For this purpose, we used RNA-seq expression data available from the International Cancer Genome Consortium (International Cancer Genome Consortium *et al.*, 2010). Normalized HTseq counts and clinical data were retrieved from the LIHC-US project[1] (NCI, TCGA-LIHC). These files were downloaded on 2016-07-19, corresponding to release 21. At this date, LIHC-US dataset comprised 294 donors and 345 samples; among them, we selected samples corresponding to solid primary tumors, based on clinical data, by selecting entries containing the expression "`Primary tumour - solid tissue`" in the specimen table (7th field). This allowed selecting one sample for each of the 294 donors. Data retrieval and filtering workflow is detailed in Supplementary File `dataset_filtering.sh`.

From this filtered dataset, we extracted a two-dimensional table of expression values (converted in $\log_2$) for 20,502 genes in 294 LIHC samples. Based on the bimodal distribution of these expression values, we discarded genes whose expression is undetectable (4,220 genes), keeping 16,282 genes. Expression values were normalized by the median value in each sample. Based on the established link between epithelial-mesenchymal transition (EMT) and tumor aggressiveness (Thiery *et al.*, 2009), we used the MSigDB (Liberzon *et al.*, 2015) set of 200 genes termed `HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION`[2] from the Broad Institute as a molecular signature of aggressiveness. From the LIHC dataset, we extracted a table of expression values for 195 entries of this

[1] All ICGC data used in this work are publicly available at `https://dcc.icgc.org/releases/release_21/Projects/LIHC-US`

[2] Id: M5930, available at `http://software.broadinstitute.org/gsea/msigdb/geneset_page.jsp?geneSetName=HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION`

EMT signature for each of the samples (5 genes were undetectable). Based on the expression values of the EMT signature, LIHC samples were classified (hierarchical clustering of euclidean distances) into three groups termed *low_EMT* (70 samples), *medium_EMT* (154 samples), and *high_EMT* (70 samples). The result of this clustering analysis is available in Supplementary Figure 1. Samples corresponding to the *medium_EMT* group were discarded and a differential expression analysis was performed by computing a non-parametric Mann-Whitney test for all the 16,282 genes between the *low_EMT* and *high_EMT* groups. p-values were adjusted for multiple analyses by the Benjamini & Hochberg method. The volcano plot of Supplementary Figure 2 represents fold-changes ($\log_2$) against adjusted p-values ($-\log_{10}$). We focused on genes with an adjusted p-value below 10^{-5}. Genes with a $\log_2(\text{fold-change})$ greater than 2 were considered as over-expressed (821 genes), whereas those with a $\log_2(\text{fold-change})$ lower than 0.5 were considered as under-expressed (1,092 genes). These 1913 differentially expressed genes, listed in Supplementary File `diffexp_filtered.csv` were subsequently used to extract a regulatory network, as explained in Section 2.2, and then used as observations for the coloring propagation process, as detailed in Section 2.3. The workflow of data clustering and differential analysis is available in Supplementary File `diffexp_and_clustering.R`.

2.2 Building the signaling network from the KEGG Pathway database

This work was performed on a human signaling network derived from the KEGG Pathway database (Kanehisa *et al.*, 2017). Human signaling pathways were fetched using the KEGG API and converted to SIF (Simple Interaction Format). This section summarizes how this network was built. A more in-depth description is available as a Supplementary Material & Methods.

2.2.1 Distinguishing genes and their products

Because the data are about gene expressions, a clear distinction was made between nodes representing genes and nodes representing proteins. In the KEGG Pathway database, this distinction is implicitly embedded in the relation types, particularly `PPrel` edges (protein-protein relations) and `GErel` edges (gene expression relations).

`PPrel` edges indicate that both source and target nodes are proteins. `GErel` edges indicate that source nodes are transcription factors and that target nodes are genes. Therefore, to explicitly differentiate genes and proteins, the source nodes of `GErel` edges were suffixed with `_prot` and the target nodes were suffixed with `_gen`. Concerning `PPrel` edges, both the source and target nodes were suffixed with `_prot`.

Furthermore, in order to link genes and their products, a relation type was added: the `GPrel` type (gene-protein relations). For each node modeling a gene, a `GPrel` edge starting from it and ending on an added node suffixed with `_prot` was added. These added nodes therefore model the corresponding gene products while the `GPrel` edges model the gene expressions themselves.

Altogether, the human signaling network can now explicitly model protein-protein interactions, gene expression regulations and gene expressions themselves, as illustrated in Supplementary Figure 3.

Finally, in order to allow the data to match their corresponding nodes in the network, a `GPrel` edge was added for each node. Except when already done due to a `GErel` edge, each node implicitly modeling a protein was put as target of a `GPrel` edge with the suffix `_prot`. By doing so, a source node was added for each of these `GPrel` edges with the suffix `_gen`: these are the corresponding genes, as illustrated in Supplementary Figure 4.

