Search of Therapeutic Targets on the Hepatocellular Carcinoma with Database Extraction and Graph Coloring Methods

Recherche de cibles thérapeutiques pour le carcinome hépatocellulaire à l’aide d’extraction de bases de données et de méthodes de coloration de graphes

Maxime FOLSCHETTE
maxime.folschette@ls2n.fr — http://maxime.folschette.name/

Current occupation: CNRS & Institut Français de Bioinformatique (IFB) & LS2N
Previous occupation: Université de Rennes 1 & IRISA & IRSET

Co-authors: MF, Vincent Legagneux, Arnaud Poret, Carito Guziolowski, Nathalie Théret Marie Lefebvre, MF, Jérémie Bourdon, Carito Guziolowski, Alban Gaignard

2019-07-01
**Introduction**

**Context**

**Hepatocellular carcinoma (HCC)**

- Most widespread liver cancer, 3rd most deadly cancer
- Mainly associated with chronic inflammation and fibrosis
- Late diagnosis and difficult to treat (resection, transplant, chemo-embolization)
- Very low survival rate

**Objectives**

- Build gene signaling networks associated with HCC aggressiveness
- Predict key molecules explaining changes in gene expression data between low and high aggressive HCC
Therapeutic Targets for Hepatocellular Carcinoma

Experimental Data

LIHC-US in ICGC [Hudson et al., 2010]

Project for liver HCC (USA)

- 294 samples with gene expression data
- Primary tumor on solid tissue only
- 20502 genes
- 16282 genes when excluding low expression

But no tumor grade annotation for these samples

⇒ We need a criterion to distinguish aggressive and non-aggressive HCC

Objectives

1) Clustering on the criterion ⇒ Two groups
2) Differential analysis on the two groups
Epithelial-Mesenchymal Transition

**Epithelial-mesenchymal transition (EMT)**

- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to remodel the extra-cellular matrix and migrate
- Invasive cancer cells ⇒ metastasis
Epithelial-Mesenchymal Transition

<table>
<thead>
<tr>
<th>Epithelial cells</th>
<th>Mesenchymal cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adhesive</td>
<td>Motile &amp; invasive</td>
</tr>
<tr>
<td>Low aggressiveness</td>
<td>High aggressiveness</td>
</tr>
</tbody>
</table>

- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to remodel the extra-cellular matrix and migrate
- Invasive cancer cells $\Rightarrow$ metastasis
- Indication of **tumor aggressiveness**

**Epithelial-Mesenchymal transition (EMT)**

**Criterion**

- Epithelial-mesenchymal transition (EMT)
- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to remodel the extra-cellular matrix and migrate
- Invasive cancer cells $\Rightarrow$ metastasis
- Indication of **tumor aggressiveness**

**EMT signature**

- Set of genes that are over-expressed during EMT
- Downloaded on GSEA [Subramanian et al., 2005]
Epithelial-Mesenchymal Transition

Epithelial-mesenchymal transition (EMT)

- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to remodel the extra-cellular matrix and migrate
- Invasive cancer cells ⇒ metastasis
- Indication of tumor aggressiveness

- EMT signature = Set of genes that are over-expressed during EMT
- Downloaded on GSEA [Subramanian et al., 2005]
ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Genes of interest

Extraction of the pathways from Kegg (Stream)

Spread coloring and make predictions (Iggy)

Robustness analysis
Therapeutic Targets for Hepatocellular Carcinoma  

Clustering

- Defining 3 sample clusters based on GSEA_EMT gene set expression values

GSEA_EMT gene set (195 genes)

294 samples (LIHC-US)

Group A = Low expression of the EMT signature
Group C = High expression of the EMT signature
Therapeutic Targets for Hepatocellular Carcinoma

Workflow of the Project

1. ICGC expression data
   Clustering on EMT signature

2. Differential expression analysis
   → Genes of interest

3. Extraction of the pathways from Kegg (**Stream**)

4. Spread coloring and make predictions (**Iggy**)

5. Robustness analysis

2 groups

---

GSEA_EMT gene set (195 genes)

294 samples (LIHC-US)

1) Defining 3 sample clusters based on GSEA_EMT gene set expression values

A      B      C
Fold-change definition

- Consider groups $A$ (lowest expression of EMT) and $C$ (resp. highest)
- For each gene $g$, compute mean value for group $A$ (resp. $C$)
- Differential analysis:

$$\text{fold-change}(g) = \frac{\text{mean}_g(C)}{\text{mean}_g(A)}$$
Therapeutic Targets for Hepatocellular Carcinoma • Differential Analysis

