Prioritization of Master Regulators Through Influence Maximization

Clémence Réda, Andrée Delahaye-Duriez

NEURODIDEROT
Unité Mixte de Recherche UMR 1141
Inserm-Université Paris Diderot

Inserm
Université Paris Cité
Université Sorbonne Paris Nord
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Master regulator genes are at the top of the gene regulation hierarchy and allow to better understand regulatory dynamics. **However, current detection methods do not take into account regulatory cascades.** Here, we apply a novel method to identify master regulatory genes linked to epilepsy.

I. Reproducible Inference of a Boolean Network

First, our work focus on combining public data sources to design an end-to-end pipeline for the **synthesis of a dynamic gene regulatory network**, **starting from a subset of genes**. This network models the regulatory dynamics in a well-chosen cell line.

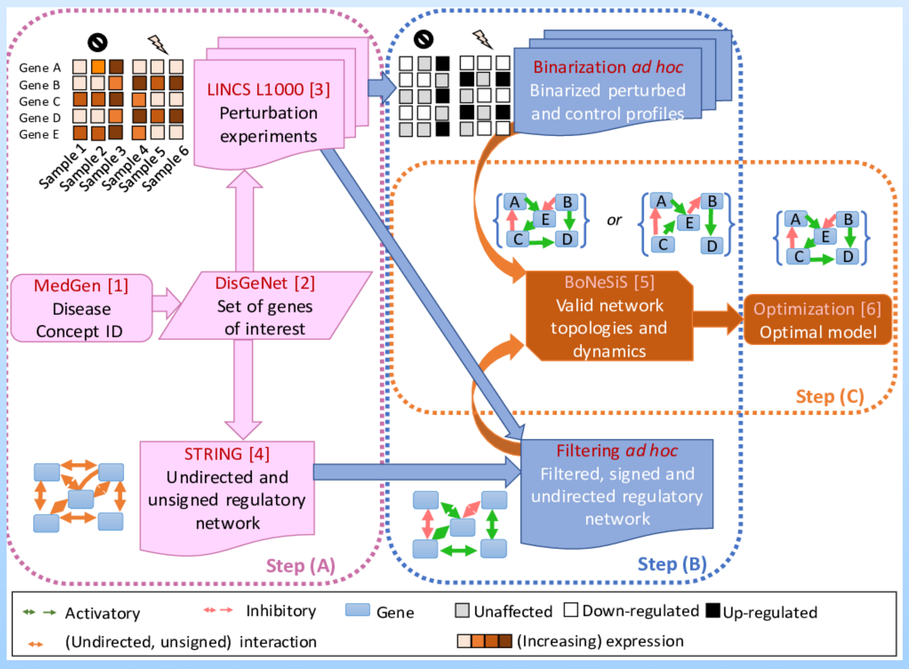

Figure 1. The Boolean network identification pipeline. **(A) data collection. (B) data processing. (C) inference.**

In the application to **epilepsy**, we considered the gene module M30 associated with epileptic *de novo* mutations [7], and gene expression data from neural progenitor and neuroblastoma cell lines.