2.2.2 Selecting functional interactions

As explained in Section 2.3, a next step consists in running the predictive tool Iggy (Thiele *et al.*, 2015) in order to infer the state of unobserved nodes, namely the genes devoid of data and all the proteins. To infer the state of a node, Iggy uses the state of its successors together with the sign of the edges linking them: edges have to be signed.

To do so, in the KEGG Pathway database and in addition to their relation types, the edges are annotated with keywords bringing details about the modeled interactions. Therefore, edge signs were inferred using these keywords.

Note that using this approach on a network where nodes modeling genes and nodes modeling proteins are distinct allows to predict protein activities from gene data. It can be insightful because gene expression does not systematically imply protein activity.

2.2.3 Extracting regulatory signaling pathways

Once the signed human signaling network obtained, we observe that only 209 genes, listed in Supplementary Table 1, are found in KEGG, from the initial list of 1913 differentially expressed genes between aggressive and non-aggressive tumors. Only the signaling pathways regulating these differentially expressed genes were extracted, that is, the upstream paths of the nodes modeling these differentially expressed genes. A tool named *Stream*[3], especially designed for that purpose, has been developed and used, providing the regulatory pathways on which predictions about nodes activity were performed.

2.3 Principle of the Predictions with Iggy

The information provided by the differential analysis of Section 2.1, regarding over-expression and under-expression of some genes, can be directly mapped on the related nodes of the graph obtained in Section 2.2. This information can be regarded as a partial *coloring* of the nodes, that consists of attaching an information to some nodes about their change of expression between the non-aggressive and aggressive stages of the tumor. Here the differential analysis provides two types of colorings: "$+$" for over-expressed genes and "$-$" for under-expressed genes. However, we will also consider the "0" coloring assessing that there is no change in the expression of a component, and allow all kinds of nodes (genes, proteins and complexes) in the graph to accept such colorings.

In the following section we will describe how new colorings are assigned from the existing ones (given by the experimental data) when the topology of the graph allows it, that is, when it leaves no ambiguity. This method was implemented using the Iggy tool (Thiele *et al.*, 2015). We will apply Iggy to obtain the results of this study, given in Section 3.2.

2.3.1 Principle of the coloring propagation

As explained above, we aim at assigning three kinds of colorings to nodes in the graph: $+$ (over-expression), $-$ (under-expression) and 0 (no change). The results from Sections 2.2 and 2.1 provide us 209 of such assignments, consisting in a "$+$" for each over-expressed and "$-$" for each under-expressed gene that is found in the KEGG network. These assignments are called *observations* as they correspond to experimental results. Iggy enumerates all possible colorings of all nodes and filters out those that do not respect the following criteria:

- The observations must keep their initial colorings.
- Each coloring $+$ or $-$ must be justified by at least one predecessor.
- Each coloring 0 must have only predecessors colored as 0 or a couple of $+$ and $-$ colored predecessors.

[3] Available at: `https://github.com/arnaudporet/stream`

When the colorings are compatible with the criteria above, given a set of observations, then in general we obtain many colorings. However, some nodes are colored the same across all colorings; these are then called *predictions* because their coloring is certain.

In the case that some observations are not compatible, that is, there are some observations which only generate colorings that invalidate the sign-consistency criteria presented before, we obtain a conflict. One way to fix such conflicts is to add artificial interactions in the network. Iggy allows to add a minimal number of such repairs, called *minimal correction set* (MCoS, see Thiele *et al.* (2015)). If several possibilities of repairs are possible, Iggy will compute them all and the final set of predictions will correspond to the union of the predictions obtained after each possible repair.

The workflow to call Iggy is given in Supplementary File `run-iggy.sh`.

2.4 Computational Validation of the results

In order to create the sets of genes of interest in Section 2.1, we used thresholds of $+2$ and -0.5 on the value of $\log_2(\text{fold-change})$. In this section, we aim at checking if these thresholds are justified. To do this, we computed "sub-predictions", that is, predictions on the same extracted graph of Section 2.2 but with subsets of observations. To generate these subsets of observations, we considered a range of samplings, from 10% to 95% of the complete observation set, with a step of 5%. For each sampling of $x\%$, 100 experiments were conducted, where an experiment consisted in randomly picking $x\%$ of the over-expressed observations $(+)$ and $x\%$ of the under-expressed observations $(-)$, and computing the predictions on this subset of observations. The results are 1,800 such subsets of observations, and as many computed sets of predictions on the nodes of the graph, hereafter called *sub-predictions*. These sub-predictions have been exploited in two ways:

1. by comparing said sub-predictions with the available gene expression data from ICGC that were already used for the differential analysis (Section 2.4.1), and
2. by comparing said sub-predictions with the final predictions obtained with 100% of observations to witness their variability (Section 2.4.2).

Both approaches are explained below and performed by Supplementary File `run-validation.sh`.