-log(padj, 10)

Under-expressed

Over-expressed

Genes

EMT signature

Maxime FOLSCHETTE

10/32

Bioss-MP 2019 — 2019-07-01
Genes of Interest

Genes of interest

- 821 up-regulated genes
- 1092 down-regulated genes

= 1913 genes

Objectives

1) Extract a graph from Kegg [Kanehisa et al., 2017] using these genes, with the tool Stream
2) Coloring + predictions with Iggy [Thiele et al., 2015]
Workflow of the Project

1. **ICGC expression data**
   - Clustering on EMT signature

2. **Differential expression analysis**
   - Genes of interest

3. **Extraction of the pathways from Kegg**
   - (Stream)

4. **Spread coloring and make predictions (Iggy)**

5. **Robustness analysis**

- 2 groups
- ≈ 2,000 genes
### Kegg [Kanehisa et al., 2017]

- Homogeneous data
- Categories: 2. Genetic Information Processing  
  3. Environmental Information Processing  
  4. Cellular Processes  
  5. Organismal Systems
- Already formatted and curated by Arnaud Poret

SIF format: \( A \xrightarrow{+/-} B \)  
“\( A \) positively/negatively influences \( B \)”

- Genes (XXX_gen)  
- Proteins (XXX_prot)  
- Complexes (XXX:YYY:ZZZ)

### Stream (Arnaud Poret)

- Ad-hoc program for upstream graph extraction  
- Extract the part of the graph for which we have expression data (25%)
Graph content:
- 3'383 nodes
- 13'771 edges
  - 11'661 activations
  - 2'110 inhibitions

1913 genes from the differential expression
Only 209 are found in Kegg:
- 138 up-regulated
- 71 down-regulated
- 3174 new nodes

Nodes with up to:
- 92 incoming influences
- 79 outgoing influences
→ Nodes with a lot of impact on the network
Therapeutic Targets for Hepatocellular Carcinoma

Workflow of the Project

ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Genes of interest

Extraction of the pathways from Kegg (Stream)

Spread coloring and make predictions (Iggy)

Robustness analysis

2 groups

≈ 2'000 genes

≈ 3'400 nodes, 14'000 edges
Graph Coloring

- Coloring = information attached to nodes about over- or under-expression
  \( \begin{align*} 
  X &= \text{over-expressed} \\
  Y &= \text{under-expressed} 
  \end{align*} \)

Given by the experimental data
Graph Coloring

- Coloring = information attached to nodes about over- or under-expression
  - $X$ = over-expressed
  - $Y$ = under-expressed

Given by the experimental data

- Prediction = a node that is always colored the same
  - Here, only 1 prediction: $D$

All computed by Iggy [Thiele et al., 2015] (Answer Set Programming)
Graph Coloring

- Coloring = information attached to nodes about over- or under-expression
  - $X$ = over-expressed
  - $Y$ = under-expressed

Consistent

- Computed by Iggy [Thiele et al., 2015] (Answer Set Programming)
Graph Coloring

- Coloring = information attached to nodes about over- or under-expression
  - $X$ = over-expressed
  - $Y$ = under-expressed

![Graph Coloring Diagram]

- Consistent
- Prediction = a node that is always colored the same
- All computed by Iggy [Thiele et al., 2015] (Answer Set Programming)
Graph Coloring

- Coloring = information attached to nodes about over- or under-expression
  - \( X \) = over-expressed
  - \( Y \) = under-expressed

\[
\begin{align*}
\text{C} & \quad \text{D} \\
\text{A} & \quad \text{B}
\end{align*}
\]

- Consistent

\[
\begin{align*}
\text{C} & \quad \text{D} \\
\text{A} & \quad \text{B}
\end{align*}
\]

- Consistent

\[
\begin{align*}
\text{C} & \quad \text{D} \\
\text{A} & \quad \text{B}
\end{align*}
\]

- Inconsistent
Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

\( X \) = over-expressed

\( Y \) = under-expressed

Consistent

Inconsistent

Compute all colorings without inconsistencies

Prediction = a node that is always colored the same

Here, only 1 prediction:

\( D \)

All computed by Iggy [Thiele et al., 2015] (Answer Set Programming)
Graph Coloring

• Coloring = information attached to nodes about over- or under-expression
  \[ \text{X} = \text{over-expressed} \quad \text{Y} = \text{under-expressed} \]

\[
\begin{array}{c}
\text{C} \quad \text{D} \\
\downarrow \quad \downarrow \\
\text{A} \quad \text{B}
\end{array}
\quad
\begin{array}{c}
\text{C} \quad \text{D} \\
\downarrow \quad \downarrow \\
\text{A} \quad \text{B}
\end{array}
\]