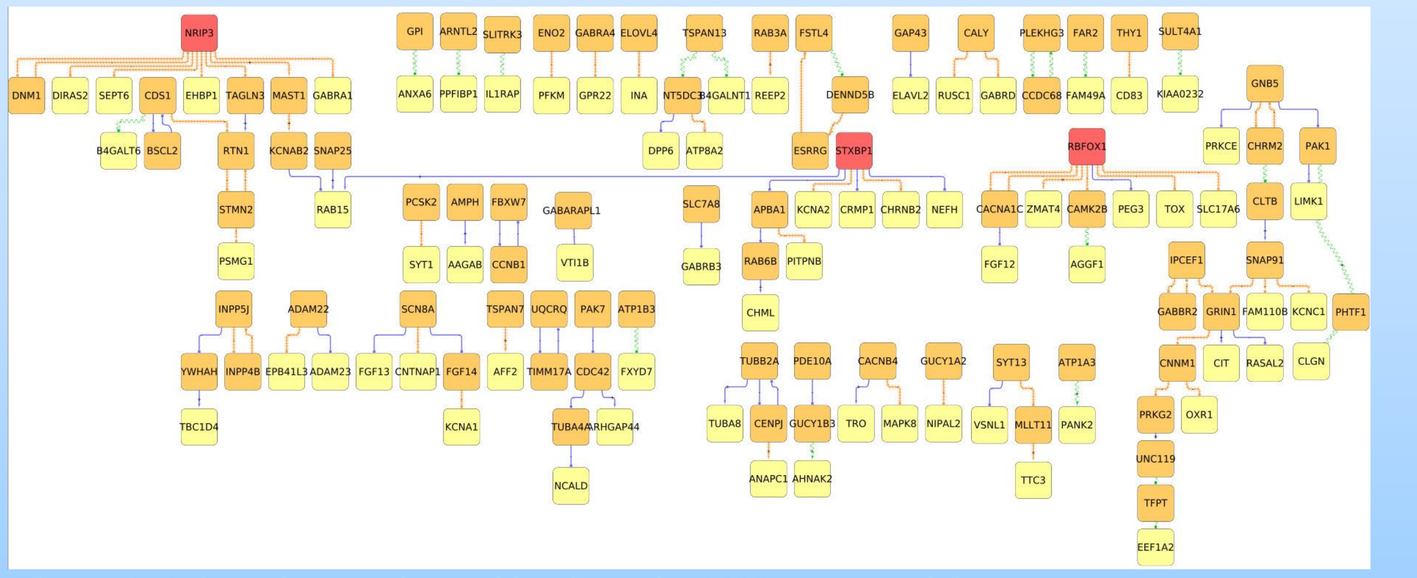

Figure 2. Inferred network resulting from the M30 gene module.

II. Dynamic Detection of Master Regulators

Second, we defined the concept of "gene influence", in terms of **transcriptomic impact of gene perturbation in this network**, called spread value. For any gene n and initial state i,

$$\mathrm{SV}_{\mathcal{B}}(\{n\}, i) = 1 - \max\left\{\mathcal{S}\left(a^1_{|\mathcal{O}}, a^2_{|\mathcal{O}}\right) : a^1, a^2 \in \mathcal{A}(i,\emptyset) \times \in \mathcal{A}(i, \{(n, \neg i[n])\})\right\}$$

where

- $\mathcal{O}$ — set of output genes
- $a^1_{|\mathcal{O}}, a^2_{|\mathcal{O}}$ — attractor states restricted to output genes
- $\mathcal{S}$ — similarity function between attractor states
- $\mathcal{A}(i,\emptyset)$ — set of attractors reachable from state i without perturbation
- $\mathcal{A}(i, \{(n, \neg i[n])\})$ — set of attractors reachable from state i under the perturbation of n in the opposite direction than its level in state i

This spread value can be extended to sets of genes and of initial states.

In the application to **epilepsy,** we selected as initial states profiles from human hippocampi afflicted with temporal lobe epilepsy in [8]. A similarity function was selected to penalize in a symmetric way differences in zeroes and ones,

$$\mathcal{S}(a^1, a^2) = 1 - \frac{1}{d}\sum_{i=1}^{d} |a^1[i] - a^2[i]|$$

Finally, we applied an **influence maximization algorithm** [9] to retrieve genes with highest regulatory influence on the remainder of the network.

Input: $\mathcal{B}$ a Boolean network on node set V ; K the minimal number of simultaneous perturbations on the network ; $\mathcal{I}$ set of initial Boolean states
Initialize $\mathcal{N} = \emptyset$, $k = 0$
repeat
  $k \leftarrow k + 1$
  # Adding to set $\mathcal{N}$ nodes that maximize the spread value
  $\mathcal{N} \leftarrow \mathcal{N} \cup N_k$, where $N_k \leftarrow \arg\max_{n \in V \setminus \mathcal{N}} \mathrm{SV}_{\mathcal{B}}(\mathcal{N} \cup \{n\}, \mathcal{I})$
until $k = K$ or $\max_{n \in V \setminus \mathcal{N}} \mathrm{SV}_{\mathcal{B}}(\mathcal{N} \cup \{n\}, \mathcal{I}) \leq \mathrm{SV}_{\mathcal{B}}(\mathcal{N}, \mathcal{I})$
Output: $\mathcal{N}$

Algorithm 1. Influence maximization algorithm for Boolean networks.

The iteratively built set $\mathbf{N_K}$ is the set of **possible K-sized gene subsets to simultaneously perturb on the network**, such that the set of attractors reachable from initial set **I** is greatly modified.

This approach, which combines the synthesis of a Boolean network and influence maximization, **can generically be applied to any disease**.

To adapt this pipeline to another disease, one needs to change the gene subset and the cell line(s) on which the network should be built, as well as the set of initial network states for the detection of master regulators.