2.4.1 Recovery Rate of the Sub-predictions

We computed a normalized score by counting the number of predictions matching the related experimental fold-change from the ICGC data. For each experiment result, this score s is given by the formula: $s = m/t$ where m is the number of matching predictions, that is, positive predictions with positive fold-changes and negative predictions with negative fold-changes, and t is the total number of predictions. This allows us to assess the ability of our model to recover from missing information (here, observations).

2.4.2 Stability of the sub-predictions

In order to look at the stability of the predictions made on subsets of observations, we also compared them to the final predictions using 100% of the observations. For each predicted node in the 100% sampling set, and for each of its corresponding sub-prediction in a lower sampling set:

- If the node is predicted and the prediction matches the one at 100% sampling, this is considered a "good" prediction.
- If the node is predicted but the prediction is not the same as for 100% sampling, this is considered a "bad" prediction, thus representing mathematical non-monotonicity and biological sensitive components or potential targets.
- If the node is not predicted, this is called a "missing" prediction.

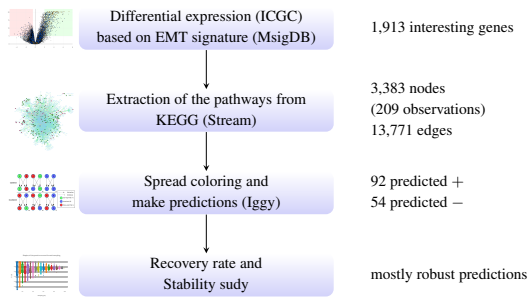

Fig. 1. Complete workflow of the method presented in this paper

Counting the elements and observing the evolution of these categories allows us to witness if lower samplings converge to the final sampling or not, independently of any exterior data such as expression data.

2.5 Summary of the Workflow

Based on the material and methods presented in this section, the graph of Figure 1 sums up the complete workflow that will be used and which results are presented in the following.

3 Results

3.1 Integration of ICGC Gene Expression in Signaling and Regulatory Network

The KEGG graph described in Section 2.2 is too large to be handled by our tools on a standard computer, as it contains a lot of components which are not involved in the mechanisms studied here. Therefore, we extracted the subgraph composed of the upstream regulators of the 209 differentially expressed genes obtained in Section 2.1 and listed in Supplementary Table 1. In practise, it consists in filtering out all components (genes, proteins, complexes) that are not upstream of this list of up- and down-regulated genes. We also filtered out the 4,220 genes whose expression is undetectable, as mentioned in Section 2.1, along with proteins and complexes containing the product of one of these genes. This extraction was implemented in the Stream tool (see Section 2.2). The final extracted graph is about a quarter of the size of the initial graph, as it contains 13,771 interactions and 3,383 components. Among them, we can find 209 of the initial 1,913 genes, meaning that 1,704 were not found in the initial KEGG graph. This is summarized in Table 1 below and the final graph is depicted in Supplementary Figure 5, and available in Supplementary File `graph.sif` in SIF format and in Supplementary File `graph.cys` as a Cytoscape session.

	Nodes	**Edges**
a) KEGG extraction	8,861	41,546
b) Queried genes extraction	3,383	13,771

Table 1. Statistics on the graph obtained a) by the extraction of sections 2, 3, 4 and 5 of KEGG, and b) after filtering out the non-upstream regulators of the 209 differentially expressed genes along with genes of undetectable expression.

The final graph contains mostly activations (11,661 versus 2,110 inhibitions) which follows the same observations as the labeled edges of

KEGG. Only 209 nodes have observations attached to them, provided by the differential analysis of Section 2.1, leaving most nodes unobserved and subject to computational predictions. Finally, the presence of nodes gathering a lot of incoming or outgoing interactions is noteworthy:

- The biggest in-degree is 92 (concerning nodes `PIK3R6_prot`, `PIK3CG_prot` and `PIK3R5_prot`);
- The biggest out-degree is 79 (concerning nodes `PRKACB_prot`, `PRKACA_prot`);
- The two nodes `MAPK3_prot` and `MAPK1_prot` both have the maximal total degree of 107, with 56 incoming and 51 outgoing interactions.

Such "hub" nodes, having an influence to and from a lot of other components, have a high impact on the rest of the network and produce less consensual colorings.

3.2 Computational Predictions & ICGC Validation

Launching Iggy on the graph of Section 2.2 and the observations derived from the differential analysis of Section 2.1 returns 146 predictions, among which:

- 92 are over-expressions (+),
- 54 are under-expressions (−),
- none of them is a no-change (0).

The list of all predictions is given in Supplementary Table 2 and plotted on on the KEGG graph in Supplementary Figure 6 and on the volcano plot of gene differential expression in Supplementary Figure 7. Furthermore, Iggy computes one minimal correction set (MCoS) on the graph because the observation data is slightly inconsistent: an influence from an unknown node is added on `PMAIP1_gen` to restore consistency, as shown in Supplementary Figure 8. In the end, 3,026 nodes remain not observed nor predicted. Iggy takes one minute to compute these results on a standard laptop computer[4].