Consistent
Consistent
Inconsistent
Inconsistent

• Compute all colorings without inconsistencies

• Prediction = a node that is always colored the same
  Here, only 1 prediction: \( \text{D} \)

• All computed by \textit{Iggy} [Thiele et al., 2015] (Answer Set Programming)
Knowledge from experiments:
- 138 up-regulated
- 71 down-regulated

Computational predictions:
- 92 predicted
  - 24 non-trivial
- 54 predicted
  - 33 non-trivial

70% more information compared to only knowledge from experiments
Computational predictions (results of Iggy)

\[ \log_2(\text{fold-change}) \]

\[-\log_{10}(\text{Padj})\]
Therapeutic Targets for Hepatocellular Carcinoma

Workflow of the Project

ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Genes of interest

Extraction of the pathways from Kegg (Stream)

Spread coloring and make predictions (Iggy)

Robustness analysis

2 groups

≃ 2'000 genes

≃ 3’400 nodes, 14’000 edges

Predictions:
92 + and 54 -
Matching between computational predictions and ICGC expression data:

- 209 inputs
  - 124 match
    - 36 non-trivial
  - 17 do not match
    - 16 non-trivial
  - 5 not found in ICGC data

88% matching
69% non-trivial

→ Good overlap
Cross-Validation

**Sampling**
- Consider a range of samplings (10%, 15%, 20%, ... 95%)
- Randomly pick x% of under- and over-expressed genes (observations)
- Compute the predictions on this sample; repeat 100 times

**Score compared to the original data**
- Compare the predictions to the original ICGC data
  - Scores converge to the final score at 100%

**Robustness of the prediction of each node**
- Compare the predictions to the final sampling of 100%
  - Not a lot of variability in the prediction types → Robust
Workflow of the Project

ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Genes of interest

Extraction of the pathways from
Kegg (Stream)

Spread coloring and
make predictions (Iggy)

Robustness analysis

2 groups

≃ 2’000 genes

≃ 3’400 nodes, 14’000 edges

Predictions:
92 + and 54 −

Robust predictions
### Prediction Results

#### New results compared to ICGC: complexes

**Complexes predicted:**

- NFKB1::BCL3 (+)
- NFKB2::RELB (+)
- JUND::NACA (−)

#### Results conflicting with ICGC data

Computational predictions which are **different from differential analysis:**

- BAK1_gen, BMP4_gen, CREB1_prot, EIF4EBP2_prot, IGFBP3_gen, IGFBP3_prot, NR0B2_gen, NR0B2_prot, NR1H4_gen, NR1H4_prot, NR3C2_gen, NR3C2_prot, SESN3_gen, SESN3_prot, THBS1_gen, TNFRSF10A_gen, TP53_prot
Therapeutic Targets for Hepatocellular Carcinoma

Results

Biological Validation

Validated by experimental data

Link with cancer validated by literature

New knowledge in aggressive HCC:

• up-regulation of NFKB2::RELB
• down-regulation of JUND::PACA
Therapeutic Targets for Hepatocellular Carcinoma

Results

Biological Validation

Validated by experimental data

- New knowledge in aggressive HCC:
  - up-regulation of NFKB2::RELB
  - down-regulation of JUND::PACA
Results

Biological Validation

- New knowledge in aggressive HCC:
  - up-regulation of NFKB2::RELB
  - down-regulation of JUND::PACA

- Validated by experimental data
- Link with cancer validated by literature
Results

Biological Validation

- Validated by experimental data
- Link with cancer validated by literature
- New knowledge in aggressive HCC:
  - up-regulation of NFKB2::RELB
  - down-regulation of JUND::PACA
Using Pathway Commons
### Pathway Commons [Cerami et al., 2010]

- A gathering of **25 pathway databases**
- Contains: PID, Kegg, Reactome, CTD, Panther, ...

