NORDic: a Network-Oriented package for the Repurposing of Drugs

Clémence Réda [1] and **Andrée Delahaye-Duriez** [1,2,3]

1 Université Paris Cité, Neurodiderot, Inserm, F-75019 Paris **2** Université Sorbonne Paris Nord, UFR de santé, médecine et biologie humaine, F-93000 Bobigny **3** Unité fonctionnelle de médecine génomique et génétique clinique, Hôpital Jean Verdier, AP-HP, F-93140 Bondy

Introduction

Genes, proteins and messenger RNAs are shown to interact on each other in order to modulate gene activity. Gene regulatory networks, which are graphs connecting biological entities according to their known regulatory interactions, are useful models that enable a better understanding of those regulatory mechanisms (Karlebach & Shamir, 2008).

In particular, one type of gene regulatory networks, called Boolean networks, allows the definition of so-called regulatory functions (Kauffman, 1969; Thomas, 1973). Those functions are specific to each node in the graph, and determine the activity of this node according to its regulators. Those functions are defined on the Boolean domain (**True** or **False**), meaning that we only consider binary gene activities. Subsequently, studying this type of networks as a dynamical system remains rather tractable (Moon et al., 2022). The potential applications are numerous. Taking into account network dynamics should improve tools originally developed using non Boolean networks for the identification of interesting biomarkers (Nicolle et al., 2015) or drug repurposing.

However, the construction and analysis of Boolean networks become extremely tedious and time-consuming in the absence of experimental data or when considering the activity of a large number of genes at a time (Collombet et al., 2017). Moreover, the identification of interesting drug targets, via the detection of master regulators suffers from not exploiting the full network topology, which might account for toxic unexpected side effects (Bolouri & Davidson, 2003; Huang et al., 2019). Finally, regulatory mechanisms at transcriptomic level are inherently stochastic (Raj & Van Oudenaarden, 2008). As a consequence, naive algorithms for Boolean network-based *in silico* drug repurposing rely on testing a given drug a large number of times, in order to get a good estimate of its effect on gene activity. Such methods might resort to the simulation of drug treatment on Boolean network in either a patient-specific approach (Montagud et al., 2022), or by ignoring the stochastic part of gene regulation.