III. Validation of master regulator candidates

Fig. 3 (**left**) displays the Spearman's ρ correlation heatmap between spread values, network centrality measures (Control Centrality [10]), and scores associated with the pathogenicity of genes (pLI [11], RVIS [12], EDS [13]) in M30. **Spread values are consistent and both correlated with network-dependent measures** (strongly) **and gene pathogenicity measures.**

Morever, a over-representation enrichment analysis (ORA) shows that **top genes for spread** (Fig. 4) **are significantly enriched in epilepsy-related terms at level 5%, w.r.t. the whole M30 module** (Fig. 3, **right**).

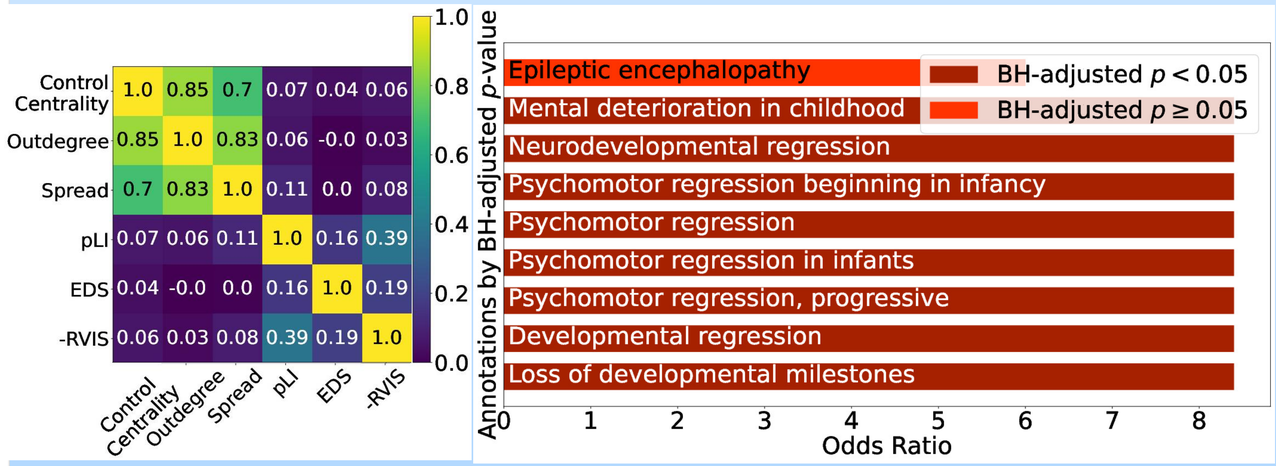

Figure 3. Correlation with network centrality and gene pathogenicity measures (**left**), pathway enrichment analysis (ORA) on candidates (**right**).

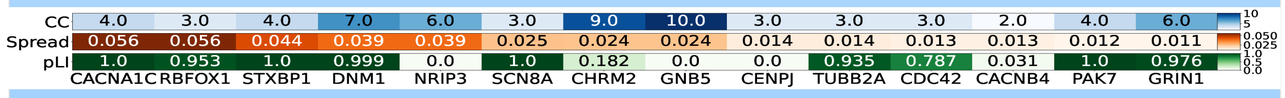

Figure 4. Top genes for spread value (**center**) with associated Control Centrality (CC, **top**) and pLI score (**bottom**).

[1] Doğan et al. (2014) DOI: 10.1016/j.jbi.2013.12.006. [2] Piñero et al. (2020) DOI: 10.1093/nar/gkz1021. [3] Subramanian et al. (2017) DOI: 10.1016/j.cell.2017.10.049. [4] Szklarczyk et al. (2021) DOI: 10.1093/nar/gkaa1074. [5] Chevalier et al. (2019) DOI: 10.1109/ICTAI.2019.00014. [6] Babichev et al. (2019) **Technique of Gene Regulatory Networks Reconstruction Based on ARACNE Inference Algorithm**. [7] Delahaye-Duriez et al., (2016) DOI: 10.1186/s13059-016-1097-7. [8] Mirza et al. (2017) DOI: 10.1093/hmg/ddx061. [9] Kempe et al. (2003) **Maximizing the Spread of Influence through a Social Network**. [10] Liu et al. (2012) DOI: 10.1371/journal.pone.0044459. [11] Lek et al. (2016) DOI: 10.1038/nature19057. [12] Petrovski et al. (2013) DOI: 10.1371/journal.pgen.1003709. [13] Wang et al. (2020) DOI: 10.1016/j.ajhg.2020.01.012.

SCAN ME

GitHub code repository

Discussion

This methodology allows a reproducible detection of master regulators, introducing for the first time **a measure which takes into account transcriptional cascades on gene expression**. It reduces the amount of data needed as input, which is one of the main caveats of researching on rare diseases.