3.2.1 Comparison with expression data

Most of the predictions produced in Section 3.2 can be compared with the result of the differential analysis computed in Section 2.1, depending on the type of the node predicted:

- The 33 predicted genes can be directly compared to the corresponding fold-change, which is based on an expression analysis.
- The 110 predicted proteins can be compared to the fold-change of the corresponding gene under the assumption that protein production is correlated to gene expression.
- The 3 predictions on complexes, however, were not compared at this point to the gene expression data, but will be discussed in detail in Section 3.4.

Such comparison gives us clues about the quality of the predictions. It can be observed on the volcano plot of gene differential expression in Supplementary Figure 7, and is also depicted on the KEGG graph in Supplementary Figure 9. Among the 146 predictions, 143 have a name that is found in the ICGC data (but was not selected as the initial list of over- and under-expressed genes). If we remove all threshold and thus consider any positive fold-change as an over-expression, and any negative fold-change as an under-expression, then 82 components predicted + are coherent with the ICGC data and 8 are not; 44 components predicted − are coherent with the ICGC data and 9 are not. This ratio of 88% of matching predictions speaks in favor of our choice of applying Iggy to this specific biological system, with respect to the currently available data in KEGG and ICGC databases. The list of predictions not matching with experimental expression data is given in Supplementary Table 3.

[4] Laptop computer containing an Intel Core i7-5600U CPU with 4 threads of 2.60GHz and running Fedora 27 64 bits.

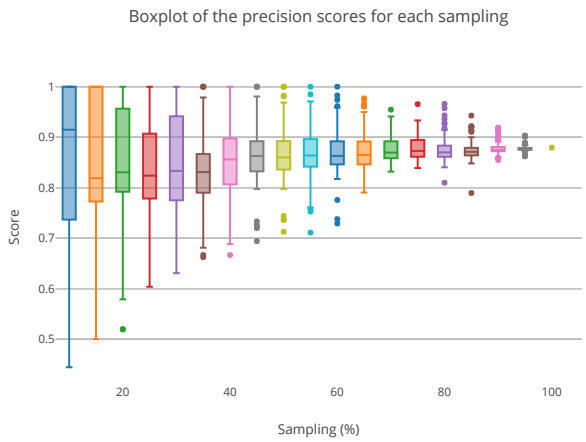

Fig. 2. Boxplots of the precision scores (ordinate) of the predictions obtained with randomly picked samplings (abscissa) of observations. Each box plot at abscissa x represents the scores of the 100 sub-predictions obtained by randomly picking $x\%$ of the observations. The point at 100% represents the score of the predictions with the complete set of observations.

3.3 Impact of data Incompleteness on Computational Predictions

This section presents the results of the two robustness analyses applied on the sampling of observations described in Section 2.4. The objective is to observe the impact of data incompleteness in our computational predictions. For this, we observed and tracked across the samples the level of precision and the quality of the information contained in the predictions.

3.3.1 Recovery Rate

The first approach (see Section 2.4.1) consists in comparing the predictions from the different samplings of observations to the available expression data by using the same dataset that was used to produce the lists of genes of interests. The plot of Figure 2 shows the box-and-whiskers diagrams corresponding to the scores of all experiments. We can observe a clear convergence of these scores towards the final score of 0.88 corresponding to the 100% sampling, which shows that our complete predictions do not lie in a local extremum.

3.3.2 Stability Study

The second approach (see Section 2.4.2) consists in observing "good", "bad" and "missing" predictions for each of the experiments (samplings $< 100\%$) compared to the 100% sampling. Figure 3 computes the minimum, maximum, median and mean of each such category. Globally, we can observe that the mean and median number of "bad" predictions, that is, predictions that are different with a subset of observations than with the complete set of observations, are really low, below 4% for all samplings. Nevertheless, some samplings show a high proportion of such "bad" predictions. Moreover, the number of "missing" predictions is very high for low samplings, which assesses that there is too little information to obtain complete results. Overall, "bad" predictions tend to decrease after the 65% sampling, along with "missing" predictions that decrease all the way, making "good" predictions mathematically increase.

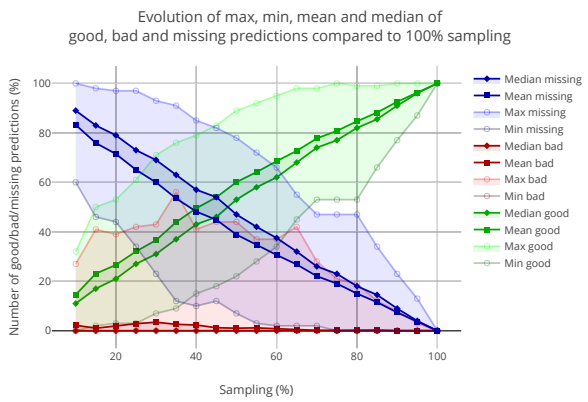

Fig. 3. Stability of the predictions for subsets of the observations, compared to the final predictions with all 100% of observations, for all samplings. "Good" predictions (matching the 100% predictions) are depicted in green, "Bad" predictions (predicted differently than the 100% predictions) in red and "Missing" predictions (not predicted) in blue. For each category, four curves are plotted representing, from top to bottom, the maximum, median, mean and minimum number of predictions of this type. Curves are normalized to the number of predictions of each "sub-prediction".