Statement of need

The development of **NORDic** relies on avoiding *ad hoc* solutions, by implementation of approaches which are relevant to all kinds of diseases regardless of the level of knowledge present in the literature. Please refer to Figure 1 for an overview of the package. Solutions proposed in this package emphasize on, first, the modularity of the methods, by providing functions which can tackle different types of regulatory dynamics for instance; second, on the transparency of the approaches, by allowing the finetuning of each method through parameters with a clearly defined impact on the result.
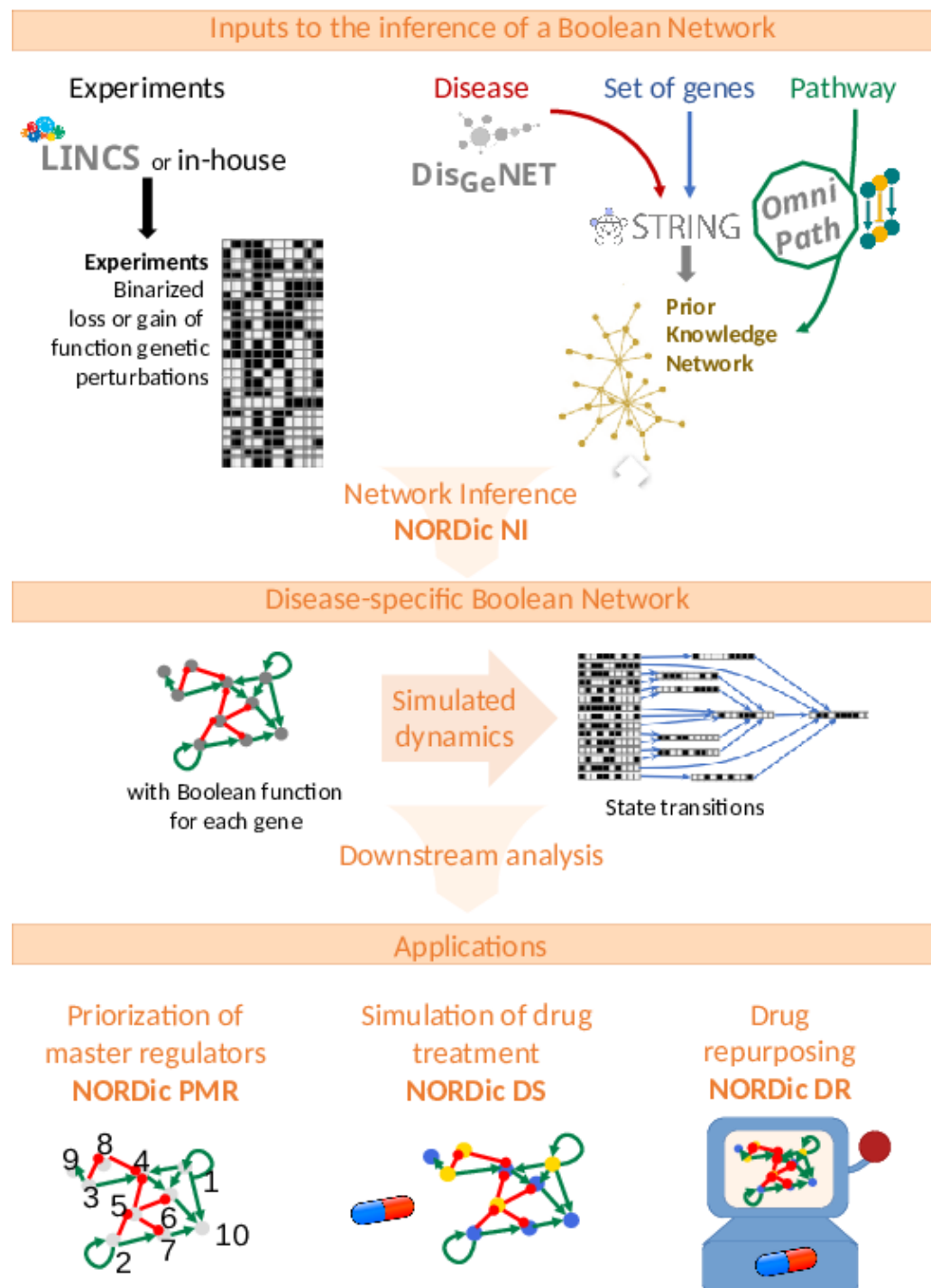
Figure 1: Overview of the different modules in NORDic.

Automated identification of disease-related Boolean networks

Most prior works about building Boolean networks assume the existence of a preselected set of known regulatory interactions and/or a set of perturbation experiments, where the gene activity of a subset of genes is measured after a single gene perturbation, for a group of genes

of interest. However, for rare diseases for instance, pinpointing a subset of genes of interest is already a difficult task by itself.

Moreover, there exist two hurdles to building Boolean networks which are specific to the Boolean framework. First, gene activity data must be binarized, meaning that one has to decide when a given gene is considered active or inactive in each sample. Such a process leads to an unavoidable loss of information. In order to avoid bias in the inference process, this step should be data-driven and user-controlled. For instance, when using PROFILE (Béal et al., 2021), a majority of genes might end up with an undetermined status –meaning that they are considered neither significantly strongly nor weakly active– which considerably undermines the input from experimental constraints.

Second, the problem of identification of a Boolean network is usually underdetermined, as there is too few of experiments and measurements in practice, compared to the size of the considered gene set.

Module **NORDic Network Identification (NI)** addresses these issues in an automated and user-controllable manner, by performing information extraction from large online sources of biological data, and data quality filtering according to user-selected parameters, which control every step of the process. As such, the hope is that **NORDic** makes the generation of disease-specific Boolean networks easier, reproducible, even in the absence of previously curated experiments, prior knowledge networks, or even a set of disease-associated genes. The pipeline implemented in **NORDic** was applied to epilepsy in a preliminary work (Réda & Delahaye-Duriez, 2022).

