Search of Therapeutic Targets in Metabolic Pathways of TGF-β Using Graph Coloring Approaches

Maxime Folschette[1, 2], Anne Siegel[1], Vincent Legagneux[2], Carito Guziolowski[3], Nathalie Théret[1, 2]

[1]Univ Rennes, Inria, CNRS, IRISA [2]IRSET Inserm UMR 1085 [3]École centrale de Nantes, LS2N

Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM) — 3rd–6th of July 2018

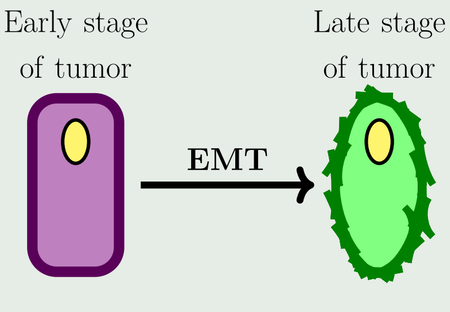
Introduction & Context

Hepatocellular carcinoma (HCC) is the most widespread type of liver cancer. Its occurrence is associated to chronic hepatitis (virus, alcohol, steatosis, etc.) that often happen with the development of fibrosis which, in its terminal phase, cirrhosis, is a major cause of HCC. Because of a late detection, the prognostic is very bad with a survival rate ranging from weeks to months. The aim of this work is to identify new invasiveness markers by exploiting the knowledge stored into **ontological and experimental databases**.

Transforming Growth Factor β (TGF-β)

The **TGF-β** is a protein implied in the signaling of many biological functions (cellular growth, differenciation, etc.). It plays a major role in the tumor progression of HCC by especially triggering the fibrosis, the **epithelial-mesenchymal transition** (EMT) and thus the tumor invasion.

Epithelial-Mesenchymal Transition (EMT)



The **EMT** happens when **epithelial** cancerous cells transform into **mesenchymal** cells, able to reshape the extra-cellular matrix and move in it, thus increasing tumor aggressiveness by creating **metastasis**. We use an **EMT signature** which is a set of genes (including TGF-β) known to be over-expressed during and after the EMT and thus constitutes a good marker to discriminate two tumor stages: early and non-invasive versus late and invasive. The signature is proposed by the Broad institute and is available in **MSigDB** [Subramanian et al., 2005].

Expression data (ICGC)

EMT signature (MSigDB)

Clustering & fold-change

Partial coloring: + / −

Genes of interest: 910

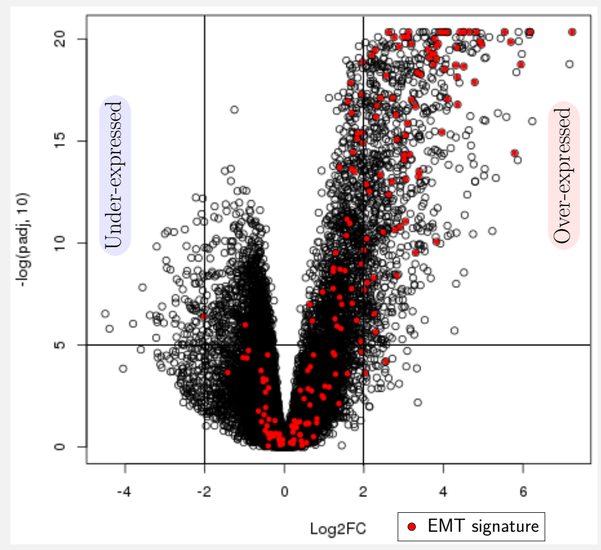
1. Select Differently Expressed Genes in Invasive Tumors

Data: **294 samples** extracted from liver cancers taken from the project LIHC-US of the database **ICGC** [Hudson et al., 2010].

Objective: Discriminate two groups (early and late stages of tumor) to carry a **differential analysis**, and establish two sets of genes that are over- and under-expressed in the late group compared to the early group.

Method: **Clustering study** based on the **EMT signature** of **MSigDB** (H-v6.1). The two sets are extracted by setting thresholds for the P-value ($P_{adj} < 10^{-5}$) and the fold-change ($\log_2(FC) > 2$ for the over-expression and $\log_2(FC) < -2$ for the under-expression).