3.3.3 Insights of the Stability Results

The analyses of the experiments shown in the previous subsections show that the "badly" predicted components for subsets of observations are always the same 28 nodes, listed in Supplementary Table 4. These nodes belong to the same region of the graph, which is depicted in Supplementary Figure 8. Actually, a group of 27 of these nodes are strongly linked and always change their coloring together. When searching inside the graph topology, one can remark that this group is tightly linked to the node `TP53_prot`, which is also part of the group. This protein acts as a "hub" inside the graph, having a high degree (25 ingoing and 28 outgoing edges). It therefore controls closely, if not directly, a lot of other components that change their sign as soon as it does so, rendering the whole group of predictions unstable. The reason of this instability is that `TP53_prot` directly influences node `PMAIP1_gen` which is involved in the only MCoS repair in our graph: the node `PMAIP1_gen` is indeed observed as over-expressed (+) but 3 other under-expressed (−) observations contradict this one: `CCNG1_gen`, `SHISA5_gen` and `TP73_gen`. This leads to an inconsistency, as explained in Section 3.2. The repair here consists in adding an edge towards `PMAIP1_gen` that models missing information, in order to remove this inconsistency, as shown in Supplementary Figure 8. In practise, this renders `PMAIP1_gen` "silent" regarding `TP53_prot`, which then takes the coloring of the other observations (under-expression). Nevertheless, when picking random sets of observations, we sometimes fall in cases where among these 4 observations, only `PMAIP1_gen` is selected; in this case, no repair is needed and `TP53_prot` is predicted as over-expressed, also leading to 26 different predictions in downstream nodes.

Finally, the last unstable node is `PMAIP1_prot`: in the case where `PMAIP1_gen` is part of the randomly picked observations, it is straightforwardly predicted over-expressed while in the converse case, where `PMAIP1_gen` is not part of the observations, it is indirectly influenced by `TP53_prot` and thus predicted under-expressed.

Such unstable predictions can be regarded as not very robust because they are changeable depending on the number of observations taken into account. On the other hand, all other predicted components are stable and can be considered as robust since, when they are predicted, their prediction matches the one obtained using all the observations. The list of stable and unstable predictions is given in Supplementary Table 4.

3.4 Biological Validation of the Computational Results

Among the computational predictions given in Section 3.2, some of them are of particular interest in regard to the expression data from ICGC. In this section, we detail and validate them biologically.

3.4.1 Activation of NFκB signaling in aggressive HCC

Based on the regulatory model (see Supplementary Figure 5) and differential expression of mRNA between low and high aggressive HCC (see Section 2.1), the algorithm Iggy predicts the activation of complexes NFKB1::BCL3 and NFKB2::RELB and the deactivation of complex JUND::NACA. This is a novel information since it was not present in the initial experimental data of gene expression.

Among them, two complexes are related to NFκB signaling and are predicted as activated: NFKB1::BCL3 and NFKB2::RELB. NFKB1, NFKB2 and RELB are three subunits of the transcription factor complex nuclear factor-kappa-B (NFκB) which consist in a homo- or heterodimeric complex formed by Rel-like domain-containing proteins p65 (RelA), RelB, c-Rel, p50 (NFKB1), and p52 (NFKB2). The NFκB signaling system acts through canonical and non canonical pathways which are induced by different extracellular signals (Shih *et al.*, 2011). The canonical pathway can be induced by TNF-α, IL-1 or LPS stimulation and requires NF-kappa-B essential modulator (NEMO) while the non-canonical pathway is induced by other ligands such as CD40 ligand (CD40L), receptor activator of nuclear factor kappa-B ligand (RANKL), B-cell activating factor (BAFF) and lymphotoxin beta (LTb). Upon ligand binding to its receptor, the signaling cascades control the degradation of IkB proteins (inhibitor of NFκB) and precursor processing including NFKB1 (p105) and NFKB2 (p100) which are proteolytically activated to p50 and p52 respectively. B-cell chronic lymphatic leukemia protein 3 (Bcl3) is a member of IkB family that are inhibitors of NFκB members. BCL3 associates with NF-kappa B in the cytoplasm and prevents nuclear translocation of the NFKB1 (p50) subunit. When phosphorylated, BLC3 is activated and associates with NFKB1 in the nucleus to regulate NFκB target genes (Wang *et al.*, 2017). NFκB system is involved in the regulation of numerous biological processes including inflammation, cell survival and development. Regarded as protective against aggression from environment in normal physiology, alteration of NFκB signaling pathways has been associated with various diseases such as inflammatory disease and cancer (Concetti and Wilson, 2018; Cildir *et al.*, 2016). In HCC, NFκB pathway was shown to be deregulated in tumor and underlying fibrotic livers (Tai *et al.*, 2000; Yokoo *et al.*, 2011). Notably, increased expression of p50 and BCL3 has been reported in tumors compared with adjacent tissues (O'Neil *et al.*, 2007) and p50 expression was associated with early recurrence of HCC (Yokoo *et al.*, 2011).