Prioritization of master regulators in Boolean networks

The identification of master regulators might relate to the disease onset or affected biological pathways of interest. Most prior works (Wu et al., 2018) either only leverage topological information about the network, and do not take into account gene activity data related to the disease; or do not take into account regulatory effects which trickle down the network, beyond the targets directly regulated by the gene (Zerrouk et al., 2020). That is, the gene activity context does not impact the genewise values computed on the network.

Module **NORDic PMR** detects master regulators in a Boolean network, given examples of gene activity profiles from patients. In contrast to prior works, the score assigned to (groups of) master regulators takes into account the network topology as well as its dynamics with respect to the diseased profiles. The approach, based on a machine learning algorithm solving the influence maximization problem (Kempe et al., 2003), is described in Réda & Delahaye-Duriez (2022).

Novel approaches for scoring drug effects & repurposing drugs

NORDic also proposes to tackle two problems related to drug repurposing: first, drug scoring, based on its ability to reverse the diseased gene activity profile (**NORDic DS**); second, the computation of an online sampling procedure which determines which drugs to test during drug screening for repurposing, in order to guarantee a bound on the error in recommendation, while remaining as sample-efficient as possible (**NORDic DR**).

There exist other approaches performing signature reversion, as mentioned in introduction. However, module **NORDic DS** (since version 2.0 of **NORDic**) is the first package to implement drug scoring based on Boolean networks, which can apply to any disease –for instance, it does not need the definition of specific biological phenotypes that should be observed after exposure to treatment (Montagud et al., 2022). The method implemented in **NORDic DS** is described in Réda (2022).

Similarly, module **NORDic DR** is the first approach that aims at solving the lack of guarantees in recommendation error. **NORDic DR** relies on bandit algorithms, which are sequential

reinforcement learning algorithms that enable the recommendation of most efficient drugs. Based on Boolean network simulations performed on the fly, those algorithms can adaptively select the next drug to test in order to perform recommendations with as few samples as possible. Algorithms implemented in **NORDic DR** are described and theoretically analyzed in Réda, Kaufmann, et al. (2021) (for the *m-LinGapE* algorithm), and in Réda, Tirinzoni, et al. (2021) (*MisLid* algorithm).

Extraction of information from large public data sets & simulation module

In all four present modules in **NORDic**, helper functions in module **NORDic UTILS** are implemented in order to extract and curate data in a transparent way from the LINCS L1000 (Subramanian et al., 2017), OmniPath (Türei et al., 2016), DisGeNet (Piñero et al., 2016) and STRING (Szklarczyk et al., 2021) databases. **NORDic** also proposes a simulation module, which allows to test the accuracy of the predictions made by the network compared to known measurements. This module also enables the study and the visualization of the behaviour of the network under various perturbations and types of regulatory dynamics.

Summary

Building a representation of gene interactions and their influences on gene activity, in an automated and reproducible way, helps to model more complex diseases and biological phenomena on a larger set of genes. These models might speed up the understanding of the gene regulation hierarchy by bioinformaticians and biologists; and allow to predict novel drugs or gene targets which might be investigated later for healthcare purposes. In particular, the network-oriented approach might be able to predict off-targets. The **NORDic** Python package aims at tackling those problems, with a focus on reproducibility and modularity. It primarily relies on popular formats for network description files, such as the .bnet format. Moreover, **NORDic** enables further study of the network in Cytoscape, by providing a direct conversion to .sif formats, along with a dedicated style file. The different pipelines present in **NORDic** produce intermediary files, which might be checked by the user, and can be fed again to the pipeline in order to reproduce the results.

To get started with the different modules proposed in **NORDic**, please check out the tutorials (Jupyter notebooks) on the GitHub repository (Réda & Delahaye-Duriez, 2023), which provides an application to a disease called Congenital Central Hypoventilation Syndrome (CCHS).