### Benefits

- Common ontology (BioPAX)
- Freely available *via* a SPARQL endpoint (programmatic queries)
- Much more data than Kegg alone ⇒ better coverage

### Drawbacks

- **Very heterogeneous data**, duplicated entities
- The BioPAX ontology is big and difficult to use
- Made for biologists rather than modelers or computer scientists

→ Could Pathway Commons be used instead of Kegg?
The BioPAX Ontology

- NOG
- bp:displayName

- R1_controller
  - bp:controller
  - bp:controlled
  - bp:controlType
  - bp:TemplateReactionRegulation
  - bp:controlType

- R1
  - bp:controlType
  - rdf:type
  - bp:controlled
  - bp:TemplateReactionRegulation
  - bp:controlled

- R1_controlled
  - bp:participant

- P1_protein
  - bp:displayName
  - LAMA2

- AAP/Aluminum
  - bp:displayName

- CTD datasource
  - bp:dataSource
  - bp:TemplateReactionRegulation
  - bp:dataSource

- R2_controller
  - bp:controller

- R2
  - bp:controlType
  - rdf:type
  - bp:controlled

- R2_controlled
  - bp:participant

- INHIBITION
**BRAvo [Lefebvre et al., 2017]**

- **Interrogates** Pathway Commons with SPARQL queries
- Upstream graph reconstruction
- Written in Python; available soon (open source)

**Benefits of BRAvo**

- Fast reconstruction: 10 mins for a graph with 1402 nodes
- Unification of duplicated nodes based on synonyms
- Regulation or signaling
- Source selection
- SIF output
910 genes of interest
Regulation graph

**Graph content:**
- 1,402 nodes
- 2,804 edges

641/910 genes found

**Computational predictions:**
- 40 predicted +
- 32 predicted −

12% more information

Using additional synonyms unification:
- 26 match
- 28 do not match

47% matching
## Summary & Conclusion

### Summary

- Clustering + diff analysis: 2 lists of over- and under-expressed genes
- Graph extracted from Kegg: regulation + signaling
- 146 computational predictions (57 non-trivial)
- Computational & biological validations

### Ongoing work

- General pipeline of the whole method
- Try other sources (Pathway Commons with BRAvo)
- Finalizing manuscripts; submission soon

### Other objectives (to do)

- Explore survival curves compared to most robust genes
- Search for proliferation signatures
- Try the same workflow on a different type of cancer (breast?)
Hepatocellular carcinoma computational models identify key protein-complexes associated to tumor progression

Vincent Legagneux: Inserm, Irset (Rennes)
Arnaud Poret: LS2N (Nantes)
Carito Guziolowski: École centrale de Nantes, LS2N (Nantes)
Nathalie Théret: Inserm, Irset, IRISA (Rennes)

Special thanks to Anne Siegel

BRAvo: Regulatory and signaling network assembly from Pathway Commons

Marie Lefebvre: INRA (Bordeaux)
Jérémie Bourdon: Université de Nantes, LS2N (Nantes)
Carito Guziolowski: École centrale de Nantes, LS2N (Nantes)
Alban Gaignard: CHU de Nantes, Institut du Thorax, LS2N (Nantes)
Bibliography

Pathway Commons, a web resource for biological pathway data.
Nucleic acids research, 39.
http://www.pathwaycommons.org/.

Hudson et al. (2010).
International network of cancer genome projects.
http://icgc.org/.

KEGG: new perspectives on genomes, pathways, diseases and drugs.

Lefebvre, M., Bourdon, J., Guziolowski, C., and Gaignard, A. (2017).
Regulatory and signaling network assembly through linked open data.
In Journées Ouvertes en Biologie, Informatique et Mathématiques.
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/.

Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies.
BMC Bioinformatics, 16(1).
http://bioasp.github.io/iggy/.
**NFkB1::BCL3 (++)**

Part of the NF\(\kappa\)B pathway
- Bcl3 alters NF\(\kappa\)B signaling pathways, associated with cancer
- NFkB1 (p105) is activated to p50 \(\Rightarrow\) early recurrence of HCC
- Increased p50 and Bcl3 reported in tumors

**NFkB2::RELB (++)**

Activates CCL19 and CCL21
- Increase of CCL19 and CCL21 validated in experimental data

**JUND::NACA (−)**

- Regulates BGLAP (osteocalcin), down-regulated in the serum of HCC patients
- Regulates LRP5, validated in experimental data (decrease)
- Decrease of JUND and NACA validated in experimental data
Hub example: TP53_prot

18 predictions directly depend of TP53_prot
Initial ICGC data, EMT signature & genes found in Kegg

\[
\log_2(\text{fold-change})
\]

\[
-\log_{10}(\text{Padj})
\]
Boxplot of the scores for each sampling
Evolution of max, min, mean and median of good, bad and missing predictions compared to 100% sampling
Trivial Predictions

“Trivial” prediction

- Protein predicted the same as its observed gene
- Rarely brings new information
- Useful for validation