In order to evaluate our predictions about the activation of NFKB1::BCL3 and NFKB2::RELB complexes, we thought to search for expression of genes regulated by these complexes. For that purpose, we take advantage of the NFκB-dependent signature available in MSigDB (Subramanian *et al.*, 2005; Liberzon *et al.*, 2011). We selected the `HALLMARK_TNFA_SIGNALING_VIA_NFKB`[5] signature which contains 200 genes regulated by NFκB in response to TNF. As shown in Supplementary Figure 10A, we demonstrated that these genes were more expressed in high aggressive HCC when compared with low aggressive ones supporting the activation of NFκB signaling. More

[5] Id: M5890, available at `http://software.broadinstitute.org/gsea/msigdb/geneset_page.jsp?geneSetName=HALLMARK_TNFA_SIGNALING_VIA_NFKB`

specifically, we searched for expression of genes targeted by NFκB-non-canonical pathway, including the cytokines CCL19 and CCL21. These genes are regulated through the activation of NFKB2::RELB complexes and their expression was increase in high aggressive HCC thereby confirming the prediction (Supplementary Figure 10B).

Another prediction was the down-regulation of JUND::NACA complex that was previously demonstrated to regulate osteocalcin (Akhouayri *et al.*, 2005). This prediction is mainly conditioned by osteocalcin (BGLAP) expression data that was found down-regulated in the aggressive HCC (-1.3 fold-change between aggressive versus non-aggressive HCC). Such observations are in accordance with previous reports showing that osteocalcin was down-regulated in the serum of HCC patients when compared with healthy controls (Liu *et al.*, 2015). As shown in Supplementary Figure 11A, we showed that both JUND and NACA gene expressions were down-regulated in aggressive HCC supporting the prediction of down-regulation of the complexes JUND::NACA. Importantly, the targets of JUND::NACA complex including LRP5 and LRP6 genes were predicted as down-regulated by our model (see Supplementary Figure 12). The down-regulation of LRP5 in aggressive HCC was validated in HCC data but was not significant for LRP6 probably due to the low level of gene expression (Supplementary Figure 11B). According with this, the up-regulation of LRP6 through JUND::NACA complexes was clearly demonstrated in osteoblasts (Pellicelli *et al.*, 2018).

To conclude, model predictions were validated by data analyses and are in accordance with the literature. This is the first report describing the activation of NFKB2::RELB complex and the down-regulation of JUND::PACA complex in aggressive HCC.

4 Discussion & Conclusion

The understanding of tumor progression dynamics is extremely difficult when considering the snap-shot nature of data from patients. However, compiling information from a wide spectrum of tissue samples can be used for modeling evolutive stories. The complexity of molecular events implicated in hepatocellular carcinoma progression is directly associated with its various etiologies that differently contribute to tumor initiation, growth and evasion. During last decades, multiscale omics data analysis of genome and proteome allowed to explore molecular networks associated with HCC and mathematical models have been developed namely to predict cancer cell behavior (D'Alessandro *et al.*, 2013). Accordingly, an elegant discrete model was developed by Steinway *et al.* (2014) to explore TGF-β signaling pathway during epithelio-mesenchymal transition in HCC. However, HCC results from complex interactions between the tumor cells and the microenvironment involving stromal cells and extracellular matrix. Molecular biological data from tumor tissues recapitulate all this information and we need to build an unique large-scale model without *a priori* to take into account such complexity. For that purpose, we propose here an original approach aiming at integrating experimental data on a regulatory graph extracted from the KEGG database to predict new markers and regulators of HCC progression.

Based on EMT gene expression signature from MSigDB (Subramanian *et al.*, 2005) we first separated low from high aggressive HCC samples stored in the ICGC database (International Cancer Genome Consortium *et al.*, 2010) and next we sought to predict the regulatory pathways implicated in this transition. For that purpose, we built a model by querying the KEGG database using the KEGG API to extract an initial network. We have implemented a tool, Stream, to allow us extracting a directed and signed sub-network, from the previously obtained network, by using the up-stream events of a list of *target genes*. Importantly, our modeling choices allowed us to connect protein complexes to their members, and to label network nodes of type *gene* and *protein*. This separation of concepts is particularly valuable when modeling gene expression.