Acknowledgements

References

Béal, J., Pantolini, L., Noël, V., Barillot, E., & Calzone, L. (2021). Personalized logical models to investigate cancer response to BRAF treatments in melanomas and colorectal cancers. *PLOS Computational Biology*, *17*(1), e1007900. https://doi.org/10.1371/journal.pcbi.1007900

Bolouri, H., & Davidson, E. H. (2003). Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics. *Proceedings of the National*

Academy of Sciences, *100*(16), 9371–9376. https://doi.org/10.1073/pnas.1533293100

Collombet, S., Oevelen, C. van, Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences*, *114*(23), 5792–5799. https://doi.org/10.1073/pnas.1610622114

Huang, Y., Furuno, M., Arakawa, T., Takizawa, S., Hoon, M. de, Suzuki, H., & Arner, E. (2019). A framework for identification of on-and off-target transcriptional responses to drug treatment. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-54180-4

Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, *9*(10), 770–780. https://doi.org/10.1038/nrm2503

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, *22*(3), 437–467. https://doi.org/10.1016/0022-5193(69)90015-0

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146. https://doi.org/10.1145/956750.956769

Montagud, A., Béal, J., Tobalina, L., Traynard, P., Subramanian, V., Szalai, B., Alföldi, R., Puskás, L., Valencia, A., Barillot, E., & others. (2022). Patient-specific boolean models of signalling networks guide personalised treatments. *Elife*, *11*, e72626. https://doi.org/10.1101/2020.05.27.119016

Moon, K., Lee, K., & Paulevé, L. (2022). Computational complexity of minimal trap spaces in boolean networks. *arXiv Preprint arXiv:2212.12756*. https://doi.org/10.48550/arXiv.2212.12756

Nicolle, R., Radvanyi, F., & Elati, M. (2015). CoRegNet: Reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics*, *31*(18), 3066–3068. https://doi.org/10.1093/bioinformatics/btv305

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. I. (2016). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, gkw943. https://doi.org/10.1093/nar/gkw943

Raj, A., & Van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, *135*(2), 216–226. https://doi.org/10.1016/j.cell.2008.09.050

Réda, C. (2022). *Combination of gene regulatory networks and sequential machine learning for drug repurposing* [PhD thesis]. Université Paris Cité.

Réda, C., & Delahaye-Duriez, A. (2022). Prioritization of candidate genes through boolean networks. *Computational Methods in Systems Biology: 20th International Conference, CMSB 2022, Bucharest, Romania, September 14–16, 2022, Proceedings*, 89–121. https://doi.org/10.1007/978-3-031-15034-0_5

Réda, C., & Delahaye-Duriez, A. (2023). Network oriented repurposing of drugs (NORDic) version 2.4.1. In *GitHub repository*. https://github.com/clreda/nordic; GitHub. https://doi.org/10.5281/zenodo.7239047

Réda, C., Kaufmann, E., & Delahaye-Duriez, A. (2021). Top-m identification for linear bandits. *International Conference on Artificial Intelligence and Statistics*, 1108–1116. https://doi.org/10.48550/arXiv.2103.10070

Réda, C., Tirinzoni, A., & Degenne, R. (2021). Dealing with misspecification in fixed-confidence linear top-m identification. *Advances in Neural Information Processing Systems*, *34*, 25489–25501. https://doi.org/10.48550/arXiv.2111.01479

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., & others. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, *171*(6), 1437–1452. https://doi.org/10.1016/j.cell.2017.10.049

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., & others. (2021). The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, *49*(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074

Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, *42*(3), 563–585. https://doi.org/10.1016/0022-5193(74)90172-6

Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, *13*(12), 966–967. https://doi.org/10.1038/nmeth.4077

Wu, L., Li, M., Wang, J., & Wu, F.-X. (2018). CytoCtrlAnalyser: A cytoscape app for biomolecular network controllability analysis. *Bioinformatics*, *34*(8), 1428–1430. https://doi.org/10.1093/bioinformatics/btx764

Zerrouk, N., Miagoux, Q., Dispot, A., Elati, M., & Niarakis, A. (2020). Identification of putative master regulators in rheumatoid arthritis synovial fibroblasts using gene expression data and network inference. *Scientific Reports*, *10*(1), 16236. https://doi.org/10.1038/s41598-020-73147-4