Results: **821 over-expressed genes** and **89 under-expressed genes**, among which 80 are part of the EMT signature. The **volcano plot** on the right represents the P-value as a function of the fold-change in logarithmic scales. The genes of the EMT signature are marked in red.



Querying of Pathway Commons (BRAvo)

Graph of regulatory pathways: 1197 nodes, 10551 edges

2. Extract a Regulatory Graph from Pathway Commons

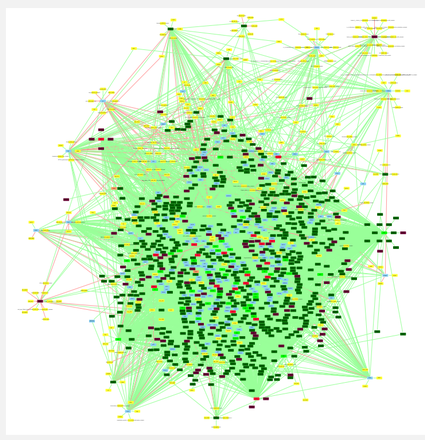
Data: Previous 910 seed genes (821 over- and 89 under-expressed genes)

Objective: Extract a **regulatory graph** from **Pathway Commons**, which is a gathering of 25 **pathway databases** [Cerami et al., 2010].

Method: The tool **BRAvo** [Lefebvre et al., 2017] allows to interrogate this database using SPARQL queries. From the 910 seed genes, it computes all (direct and indirect) upstream regulations.

Results: A graph in SIF format containing 1197 nodes and 10551 edges. Only 645 genes of the initial 910 genes were found.

Difficulties: **Pathway Commons** is very heterogeneous and the result requires postprocessing. Some bases should also be excluded from the search.



Consistency rules (Iggy)

Predictions: 0 22% | + 8.7% | − 2.4%

3. Apply Coloring Propagation to Obtain Gene Predictions

Data: Previous graph & sets of 821 over- and 89 under-expressed seed genes

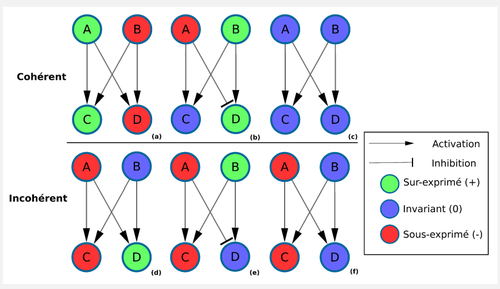
Objective: Infer if the new nodes in the graph should be **over-expressed** (+), **under-expressed** (−) or **invariant** (0) based on the value of the seed genes.

Method: The tool **Iggy** [Thiele et al., 2015] achieves this with consistency rules:

- An affectation (+, − or 0) must be explained by at least one predecessor;
- There must be a path between an affectation + or − and one of the 910 seed genes;
- Without such a path or in case of contradictory influences, coloring 0 can be affected.

A **prediction** is a node affectation that is common to all compatible colorings.

Results: 185 (34%) non-seed nodes are predicted, and 61 (11%) as "+" or "−".



Exploitation Validation

Other outcomes

Key controllers

Dynamical analysis

4. Exploitation, Validation & Outlooks

Exploitation: The obtained predictions may be suitable as new tumor invasiveness markers or therapeutic targets. We can also search for the **regulating miRNA** and point to those of particular interest.

Validation: Cross-validation (by hiding a part of the observations and checking the new predictions) would allow to assess the robustness of our approach.

Outlooks:

- Searching for **key controllers**: minimal sets of genes that have a maximum impact on predictions.
- **Dynamical analysis** on attractors reachability [Poret and Guziolowski, 2018].

References

Cerami, E. G. et al. (2010). Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39. http://www.pathwaycommons.org/.

Hudson, T. J. et al. (2010). International network of cancer genome projects. *Nature*, 464. http://icgc.org/.

Lefebvre, M. et al. (2017). Regulatory and signaling network assembly through linked open data. In *Journées Ouvertes en Biologie, Informatique et Mathématiques*. Demo paper. https://github.com/symetric-group/bionets-demo.

Poret, A. and Guziolowski, C. (2018). Therapeutic target discovery using boolean network attractors: improvements of kali. *Royal Society Open Science*, 5(2).

Subramanian, A. et al. (2005). Gene Set Enrichment Analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. of the Nat. Ac. of Sci.*, 102(43). http://software.broadinstitute.org/gsea/.

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