The publicly available knowledge base KEGG, gathering curated signaling and regulatory processes, is well structured to automatically extract mechanistic models from it. In particular: (i) the information concerning gene transcription and signaling modifications is differentiated, (ii) the network nodes identifiers are unique, and (iii) the biological processes, such as phophorylation or gene-regulation, are clearly represented.

Using Iggy, it was possible to confront the logic of a large-scale KEGG network (3,383 nodes, 13,771 edges) to the expression of genes differentially expressed between aggressive and non-aggressive HCC. In this context, we were able to propose an integrated model of HCC progression and to predict the regulation of new biomolecules including genes, proteins and complexes. A major finding is that the model predicted the behavior of 146 network components that were associated with the progression of tumors. 88% of the computation model predictions were validated with the ICGC data-set and by using cross-validation techniques, thereby demonstrating the quality of the model. Conversely, 12% of the predictions did not match the experimental data, however 10 of these components are part of gene/protein couples leading to linked predictions. In addition, all of these components but one had a low expression change (less than 1 in absolute value) along with a high p-value (above 10^{-2}) that might explain the inconsistency. The remaining one is `THBS1_gen` (thrombospondin 1 gene) with a fold-change of 1.996, and is also part of the cluster of unstable predictions depicted in Section 3.3.3. Indeed, we discovered a subset of 28 network nodes that were very sensitive to the experimental data. That is, they were strongly constrained by a subset of experimental observations. We notice that these nodes behave as hubs in the network, and can be candidate to experimental stimulation or inhibition in order to affect the system behavior.

The most interesting prediction was the activation of protein complexes related to NFκB signaling since complexes formation is directly responsible for signal transduction (O'Dea and Hoffmann, 2010). While the role of NFκB signaling pathway has been widely documented in chronic liver disease (Luedde and Schwabe, 2011), the activation of NFKB1/BCL-3 complexes in aggressive HCC has never been reported. The IκB protein BCL-3 acts both as a co-activator that form complexes with NFKB1(p50) dimers to promote genes (Chang and Vancurova, 2014) and as a co-repressor of gene transcription by stabilizing P50 homodimers on DNA promoters (Collins *et al.*, 2014). Predicted activation of such complexes in aggressive HCC revealed the ambivalent role of NFKB-mediated inflammatory response during the course of tumor progression (Seki and Brenner, 2007).

The present study is general to be applied to other biological data from cancers or other disease. In the future, we would like to use logic programming to target the combinatorics of sensitive regions in a regulatory graph with respect to gene expression profiles, in order to propose regulatory elements for clinical therapy. Another perspective is to apply our method to subsets of patients, and observe if there are clusters of patients that have specific computational model signatures for HCC progression.

Acknowledgements

The authors wish to thank Anne Siegel for her fruitful discussions and comments. The authors also acknowledge the Université Bretagne Loire for the funding, and the GenOuest bioinformatics core facility[6] for providing the computing infrastructure.

[6] `https://www.genouest.org`

References

Akhouayri, O. *et al.* (2005). Sequence-specific DNA binding by the alphaNAC coactivator is required for potentiation of c-Jun-dependent transcription of the osteocalcin gene. *Mol. Cell. Biol.*, **25**(9), 3452–60.

Cerami, E. G. *et al.* (2010). Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, **39**(suppl_1), D685–D690.

Chang, T.-P. and Vancurova, I. (2014). Bcl3 regulates pro-survival and pro-inflammatory gene expression in cutaneous T-cell lymphoma. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1843**(11), 2620 – 2630.

Cildir, G. *et al.* (2016). Noncanonical NF-κB Signaling in Health and Disease. *Trends Mol Med*, **22**(5), 414–429.

Collins, P. E. *et al.* (2014). Inhibition of Transcription by B Cell Leukemia 3 (Bcl-3) Protein Requires Interaction with Nuclear Factor κB (NF-κB) p50. *Journal of Biological Chemistry*, **289**(10), 7059–7067.

Concetti, J. and Wilson, C. L. (2018). NFKB1 and Cancer: Friend or Foe? *Cells*, **7**(9).

D'Alessandro, L. *et al.* (2013). Hepatocellular carcinoma: a systems biology perspective. *Frontiers in Physiology*, **4**, 28.

Fabregat, A. *et al.* (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**(D1), D649–D655.

Giannelli, G. *et al.* (2016). Role of epithelial to mesenchymal transition in hepatocellular carcinoma. *J. Hepatol.*, **65**(4), 798–808.

Global Burden of Disease Liver Cancer Collaboration, Akinyemiju, T., *et al.* (2017). The Burden of Primary Liver Cancer and Underlying Etiologies From 1990 to 2015 at the Global, Regional, and National Level: Results From the Global Burden of Disease Study 2015. *JAMA Oncol*, **3**(12), 1683–1691.

International Cancer Genome Consortium, Hudson, T. J., *et al.* (2010). International network of cancer genome projects. *Nature*, **464**, 993.

Kanehisa, M. *et al.* (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**(D1), D353–D361.

Khemlina, G. *et al.* (2017). The biology of Hepatocellular carcinoma: implications for genomic and immune therapies. *Mol. Cancer*, **16**(1), 149.

Kim, J. *et al.* (2010). Epithelial-mesenchymal transition gene signature to predict clinical outcome of hepatocellular carcinoma. *Cancer Sci.*, **101**(6), 1521–8.

Liberzon, A. *et al.* (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**(12), 1739–40.

Liberzon, A. *et al.* (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**(6), 417–425.

Liu, Y. *et al.* (2015). Relationships between the Osteocalcin gene polymorphisms, serum osteocalcin levels, and hepatitis B virus-related hepatocellular carcinoma in a Chinese population. *PLoS ONE*, **10**(1), e0116479.

Luedde, T. and Schwabe, R. F. (2011). NF-κB in the liver—linking injury, fibrosis and hepatocellular carcinoma. *Nature Reviews Gastroenterology & Hepatology*, **8**, 108–118.

Melas, I. N. *et al.* (2013). Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Comput. Biol.*, **9**(9), e1003204.

Mi, H. *et al.* (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**(D1), D183–D189.

Neaves, S. R. *et al.* (2018). Reactome Pengine: a web-logic API to the Homo sapiens reactome. *Bioinformatics*, **34**(16), 2856–2858.

O'Dea, E. and Hoffmann, A. (2010). The Regulatory Logic of the NF-κB Signaling System. *Cold Spring Harbor Perspectives in Biology*, **2**(1).

O'Neil, B. H. *et al.* (2007). Expression of nuclear factor-kappaB family proteins in hepatocellular carcinomas. *Oncology*, **72**(1-2), 97–104.

Pellicelli, M. *et al.* (2018). Lrp6 is a target of the PTH-activated αNAC transcriptional coregulator. *Biochim Biophys Acta Gene Regul Mech*, **1861**(2), 61–71.

Saran, U. *et al.* (2016). Hepatocellular carcinoma and lifestyles. *J. Hepatol.*, **64**(1), 203–14.

Schulze, K. *et al.* (2016). Genetic profiling of hepatocellular carcinoma using next-generation sequencing. *J. Hepatol.*, **65**(5), 1031–1042.

Seki, E. and Brenner, D. A. (2007). The role of NF-κB in hepatocarcinogenesis: Promoter or suppressor? *Journal of Hepatology*, **47**(2), 307–309.

Shannon, P. *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**(11), 2498–2504.

Shih, V. F.-S. *et al.* (2011). A single NFκB system for both canonical and non-canonical signaling. *Cell Res.*, **21**(1), 86–102.

Steinway, S. N. *et al.* (2014). Network Modeling of TGFβ Signaling in Hepatocellular Carcinoma Epithelial-to-Mesenchymal Transition Reveals Joint Sonic Hedgehog and Wnt Pathway Activation. *Cancer Research*, **74**(21), 5963–5977.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(43), 15545–50.

Tai, D. I. *et al.* (2000). Constitutive activation of nuclear factor kappaB in hepatocellular carcinoma. *Cancer*, **89**(11), 2274–81.

Thiele, S. *et al.* (2015). Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. *BMC Bioinformatics*, **16**(1).

Thiery, J. P. *et al.* (2009). Epithelial-mesenchymal transitions in development and disease. *Cell*, **139**(5), 871–90.

Turei, D. *et al.* (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, **13**(12), 966–967.

Veber, P. *et al.* (2004). Complex qualitative models in biology: A new approach. *ComPlexUs*, **2**, 140–151.

VoPham, T. *et al.* (2018). Ambient PM2.5 air pollution exposure and hepatocellular carcinoma incidence in the United States. *Cancer Causes Control*, **29**(6), 563–572.

Wang, V. Y.-F. *et al.* (2017). Bcl3 Phosphorylation by Akt, Erk2, and IKK Is Required for Its Transcriptional Activity. *Mol. Cell*, **67**(3), 484–497.e5.

Wu, G. and Stein, L. (2012). A network module-based method for identifying cancer prognostic signatures. *Genome Biol.*, **13**(12), R112.

Yamada, S. *et al.* (2014). Epithelial to mesenchymal transition is associated with shorter disease-free survival in hepatocellular carcinoma. *Ann. Surg. Oncol.*, **21**(12), 3882–90.

Yan, L. *et al.* (2018). Relationship between epithelial-to-mesenchymal transition and the inflammatory microenvironment of hepatocellular carcinoma. *J. Exp. Clin. Cancer Res.*, **37**(1), 203.

Yokoo, H. *et al.* (2011). Clinicopathological significance of nuclear factor-κB activation in hepatocellular carcinoma. *Hepatol. Res.*, **41**(3), 240–9.